AN EVALUATION METHOD FOR FEATURE RANKINGS

Ivica Slavkov

Doctoral Dissertation
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia, July 2012

Evaluation Board:

Asst. Prof. Bernard Ženko, Chairman, Jožef Stefan Institute, Ljubljana, Slovenia

Assoc. Prof. Dr. Marko Robnik-Šikonja, Member, Faculty of Computer and Information Science, University of Ljubljana, Slovenia

Dr. Benedikt Brors, Member, German Cancer Research Center, Heidelberg, Germany

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

Ivica Slavkov

AN EVALUATION METHOD FOR FEATURE RANKINGS

Doctoral Dissertation

METODA VREDNOTENJA UREJENOSTI ZNAČILK

Doktorska disertacija

Supervisor: Prof. Dr. Sašo Džeroski

Ljubljana, Slovenia, July 2012

To my parents

На моите родители

Contents

Abstract

Feature ranking is the machine learning task of inducing an ordering of features in a given dataset according to some notion of relevance. We consider the feature ranking task in the context of supervised learning, where the notion of feature relevance is defined with respect to a target concept. Feature ranking is rarely considered as a standalone task, since it usually precedes some other machine learning task, such as learning predictive models. Because of that, the evaluation of feature ranking algorithms, in a real-world setting, is always performed in a utility oriented manner.

The main focus of the work presented in this thesis is the evaluation of feature rankings. More specifically, three goals related to the evaluation of feature rankings are addressed. The first is to formally define and quantify a ground truth feature ranking. As a result we obtain a synthetic controlled setting that can be used to evaluate feature rankings. The second goal is to propose an evaluation method for feature rankings. This method should be able to estimate and compare the quality of feature rankings produced by different approaches. The third goal is to identify practical scenarios where feature ranking evaluation is needed and to demonstrate the usefulness of the proposed method.

We first formally define the feature ranking task. Based on definitions given in the literature, in the context of feature selection, the purpose of feature ranking is to solve the all-relevant feature selection problem and provide a correct ordering of the relevant features. The output of the feature ranking task is an ordered set of features, called a feature ranking.

Feature ranking is based on the definition of relevance. We thus provide a survey of the various definitions and measures of feature relevance that exist in the literature. A common deficiency of most of these feature relevance measures is that they treat each feature independently. To overcome this, we propose to quantify feature relevance by a measure based of feature interactions, grounded in information theory. By using this definition of relevance, we are able to produce a ground truth ranking and define a synthetic controlled setting for evaluating feature rankings.

Next, we propose an evaluation method for feature rankings. Intuitively, the method is based on the utility of feature ranking as a filter method for feature selection. It is a formalised algorithmic procedure based on stepwise construction of k-ranked feature subsets and subsequent construction of predictive models. The output of the evaluation method are the so-called error curves that show how the relevant features are distributed across a feature ranking. These curves can be further used to comparatively evaluate the quality of different feature ranking methods.

The proposed ranking evaluation method is tested in a controlled setting, where the definition of ground truth ranking is used. The experiments are based on adding different levels of noise to the known ground truth ranking and evaluating them with the proposed method. Both uniform and non-uniform noise addition to the ground truth ranking is considered. The results show that the evaluation method can successfully detect the decrease in quality of the feature rankings as higher and higher levels of noise are considered.

We then perform three empirical studies, covering different domains. The first investigates the behaviour of different feature ranking algorithms on both synthetic and various real datasets. The results reveal that ReliefF is the best performing method on the synthetic

data. On real data, there is no statistically significant difference in quality of the rankings produced by ReliefF and the information gain.

The second study investigates the usefulness of feature ranking ensembles (FREs). The aim of the analysis is to find under what conditions FREs produce rankings with better quality, as compared to the rankings produced by individual ranking methods. The results show that constructing FREs has a profound effect only when considering unstable base ranking methods, such as the ones provided by Random Forests.

The last study is from the domain of cancer research, more specifically the area of embryonal tumours (ETs). The aim of the analysis is to identify key genes involved in tumour aggressiveness. Our method helped in this task by identifying the feature ranking method that produces the best ranked list of candidate genes. This was also confirmed by a subsequent construction of gene networks for the top-ranked genes.

Finally, we conclude the thesis by discussing several possibilities for further work in this area. The first and the simplest extension would be in the experimental evaluation performed. Namely, in our work we considered only classification and regression target concepts. We would like to extend this to also include experiments performed on structured targets. The second line of further development would be to combine several aspects of feature rankings, including stability, when evaluating feature ranking methods. The final direction for further work is to evaluate feature rankings by considering different utilities, besides their predictive performance. An example for this was in the last case study conducted, when the feature rankings were used to construct gene networks.

Povzetek

Urejanje značilk (ang. feature ranking) je naloga strojnega učenja pri kateri želimo značilke iz dane množice podatkov urediti glede na neko mero pomembnosti. V disertaciji obravnavamo urejanje značilk v okviru nadzorovanega učenja, zato je pojem pomembnosti značilk opredeljen glede na ciljni koncept. Urejanje značilk je le redko obravnavano kot samostojna naloga, saj jo skoraj vedno izvajamo pred drugimi nalogami strojnega učenja, kot je na primer učenje napovednih modelov. Iz tega razloga v praksi vrednotenje algoritmov za urejanje značilk vselej izvajamo glede na določen ciljni problem.

Osrednje težišče raziskav, predstavljenih v tej disertaciji, je vrednotenje urejenosti značilk. Obravnavamo predvsem tri vidike njihovega vrednotenja. Prvi vidik je definiranje in kvantificiranje resnične urejenosti značilk (ang. ground truth ranking). Rezultat tega dela raziskav je kontrolirano okolje, ki ga lahko uporabimo za vrednotenje urejenosti značilk. Drugi vidik je novo razvita metoda vrednotenja urejenosti značilk, ki je sposobna oceniti in primerjati kakovost seznamov urejenih značilk dobljenih z različnimi metodami. Tretji vidik pa sta identifikacija praktičnih scenarijev, kjer vrednotenje urejenosti značilk potrebujemo, ter demonstracija koristnosti predlagane metode.

Začnemo s formalno opredelitvijo naloge urejanja značilk. Na osnovi definicij navedenih v literaturi o izbiranju značilk (ang. feature selection) opredelimo nalogo urejanja značilk kot reševanje problema izbiranja vseh relevantnih značilk in določitve njihove pravilne urejenosti. Rezultat je seznam urejenih značilk.

Urejanje značilk je pogojeno z definicijo pomembnosti. Naredimo izčrpen pregled različnih definicij in mer pomembnosti značilk opisanih v literaturi. Pri tem ugotovimo, da je skupna pomanjkljivost vseh mer pomembnosti značilk v tem, da značilke obravnavajo kot med seboj neodvisne. Da bi to pomanjkljivost presegli, predlagamo mero pomembnosti značilk z upoštevanjem interakcij med značilkami, kot jih lahko določimo na osnovi informacijske teorije. Z uporabo te definicije pomembnosti lahko rekonstruiramo resnično urejenost značilk in tako vzpostavimo umetno nadzorovano okolje za vrednotenje urejenosti značilk.

V drugem koraku predlagamo novo metodo za vrednotenje urejenosti značilk. V intuitivnem smislu metoda temelji na uporabnosti urejanja značilk kot filtrirne metode (ang. filter method) pri izbiranju značilk. Gre za formaliziran algoritemski postopek, ki temelji na postopni gradnji podmnožic značilk in učenju napovednih modelov na teh podmnožicah. Rezultat metode vrednotenja so t.i. krivulje napak (ang. error curves), ki prikazujejo kako (ali na kakšen način) so posamezne značilke razporejene znotraj danega seznama značilk. Omenjene krivulje lahko nadalje uporabimo za primerjalno vrednotenje kakovosti različnih metod urejanja značilk.

Metodo vrednotenja urejenosti testiramo v nadzorovani situaciji, ki temelji na omenjeni definiciji resnične urejenosti značilk. V poskusih znani resnični urejenosti značilk dodajamo različne nivoje šuma ter jih nato vrednotimo s predlagano metodo. Preizkusili smo tako enakomerno kot spremenljivo dodajanje šuma. Rezultati kažejo, da metoda uspešno zazna padec kvalitete urejenosti značilk pri uporabi višjih nivojev šuma.

Zadnji prispevek disertacije zajema tri empirične študije, ki se navezujejo na različne domene. Prva študija raziskuje vedenje različnih algoritmov za urejanje značilk, tako na sintetičnih kot na realnih podatkih. Rezultati kažejo, da je metoda ReliefF najboljša pri

urejanju značilk v sintetičnih podatkih. Pri realnih podatkih ni bilo nobene statistično pomembne razlike v kvaliteti urejenosti seznamov zgrajenih z metodo ReliefF in metodo računanja informacijskega prispevka (ang. information gain).

Druga študija preučuje uporabnost ansamblov seznamov urejenih značilk (ang. feature ranking ensembles – FREs). Cilj analize je bilo ugotoviti pod kakšnimi pogoji dobimo s pomočjo ansamblov sezname boljše kakovosti kot z metodami za gradnjo posameznih seznamov urejenih značilk. Rezultati kažejo, da ansambli dosegajo boljše rezultate le, če uporabljamo nestabilne osnovne metode za urejanje značilk, kot je na primer metoda naključnih gozdov.

Tretja študija sega na področje raziskav rakavih bolezni, oz. natančneje, na področje embrionalnih tumorjev (ET). Cilj analize je bil določiti ključne gene, ki so povezani z agresivnostjo tumorja. Naša metoda je v tej nalogi pomagala prepoznati metodo za urejanje značilk, ki je ustvarila najbolje urejen seznam potencialnih genov. Rezultati so bili podkrepljeni tudi s poznejšo rekonstrukcijo genskih mrež za gene, ki se nahajajo na vrhu urejenega seznama potencialnih genov.

Disertacijo zaključujemo s predstavitvijo možnosti nadaljnjega dela na tem področju. Prva in najpreprostejša razširitev se dotika področja izvedenega eksperimentalnega vrednotenja. V našem delu smo namreč upoštevali zgolj klasifikacijske in regresijske ciljne koncepte, metodologijo pa bi lahko razširili tudi na strukturirane ciljne spremenljivke. Druga možna smer razvoja tega dela bi lahko bila kombinacija več vidikov urejenosti značilk, vključno npr. s stabilnostjo seznamov urejenih značilk. Končno bi lahko pri raziskavah, poleg napovedne točnosti v okviru napovednega modeliranja, upoštevali tudi druge uporabnostne vidike urejenih seznamov značilk. Kot primer bi lahko služila zadnja empirična študija, v kateri so bile urejene značilke uporabljene za rekonstrukcijo genskih mrež.

Abbreviations

ECA	=	error curve average
ET	=	embryonal tumours
FFA	=	forward feature addition
FR	=	feature ranking
FRE	=	feature ranking ensembles
FS	=	feature selection
GT	=	ground truth
IG	=	information gain
RF	=	random forests
RFA	=	reverse feature addition
SVM-RFE	=	support vector machines - recursive feature elimination

1 Introduction

In this chapter we provide an introductory discussion on the work presented in this thesis. We begin by outlining the wider research area and the general context of this thesis. We then specify the research problems we are considering. After this is established, we present the motivation for this work and clearly state its scientific contributions. At the end of the chapter, we describe the overall organisation of the remainder of this thesis.

1.1 General Perspective

The work in this thesis can be categorised into the general research area of computer science. More specifically, it falls within the scope of artificial intelligence (McCarthy et al., 1955). Artificial intelligence is the science and engineering of making intelligent machines, especially intelligent computer programs.

Machine learning plays a central role in artificial intelligence, building also on contributions from areas such as statistics and information theory. It studies computer programs that automatically improve with experience (Mitchell, 1997). Another similar definition is given by Langley (1996), where learning is defined as the improvement of performance in some environment through the acquisition of knowledge resulting from experience in that environment.

Both definitions are based on learning from some kind of experience. If this experience is in the form of data examples (instances), this is called inductive learning (Bratko, 2000). In the most classical setting, an instance consists of several attribute-value pairs. In other words, an instance is a set of features (attributes) with a specific value for each feature.

A feature of specific interest is called a target concept, or more generally a target feature. If the objective of a learning task is to obtain models that predict the value of the target feature from the learning examples, then the task is called predictive learning (modelling) or supervised learning (Hastie et al., 2003). The target feature can be either discrete or continuous. Depending on the type of the target feature, the learning task can be classification (discrete target) or regression (continuous target).

The features contained in the dataset can be divided into several groups according to their relation to the target and the other features present in the data (Kononenko and Kukar, 2007). They can be random features that are unrelated to the target. They can be redundant features, whose information about the target is already contained in other feature(s). Correlated features are a type of redundant features, whose information is partially contained in other features. Also, it is possible to have features that are strongly interdependent, but only when considered w.r.t. the target concept. When considered individually, these features are independent from the target.

The type of features present in the data can have a large impact on the quality of the predictive models induced by a learning method (Liu and Motoda, 2008). With respect to the learning task, features can either be irrelevant or relevant. Irrelevant features do not influence or even have detrimental effect on the learning task. Random features are obviously irrelevant features. Redundant and correlated features are not strictly irrelevant because they contain information about the target. However, if they are eliminated from

the data the quality of the predictive model is unchanged. Such features are referred to as weakly relevant features (John et al., 1994). Relevant features help improve the quality of the produced predictive models and John et al. (1994) refers to them as strongly relevant features.

The task of finding and eliminating the irrelevant and/or redundant features in a dataset is called feature selection (Liu and Motoda, 2008). It is a task that usually precedes the learning of predictive models and has the goal to improve the induced model quality. If we view the process of learning as search through a space of hypotheses that explain the target concept, then the reduction of the number of features constrains this search space. Assuming that the eliminated features are irrelevant or redundant, this constraint allows the learning algorithm to focus on a search space that is more likely to contain better hypotheses, i.e., induce better predictive models.

Feature selection methods are usually divided into filter, wrapper and embedded methods (Guyon and Elisseeff, 2003). Filter methods provide a set of features to be used by the learning algorithm and are independent on the type of the learning algorithm used. Wrapper methods are actually coupled with a learning method that is used to select the optimal set of features. Embedded methods have the feature selection incorporated in the model learning phase.

Filter methods usually approach their task via a related machine learning task, namely, the task feature ranking. First, features of the dataset are ranked according to some notion of feature relevance. Then feature selection is performed by using a threshold on the number of top-ranked features, to be used by the learning method for inducing the predictive model.

The work presented in this thesis is related to the machine learning task of feature ranking, formulated within the supervised learning setting. The process of feature ranking is based on the evaluation of the relevance of individual features with respect to a target. This feature relevance can be used to establish a ranking order of the features (Duch, 2006), thus providing a *feature ranking*. The purpose of the work presented in this thesis is related to the evaluation of the correctness of the ordering of features within a feature ranking.

1.2 Motivation and Goals

Feature ranking has been primarily considered in the context of filter methods for feature selection. With the advent of high-throughput technologies in biology (Slonim, 2002), the curse of dimensionality has become an ever present problem when working with biological data. Feature selection techniques have become more and more important and the need for research in this area has increased (Saeys et al., 2007).

Not only has feature selection become more important, but the variety of applications including high-throughput data has also led to feature ranking gaining more importance as task separate from feature selection. These applications include the discovery of biomarkers for a disease, identifying drug targets, or genes that are involved in disease mechanisms. As noted by He and Yu (2010), one of the main problems with the analysis of the high-throughput data is the instability of the results. Namely, different datasets from the same problem area produce different feature rankings by using the same ranking method. Different subsamples of the same dataset also tend to produce different feature rankings. Finally, different ranking methods produce different feature rankings.

He and Yu (2010) identify three possible sources of this instability. The first is the inherent instability of some of the feature ranking algorithms. The second is the existence of multiple correlated features (genes) which are present in high-throughput biological data. The last source is the curse of dimensionality, namely the large number of features as compared to the number of instances in a dataset.

Independent of the cause of instability, when we have multiple different feature rankings, the question at hand is which feature ranking should be used for subsequent data analysis.

This leads to the more general issue of evaluating and comparing different feature rankings. Feature selection methods are evaluated in the context of their utility, namely selecting features used for inducing better predictive models. For feature ranking methods, this evaluation is less clear. To the best of our knowledge, there is no generally accepted method (in the literature) for evaluating feature rankings.

Our work focuses on this underexplored area of evaluating feature rankings. The main research goal of this thesis is to propose, analyse and apply an evaluation method for feature rankings. In order to achieve this main goal, we adress several more specific goals.

The first goal is to provide definitions of the different aspects of the machine learning task of feature ranking. The purpose of this is to clearly establish the context in which feature rankings will be evaluated. This further helps to determine the experimental setting in which the evaluation method for feature rankings will be analysed.

The second goal is the development of a feature ranking evaluation method. We propose an intuitive method, which evaluates feature rankings with the help of predictive models. The purpose of this evaluation method is to provide a qualitative and quantitative comparison between feature rankings of different quality.

Our third goal is to show that the evaluation method for feature rankings can be used for distinguishing between feature rankings of different quality. For this a controlled experimental setting is needed where the method will be analysed. The results of these experiments should demonstrate the usefulness of the feature ranking evaluation method.

Finally, the fourth goal of this thesis is to present different domains where the feature ranking evaluation method can be applied. The purpose of this is to show practical real-world scenarios where evaluating feature rankings is a relevant task. These applications should involve diverse data and show how evaluating feature rankings can aid in answering different research questions.

All of these goals have been addressed by the different contributions of this thesis. We give a summary of each of the individual contributions and relate them to the specific goals in the following text.

1.3 Contributions

Here, we present an overview of the contributions of this thesis. We organise this presentation according to the goals stated above (in Section 1.2).

The first group of contributions relates to the goal of providing an overview of different aspects of the task of feature ranking. We begin by providing surveys of different definitions in the literature about what is feature ranking. These are all given in the context of feature selection. Within this context we define that the purpose of feature ranking is to determine and provide a correct ordering of all of the relevant features present in the data. As output, the feature ranking task should provide a linearly ordered vector of features.

As feature ranking is based on a notion of relevance with respect to a target concept, we also review the various definitions of feature relevance. We provide both a survey of the axiomatic definitions of feature relevance, as well as the feature relevance indices used for quantifying it. We build upon these surveys and use a feature relevance measure based on feature interactions to define the ground truth ranking of features. This in turn helps us with the experimental work in this thesis, by providing a controlled experimental setting for evaluating feature rankings.

The second group of contributions is related to the method for evaluating feature rankings that we propose. It is an algorithmic procedure that uses a stepwise filter approach to evaluate feature rankings. Assuming a feature ranking is given as input, the method learns multiple predictive models in order to estimate the distribution of relevant features within a feature ranking. This allows us to distinguish between feature rankings of different quality, as better feature rankings would have all of the relevant features placed at the top

of the ranking. The advantages of the method are that it is intuitive and can provide both qualitative and quantitative comparison between different feature rankings.

The third group of contributions is related to demonstrating the usefulness of the proposed feature ranking evaluation method. First, the feature relevance measure based on feature interactions is used to define ground truth rankings. Next, the proposed evaluation method for feature rankings is analysed within this controlled setting. The analysis is based on uniform and non-uniform addition of noise to the known ground truth ranking. This produces noisy rankings with worsening quality. We use our feature ranking evaluation method for comparing these noisy ground truth rankings with different levels of noise. The outcome of the analysis shows that the method can clearly distinguish between the noisy feature rankings of different (worsening) quality.

The last group of contributions concerns several application scenarios for the proposed feature ranking evaluation method. Three contributions were made by addressing three different practical problems. The first is the comparative analysis of the behaviour of various feature ranking algorithms. We analyse different feature ranking approaches, with our evaluation method on both synthetic and real-world data. Although limited in many ways, the synthetic setting provides us with some conclusions about the ability of different feature ranking methods to detect feature interactions and properly rank the interacting features. The real-world data we considered includes low-dimensional UCI datasets from various domains as well as high-dimensional microarray gene expression datasets.

The second practical contribution is the investigation of the usefulness of feature ranking ensembles (FREs), which aims to establish the conditions under which constructing an aggregated feature ranking has a positive impact on feature ranking quality. For evaluating FREs, we used the synthetic setting in which the ground truth ranking was known. We used our feature ranking evaluation method to compare the individual and the aggregated feature rankings obtained with the use of FREs. The results and conclusion were in line with the previous theoretical and empirical studies on feature ranking ensembles. Ensemble rankings are more stable, but otherwise of approximately the same quality as individual rankings.

The final practical contribution is the use of the feature ranking evaluation method as a part of a biological knowledge discovery scenario. The analysis involved microarray gene expression data for embryonal tumours (ETs) and aimed at discovering genes related to tumour aggressiveness in embryonal tumours (ETs). In this context, we used our evaluation method for feature rankings to identify the feature ranking method that produces the best ranked list of genes. With the support of stability analysis, part of the top ranking genes were chosen and used for constructing gene networks based on previous biological knowledge.

1.4 Organization

We organise this thesis in several chapters. Chapter 2 presents in detail the context of this thesis. It includes a survey of definitions of feature ranking, means to detect and quantify feature relevance, as well as a detailed overview of several widely used feature ranking algorithms. Also, an overview of the different aspects of evaluating feature ranking algorithms is given.

A feature relevance measure, based on information theory, is described in Chapter 3. Its advantage over other feature relevance indices is that, when evaluating the relevance of a feature, it takes into account the whole context, i.e., considers feature interactions. This measure also allows for the formal definition of a synthetic, controlled setting, with a known ground truth ranking, useful for performing experimental evaluations of feature rankings.

The main contribution of this thesis, the feature ranking evaluation method, is presented in Chapter 4. We first give the basic intuition behind the evaluation method and then its formal algorithmic definition. We include a detailed explanation of the output that the

method provides and explain how it can be interpreted and used to compare different feature rankings.

The first set of experiments with our feature ranking evaluation method is presented in Chapter 5. The goal of these experiments is to empirically demonstrate that the feature ranking evaluation method can distinguish between rankings of different quality. Also, it contains experiments that investigate different feature ranking algorithms and their properties.

The second set of experiments is presented in Chapter 6. The experiments concern the application of different feature ranking methods and our feature ranking evaluation method to various datasets from different domains, including benchmark and practically relevant datasets. First,the usefulness of feature ranking ensembles is analysed. Next, a comparative analysis of different feature ranking algorithms, on datasets from various domains (medicine, biology, ecology, etc.) is performed. Finally, a set of experiments is performed in the domain of cancer research, more concerned with discovering key genes involved in embryonal tumour aggressiveness.

We finish the thesis by presenting our conclusions in Chapter 7. We present a detailed overview of each of the contributions of this thesis, based on the previously described methodology and the results of the experiments. Finally, we provide several ideas and directions for further work.

2 Background and Related Work

In this chapter, we present the background and the related work of this thesis. We organise the text into three separate parts, each related to feature ranking from a different aspect. We begin by defining feature ranking in Section 2.1. We also contrast and compare between the machine learning tasks of feature ranking and feature selection.

We then proceed with an overview of the different definitions of feature relevance in Section 2.2. This basics provide means to define and quantify feature relevance that is in the essence of any feature ranking process. We then provide, in Section 2.3, an overview of machine learning approaches used for inducing feature rankings from data, as well as ensemble methods for feature rankings in Section 2.3.4.

At the end, we focus on the evaluation of feature ranking approaches. We first consider one specific aspect of feature ranking, namely stability of feature ranking methods in Section 2.4. We also present an overview of how different feature ranking algorithms have been evaluated so far, in Section 2.5.

2.1 Feature Ranking and Feature Selection

In principle, feature ranking and feature selection are two different tasks. Feature ranking provides an ordered set of features according to a feature relevance function, while feature selection provides a feature subset optimal for model induction (Guyon and Elisseeff, 2003). In practise however, the two are closely related. They are both prepossessing steps before model induction and are usually considered together under the name of feature selection.

Feature selection encompasses three types of methods: filter methods, wrappers and embedded methods (Guyon and Elisseeff, 2003). Feature ranking is usually used within filter methods for feature selection. Filter methods act as a preprocessing step in the knowledge discovery process and are used to filter out irrelevant features before model induction occurs (Blum and Langley, 1997).

An alternative view of feature ranking within the context of feature selection is given by Molina et al. (2002). Feature selection algorithms are classified according to the output they yield:

- algorithms giving a (weighted) linear order of features (FR)
- algorithms giving a subset of the original features (FS)

Feature ranking algorithms are of the first type, as they provide a linear order of features. According to this definition, feature ranking subsumes feature selection, if we consider that FS algorithms provide binary weighted features.

In (Nilsson et al., 2007) feature selection algorithms are divided according to the type of feature selection problem that they solve. Namely, two types of feature selection problems are defined:

- **minimal-optimal**, finding a minimal feature set optimal for classification (FS)
- **all-relevant**, finding all features relevant to the target variable (FR)

By this definition, feature ranking can be thought of as a feature selection task, namely the *all-relevant* feature selection problem.

From all the definitions provided, it can be seen that there is a distinction in the literature between feature ranking and feature selection. However, there is an overlap in the utility, or the purpose of performing feature selection and/or feature ranking. They both are used as a step preceding model induction aiming to improve the performance of the model. Therefore, feature ranking is treated in the context of feature selection.

In our work, we also take into account this definitions. Namely, we consider that the main requirement of a feature ranking algorithm is to solve the all-relevant problem and in addition, provide as output a linear ordering of features. However, when evaluating the feature ranking methods, we additionally evaluate how well-ordered the produced rankings are.

2.2 Feature Relevance

Determining the relevance of a feature represents the basis for the process of feature ranking. Depending on the task at hand, there are various ways to define the relevance of a feature F_i. In machine learning literature, several surveys provide an overview of the definitions of feature relevance. The most notable are given by John et al. (1994), by Blum and Langley (1997) and, more recently, by Molina et al. (2002).

In a supervised setting, there is always a target concept and feature relevance is always estimated with respect to this target, F_t. Formally, Blum and Langley (1997) define this as:

Definition 1. *A feature F_i is* **relevant to a target concept** *F_t if there exsists a pair of examples A and B in the instance space such that A and B differ only in their assignment to F_i and $F_t(A) \neq F_t(B)$.*

However, this definition has several drawbacks. The first is that it looks at the relevance of the feature isolated from the context of the other features. Namely, a feature F_i might be relevant for a target concept F_t, only in context of another feature(s), for example related to a feature F_j via an XOR relation. The other drawback is feature redundancy. If we assume that the dataset contains a feature F_i', that is an exact copy of F_i, then by this definition neither F_i nor F_i' would be considered relevant.

To remedy this drawbacks, John et al. (1994) introduce two notions of feature relevance, namely *strong* and *weak* feature relevance. Both of the definitions are based on conditional probabilities and take into account the context of a feature F_i, i.e., the other features present in the dataset. If we let $\mathscr{F}_{S_i}$ be the set of all features of the dataset except F_i, then strong relevance is defined as:

Definition 2. *F_i is* **strongly relevant** *iff there exist feature values f_i, f_t and $\langle f_{S_i} \rangle$ for which $P(F_i = f_i, \mathscr{F}_{S_i} = \langle f_{S_i} \rangle) > 0$ such that:*

$$P(F_t = f_t | F_i = f_i, \mathscr{F}_{S_i} = \langle f_{S_i} \rangle) \neq P(F_t = f_t | \mathscr{F}_{S_i} = \langle f_{S_i} \rangle)$$

.

Intuitively, a strongly relevant feature can't be removed from the dataset without loss of information about the target concept. Based on this definition, the less constrained version of relevance is defined, namely weak relevance:

Definition 3. *A feature F_i is* **weakly relevant** *iff it is not strongly relevant, and there exists a subset of features $\mathscr{F}'_{S_i} \subset \mathscr{F}_{S_i}$, for which there exist some f_i, f_t and $\langle f'_{S_i} \rangle$ for which $P(F_i = f_i, \mathscr{F}'_{S_i} = \langle f'_{S_i} \rangle) > 0$ such that:*

$$P(F_t = f_t | F_i = f_i, \mathscr{F}'_{S_i} = \langle f'_{S_i} \rangle) \neq P(F_t = f_t | \mathscr{F}'_{S_i} = \langle f'_{S_i} \rangle)$$

.

In contrast to strongly relevant features, a weak relevant feature can be removed from the dataset without loss of information. However, there is a subset of features $\mathscr{F}'_{S_i}$ of the set of all features in the data, for which the feature becomes strongly relevant.

If we additionally introduce a learner L, another definition of feature relevance can be formulated, termed *incremental usefulness* (Blum and Langley, 1997) or simply *usefulness* (Caruana and Freitag, 1994).

Definition 4. *Given a sample of data S, a learning algorithm L and a feature set $\mathscr{F}_S$, feature F_i is* **incrementally useful** *to L w.r.t. $\mathscr{F}_S$, if the accuracy of the hypothesis that L produces using the feature set $\{F_i\} \cup \mathscr{F}_S$ is better than the accuracy achieved using just the feature set $\mathscr{F}_S$.*

This definition of relevance is more suited for selecting important features for a feature subset selection task than for determining the relevance of an individual feature. As noted by Caruana and Freitag (1994), all useful features are relevant, but not all relevant features are useful.

All of the definitions stated so far give just qualitative distinction between relevant and irrelevant features. There are also approaches that consider feature relevance quantitatively and aim to calculate the amount of relevance of features. One definition is based on entropy and mutual information, given by (Bell and Wang, 2000) and it is the following:

Definition 5. *Given three sets of variables $\mathscr{F}_\mathscr{X}$,$\mathscr{F}_\mathscr{Y}$ and $\mathscr{F}_\mathscr{Z}$ with a joint probability distribution P, let $I(\mathscr{F}_\mathscr{X};\mathscr{F}_\mathscr{Y}|\mathscr{F}_\mathscr{Z})$ be the mutual information between $\mathscr{F}_\mathscr{X}$ and $\mathscr{F}_\mathscr{Y}$ given $\mathscr{F}_\mathscr{Z}$, and let $H(\mathscr{F}_\mathscr{X}|\mathscr{F}_\mathscr{Y})$ be the entropy of $\mathscr{F}_\mathscr{X}$ given $\mathscr{F}_\mathscr{Y}$. If $H(\mathscr{F}_\mathscr{Y}|\mathscr{F}_\mathscr{Z}) \neq 0$, then the feature relevance of $\mathscr{F}_\mathscr{X}$ to $\mathscr{F}_\mathscr{Y}$ given $\mathscr{F}_\mathscr{Z}$, denoted $r_p(\mathscr{F}_\mathscr{X};\mathscr{F}_\mathscr{Y}|\mathscr{F}_\mathscr{Z})$, is defined as:*

$$r_p(\mathscr{F}_\mathscr{X};\mathscr{F}_\mathscr{Y}|\mathscr{F}_\mathscr{Z}) = \frac{I(\mathscr{F}_\mathscr{X};\mathscr{F}_\mathscr{Y}|\mathscr{F}_\mathscr{Z})}{H(\mathscr{F}_\mathscr{Y}|\mathscr{F}_\mathscr{Z})} = \frac{H(\mathscr{F}_\mathscr{Y}|\mathscr{F}_\mathscr{Z}) - H(\mathscr{F}_\mathscr{Y}|\mathscr{F}_\mathscr{X},\mathscr{F}_\mathscr{Z})}{H(\mathscr{F}_\mathscr{Y}|\mathscr{F}_\mathscr{Z})}$$

.

If $H(\mathscr{F}_\mathscr{Y}|\mathscr{F}_\mathscr{Z}) = 0$, then $r_p(\mathscr{F}_\mathscr{X};\mathscr{F}_\mathscr{Y}|\mathscr{F}_\mathscr{Z}) = 0$

This is a general definition of relevance of sets of features that also includes conditioning on the feature set $\mathscr{F}_\mathscr{Z}$ and is therefore referred to as *conditional relevance* (Bell and Wang, 2000). If this conditioning on $\mathscr{F}_\mathscr{Z}$ is dropped then the so-called *unconditional relevance* (Bell and Wang, 2000) is obtained.

This unconditional feature set relevance can be instantiated to include just a single feature F_i and a single target concept F_t. We refer to this instantiation as *entropic relevance* of a feature (Molina et al., 2002), given by equation:

$$r_p(F_i;F_t) = \frac{I(F_i,F_t)}{H(F_t)} = \frac{H(F_t) - H(F_t|F_i)}{H(F_t)} \tag{1}$$

Intuitively, this equation quantifies the relevance of the feature F_i as the relative reduction of entropy of F_t due to the knowledge of F_i.

An extensive overview of approaches for calculating feature relevance are given in Duch (2006). They are considered in the context of filter methods for feature selection and are grouped into three types of approaches. They are the *correlation* based, the *information theory* based and the approaches based on *distances between distributions.*

The first approach, correlation based relevance, is presented as the simplest way to estimate feature relevance. Unlike the other approaches, it does not rely on probability density estimation. Representative examples of this type of approaches for estimating feature relevance include the χ^2 statistic (Kenney and Keeping, 1951), the t-test (Gosset, 1908), Pearsons correlation coefficient (Pearson, 1896–1912).

If the probability distributions of the features can be estimated, then the relevance can be calculated as the distance between the probability distribution of the feature and the

probability distribution of the target. The distance measures that can be used include the Kolmogorov measure (Kolmogorov, 1965), the Kullback-Leibler divergence (Kullback and Leibler, 1951), the Jeffreys-Matusita distance (Jeffreys (1946), Matusita (1951)), the Vajda entropy (Vajda, 1979) and the value difference metric (VDM) (Stanfill and Waltz, 1986).

The various relevance measures based on information theory are also related to the estimation of the probability distributions. The most basic measures, mutual information and information gain are based on calculating the entropy of features. There are various modifications of these measures and they include information gain ratio, entropy distance, Mantaras distance and symmetrical uncertainty coefficient. There are also information theory relevance measures that are not based on entropy, such as average weight of evidence (Michie, 1990) and the minimum description length (MDL) measure (Rissanen, 1978).

This concludes the overview of the various approaches to define and quantify feature relevance. Both from the axiomatic definitions and the relevance measures presented it can be concluded that there are a multitude of perspectives from which feature relevance can be considered. According to the axiomatic definitions of relevance, features can vary between weakly/strongly relevant or irrelevant, while in the presence of a learner L they can be incrementally useful. Quantifying the amount of relevance of a feature w.r.t. a target is also quite diverse. The different feature relevance indices can be calculated as simple correlations, as distances between distributions, or as various information-theoretic based measures. The choice of a relevance measure is based on the specific application at hand.

2.3 Feature Ranking Algorithms

Feature ranking is a process based on determining feature relevance. In Section 2.2, we have provided an overview of the various ways to define and quantify feature relevance. An extensive overview of feature relevance indices has been given by Duch (2006), who divided them into correlation based, information-theory based and indices based on distances between distributions.

In this section, we consider several algorithmic approaches to determining feature ranking (relevance) which are widely used in data mining applications. Namely, we take a look at the Relief family of algorithms (Robnik-Šikonja and Kononenko, 2003) in Section 2.3.1, random forests (Breiman, 2001) in Section 2.3.2 and SVM-RFE (Guyon et al., 2002) in Section 2.3.3. In Section 2.3.4, we discuss feature ranking ensembles as a way of inducing feature rankings.

2.3.1 ReliefF and RReliefF

The approaches from the Relief family of algorithms for evaluating feature relevance are widely used. The original Relief algorithm was proposed by Kira and Rendell (1992) and is limited to two-class classification problems. The algorithm was extended by Kononenko (1994) to deal with multi-class problems. The extension was named ReliefF. Later, it was also adapted for regression problems (Robnik-Šikonja and Kononenko, 1997) and named RReliefF.

The basic intuition behind the Relief algorithms is to estimate the relevance of a feature according to how well it distinguishes between neighbouring instances. If the feature has different values for neighbouring instances that are of different class (nearest miss), then it is awarded a higher relevance values. However, if the values of the class for the neighbouring instances are the same (nearest hit), then the relevance value is decreased. The final relevance value awarded by the Relief algorithm is an approximation of the following difference of probabilities (Kononenko, 1994):

$$W[F] = P(\text{diff. value of } F | \text{nearest inst. from diff. class}) -$$

$$P(\text{diff. value of } F | \text{nearest inst. from same class})$$

Algorithm 1 Pseudocode for the ReliefF algorithm, taken from Robnik-Šikonja and Kononenko (2003).

Input: for each training instance a vector of feature values and the class value
Output: the vector W of estimations of the relevance of features

set all weights $W[F] = 0$
for $i = 1$ to m **do**
  randomly select and instance R_i
  find k nearest hits H_j
  for each class $C \neq class(R_i)$ **do**
    from class C find k nearest misses $M_j(C)$
  end for
  for $F = 1$ to f **do**
    $W[F] = W[F] - \sum_{j=1}^{k} diff(F,R_i,H_j)/(m \cdot k) +$
    $\sum_{C \neq class(R_i)} [\frac{P(C)}{1-P(class(R_i))} \sum_{j=1}^{k} diff(F,R_i,M_j(C))]/(m \cdot k)$
  end for
end for

The pseudocode of the ReliefF algorithm is given in Algorithm 1. The algorithm begins by selecting a random instance (R_i) and finding the nearest hits and nearest misses. Nearest hits (H_j) are instances of the same class, while nearest misses (M_j) are instances of different class(es). The function $diff(F,I_1,I_2)$ gives the difference between values of the feature F, for two instances I_1 and I_2. For nominal values it is calculated as:

$$diff(F,I_1,I_2) = \begin{cases} 0; & value(F,I_1) = value(F,I_2) \\ 1; & \text{otherwise} \end{cases}$$

and for numerical features as:

$$diff(F,I_1,I_2) = \frac{|value(F,I_1) - value(F,I_2)|}{max(F) - min(F)}$$

The relevance of the features given by the vector $W[F]$ are then calculated as the difference between two sums of differences, namely $\sum diff(F,R_i,M_j)$ and $\sum diff(F,R_i,H_j)$. The first sum represents the average contribution of all of the misses (averaged across all of the classes) and the second sum the average contribution of all of the hits. A larger average difference for the misses means that the feature distinguishes well between different classes. The average difference for the hits is indicative of how much the feature is variable across instances of the same class: a large value is not favourable.

If the target of interest is continuous, we have a regression problem. RReliefF is a variant of the Relief algorithms adapted to deal with regression problems. The details of the algorithm are given in pseudocode form in Table 2. Without going into the details of the algorithm, we will explain the basic intuition.

The main problem with regression tasks is that the predicted value $\tau(\cdot)$ is continuous and the concept of hits and misses does not apply directly. Therefore, a type of probability, or a similarity measure between two instances is introduced. Recall that the purpose of the Relief algorithm was to approximate these two probabilities:

$$P_{diffF} = P(\text{diff. value of } F | \text{nearest inst. from diff. class})$$

$$P_{diffC} = P(\text{diff. value of } F | \text{nearest inst. from same class})$$

If we also take into account the notation:

$$P_{diffC|diffA}(\text{diff. prediction} | \text{diff. value of F and nearest instances})$$

Algorithm 2 Pseudocode for the RReliefF algorithm, taken from Robnik-Šikonja and Kononenko (2003).

Input: for each training instance a vector of feature values $\mathbf{x}$ and predicted value $\tau(\mathbf{x})$
Output: the vector W of estimations of the relevance of features

set all N_{dC},$N_{dF}[F]$,$N_{dC\&dF}[F]$,$W[F]$ to 0
for $i = 1$ to m **do**
  randomly select and instance R_i
  select k instances I_j nearest to R_i
  for $j = 1$ to m **do**
    $N_{dC} = N_{dC} + diff(\tau\cdot, R_i, I_j) \cdot d(i,j)$
    for $F = 1$ to f **do**
      $N_{dF}[F] = N_{dF}[F] + diff(F, R_i, I_j) \cdot d(i,j)$
      $N_{dC\&dF}[F] + diff(\tau(\cdot), R_I, I_j) \cdot diff(F, R_i, I_j) \cdot d(i,j)$
    end for
  end for
end for
for $F = 1$ to f **do**
  $W[F] = N_{dC\&dF}[F]/N_{dC} - (N_{dF}[F] - N_{dC\&dF}[F])/(m - N_{dC})$
end for

then by using the Bayes rule, we have:

$$W[F] = \frac{P_{diffC|diffA} P_{diffA}}{P_{diffC}} - \frac{(1 - P_{diffC|diffF}) P_{diffA}}{1 - P_{diffC}}$$

The probabilities are estimated from N_{dC}, $N_{dF}[F]$ and $N_{dC\&dF}[F]$, where each of them is calculated as described in Algorithm 2.

2.3.2 Random Forests

Random forests are ensemble methods for predictive modelling. They were originally proposed by Breiman (2001). Their use as feature ranking methods has been recently empirically studied by Verikas et al. (2011).

Definition 6. *A random forest is a classifier consisting of a collection of tree-structured classifiers* $\{h(\mathbf{x}, \Theta_k, k = 1, \ldots)\}$ *where the* $\{\Theta_k\}$ *are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input* $\mathbf{x}$.

The general definition for random forests given above was provided by Breiman (2001). This definition only specifies that the random forests should consist of decision trees $\{h(\mathbf{x}, \Theta_k, k = 1, \ldots)\}$ (Breiman et al., 1984) that are generated by a randomised tree learning algorithm taking into account some kind of a random component Θ_k. The random component can be introduced when constructing the decision trees, either by randomly selecting a split (Dietterich, 2000b) from the best k splits or by selecting the best split from a randomly selected k features at each split (Amit and Geman, 1997). Randomisation can also be introduced by sampling either the data instances or the feature space and constructing decision trees on each sample. Instances can be sampled by bootstrap sampling (Breiman, 1996), while the approach for sampling the feature space is referred as "random subspace" sampling (Ho, 1998).

However, the usual approach to constructing random forests is to first perform bootstrap sampling on the data and then build decision trees for each sample. The decision trees are constructed by considering the best split at each level, from a randomly selected feature

subset. Determining the relevance of features by using random forests is based on two principles: randomisation and out-of-bag error estimates.

The rationale is the following: first, each out-bag-sample is used to estimate the out-of-bag error Err_{OOB_j} of the corresponding decision tree. Next, the feature of interest F_i, has its value randomised in the out-of-bag samples. Then, this permuted out-of-bag sample is run through the nodes of the corresponding decision tree and this out-of-bag error Err_{ij} is recorded. The feature relevance is then calculated as the average change of error from all the decision trees in the forest:

$$\text{rel}(F_i, F_t) = \frac{1}{k} \cdot \sum_{j=1}^{k} \frac{Err_{ij} - Err_{OOB_j}}{Err_{OOB}} \tag{2}$$

where k is the number of bags (boostrap samples). The intuition behind this relevance estimation is that a feature F_i is relevant to the target F_t, then the out-of-bag error should increase after permuting the values of F_i. For completeness, we present the whole algorithm (Algorithm 3) for calculating feature relevance by using random forests as provided by Kocev (2011).

Algorithm 3 Pseudocode for calculating feature relevance by using Random Forests, adapted from Kocev (2011). E is the set of training examples, k is the number of trees in the forest, D is the number of descriptive variables and $f(D)$ is the size of the feature subset that is considered at each node during tree construction.

procedure Induce_RF($E, k, f(D)$) **returns** Forest F, Relevance Vector R

$F = \emptyset$
$R = \emptyset$
for $i = 1$ **to** k **do**
    E_i = Bootstrap_sample(E)
    $T_i = Tree(E_i, f(D))$
    $F = F \bigcup T_i$
    $E_{OOB} = E \setminus E_i$
    Update_Imp(E_{OOB}, T_i, R)
end for
R = Average(R, k)
return F, R

procedure Update_Imp(E_{OOB}, T, R)

Err_{OOB} = Evaluate(T, E_{OOB})
for $j = 1$ **to** D **do**
    E_j = Randomize(E_{OOB}, j)
    Err_j = Evaluate(T, E_j)
    $R_j = R_j + \frac{1}{k} \cdot \frac{Err_j - Err_{OOB}}{Err_{OOB}}$
end for
return

procedure Average(R, k)

$R^T = \emptyset$
for $l = 1$ **to** $size(R)$ **do**
    $R_l^T = R_l / k$
end for
return R^T

2.3.3 SVM-RFE

SVM-RFE is an algorithm which provides just a ranking of features, without providing an explicit relevance value. It was proposed by Guyon et al. (2002). It is an algorithm based on recursive feature elimination (RFE) and a weighting provided by SVMs (Boser et al. (1992); Vapnik (1998) and Cristianini and Shawe-Taylor (2010)). It is based on iterative training of SVMs and removal of features with the smallest calculated weight.

Recursive feature elimination (RFE) is basically a backward feature elimination procedure that Kohavi and John (1997) used for wrappers in feature selection. It is a procedure based on the following intuition, as provided by Guyon et al. (2002):

1. Train a classifier

2. Compute a ranking criterion for all of the features

Algorithm 4 Pseudocode for the SVM-RFE algorithm, taken from Guyon et al. (2002).

Input: Training examples
1: $\mathbf{X}_0 = [\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_k, \ldots \mathbf{x}_l]^T$
2: Class labels $\mathbf{y} = [y_1, y_2, \ldots y_k, \ldots y_l]^T$
Output: Feature ranked list $\mathbf{r}$
3: Subset of surviving features: $\mathbf{s} = [1, 2, \ldots n]$
4: Feature ranked list: $r = [\,]$
5: **for** $\mathbf{s} \neq [\,]$ **do**
6: Restrict training examples to good feature indices: $X = X_0(:, \mathbf{s})$
7: Train the classifier: $\alpha = \mathrm{SVM-train}(X, \mathbf{y})$
8: Compute the weight vector of dimension length($\mathbf{s}$): $\mathbf{w} = \sum_k \alpha_k y_k \mathbf{x}_k$
9: Compute the ranking criteria: $c_i = (w_i)^2$, for all i
10: Find the feature with the smallest ranking criterion: $f = \arg\min(\mathbf{c})$
11: Update feature ranked list: $\mathbf{r} = [\mathbf{s}(f), \mathbf{r}]$
12: Eliminate the feature with the smallest ranking criterion: $\mathbf{s} = \mathbf{s}$(1:f-1,f+1:length($\mathbf{s}$))
13: **end for**

3. Remove the feature with the smallest ranking criterion

Note that, in order to obtain a feature ranking, this procedure should be used with the elimination of one feature at a time. Also, this feature elimination is not useful if coupled with a correlation-based method. This is because the correlation ranking criterion is determined with information from a single feature and irrespective of the features eliminated with RFE, it would produce the same results.

For completeness, we present the SVM-RFE algorithm in Algorithm 4. It is basically a more technically detailed version of the previously described RFE procedure. We will refrain from discussing it in details.

2.3.4 Feature Ranking Ensembles

Feature ranking ensembles (FREs) are inspired by ensemble learning (Seni and Elder, 2010), where the purpose is to combine the predictions of multiple models in one ensemble prediction. The success of ensemble methods inspired an intuitive extension of its basic principle for combining multiple baseline feature rankings. There are both empirical (Saeys et al., 2008) and theoretical studies (Jong et al., 2004) of feature ranking ensembles.

The basic intuition for constructing FREs is performed in three steps:

- bootstrap samples of the original data are produced
- by using some feature ranking method, several base feature rankings are induced
- the base rankings are combined in a single consensus ranking by using some kind of an aggregation procedure

In Jong et al. (2004), the base rankings are induced by using the ROGER (ROC-based genetic learner) algorithm (Sebag et al., 2003). The empirical study in (Saeys et al., 2008) compares several different feature ranking methods like: symmetrical uncertainty (SU) (Press et al., 1992), ReliefF, SVM-RFE and random forests. For aggregating the rankings a linear weighted function is used, of the type:

$$\mathrm{R}_{agg} = \sum_{i=1}^{n} w_i \cdot \mathrm{R}_i \tag{3}$$

The theoretical evaluation of ensembles of feature rankings (Jong et al., 2004) has shown that, under certain assumptions, if the feature rankings are consistent, then the ensemble

of feature rankings will converge. The empirical study by Saeys et al. (2008) has shown that the major advantage in using FREs is the increased robustness and stability of the consensus ranking, rather than improved quality of the feature ranking.

2.4 Stability of Feature Rankings

An important aspect of feature ranking algorithms is their stability or, more specifically, the stability of the ranked feature lists that they produce. There are several studies that propose measures for stability and also investigate it empirically, the most notable being the ones by Guzmán-Martínez and Alaiz-Rodríguez (2011), Kalousis et al. (2007), and Jurman et al. (2008). The estimation of the stability of a feature ranking algorithm is intuitively similar to the analysis of stability of classification algorithms (Turney, 1995). It is based on first inducing feature rankings, with the same algorithm, from different training instances and then comparing of these produced ranked lists.

Depending on the application several options can be considered in the comparison between the ranked lists can be between (Guzmán-Martínez and Alaiz-Rodríguez, 2011). One can compare:

- full ranked lists,
- partial ranked lists (top-k features), and
- partial sublists (feature subsets).

There is a variety of dissimilarity measures that can be used for each of these three types of comparisons. For full and partial list comparisons, the Spearman rank correlation (Saeys et al. (2008) and Kalousis et al. (2007)) coefficient or the Canberra distance (Jurman et al., 2008) can be used. For comparing partial ranked sublists, a wide variety of measures can be used: the Jaccard distance (as considered by Saeys et al. (2008) and Kalousis et al. (2007)) an adaptation of the Tanimoto distance (Kalousis et al., 2007), Kuncheva's stability index (Kuncheva, 2007), the Relative Hamming Distance (Dunne et al., 2002), consistency measures (Somol and Novovicová, 2010), Dice-Sorense's index (Loscalzo et al., 2009), Ochiai's index (Zucknick et al., 2008) or the percentage of overlapping features (He and Yu, 2010). In addition, Guzmán-Martínez and Alaiz-Rodríguez (2011) propose a measure based on Jensen-Shannon divergence (Lin, 1991) that is applicable for all of the three types of comparison.

If we denote with $\mathbf{R} = \{\mathrm{R}_1, \mathrm{R}_2, \ldots, \mathrm{R}_N\}$, the set of all feature rankings induced from the different samples of dataset $\mathscr{D}$, then the stability index $S(\mathbf{R})$ can be calculated as:

$$S(\mathbf{R}) = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} S_M(\mathrm{R}_i, \mathrm{R}_j) \tag{4}$$

This equation basically calculates the stability index, as average pairwise similarities between each pair of the feature rankings R_i. The function S_M can be any of the dissimilarity measures described above.

In our experimental work, presented in Chapter 5 and Chapter 6, for estimating the stability of the feature ranking algorithms, we use the Canberra distance as discussed by Jurman et al. (2008). The Canberra distance (Lance and Williams, 1966) and (Lance and Williams, 1967) is a weighted distance metric that puts bigger emphasis on the stability of the top ranked features. If we have two ranked lists R_A and R_B, where $\mathrm{R}(i)$ gives the rank of feature F_i, then the Canberra distance can be calculated as:

$$Ca(\mathrm{R}_A, \mathrm{R}_B) = \sum_{i=1}^{n} \frac{|\mathrm{R}_A(i) - \mathrm{R}_B(i)|}{\mathrm{R}_A(i) + \mathrm{R}_B(i)} \tag{5}$$

where n is the number of features.

In order for the distance to be applicable to partial rankings with $k < n$, the following adaptation is proposed:

$$Ca^{(k+1)}(\mathrm{R}_A, \mathrm{R}_B) = \sum_{i=1}^{n} \frac{|min\{\mathrm{R}_A(i), k+1\} - min\{\mathrm{R}_B(i), k+1\}|}{min\{\mathrm{R}_A(i), k+1\} + min\{\mathrm{R}_B(i), k+1\}} \tag{6}$$

where $Ca^{k+1} = Ca$.

Additionally, Jurman et al. (2008) provide an analytical approximation of the expected Canberra distance between completely random rankings. The approximation is dependent only on the total number of features n and the size of the top-k subset that is considered. It is given by:

$$\hat{E}\{Ca^{(k+1)}\} = \frac{(k+1)(2n-k)}{n}\log(4) - \frac{2kn+3n-k-k^2}{n} \tag{7}$$

For complete lists. the approximation becomes:

$$\hat{E}\{Ca^{(k+1)}\} = (\log(4)-1)n + \log(4) - 2 \tag{8}$$

Finally, if the Canberra distance for partial rankings is normalised with the expected Canberra distance for each value of k, a normalised stability indicator is obtained, calculated as:

$$\hat{I} = \left\{ \left(k, \frac{S_k(\mathbf{R})}{\hat{E}\{Ca^{(k+1)}\}} \right) : 1 \leq k \leq n \right\} \tag{9}$$

With this adaptation, the stability indicator becomes independent of particular values for k and n, as it represents the relative change of distance between top-k lists w.r.t. the expected top-k distance. Therefore, the values of the stability index are directly comparable for datasets with different number of features n and between different top-k sets of genes.

2.5 Evaluating Feature Rankings

From the various definitions of feature ranking, presented in Section 2.1, it is evident that feature ranking has been mostly considered and defined in the context of feature selection. As feature selection is a process that precedes model induction, it is usually evaluated in the context of the quality of the induced model. The same applies for evaluating feature rankings, as they are mostly evaluated via their utility as filter methods.

There are at least two general ways in which a feature ranking method can be evaluated. The first is in a synthetic controlled environment, where the relevant features in a dataset are known. The second is in a real-world setting, on datasets from different domains, with unknown structure of relevant features. So far, a unified framework for evaluating feature ranking has not been proposed, but there are some general trends which become apparent.

In the setting where the relevant features are known, the evaluation of feature ranking algorithms is done mainly by evaluating their capability to delineate relevant from irrelevant features. Put differently, they are evaluated in terms of their ability to solve the all-relevant feature selection problem. For example, ReliefF is evaluated by Robnik-Šikonja and Kononenko (2003) on synthetic datasets by its *separability* and *usability*. Separability is the difference between the lowest estimated relevance of the relevant features and the highest relevance of the irrelevant features, given by: $s = W_{R_{worst}} - W_{I_{best}}$. Usability is the difference between the highest estimated relevance of the relevant features and the highest estimated relevance of the irrelevant features, given by: $u = W_{R_{best}} - W_{I_{best}}$.

Jong et al. (2004) propose a statistical validation model for evaluating feature ranking algorithms on synthetic data. The evaluation is based on a ROC curve setting. Namely,

each ranking algorithm is evaluated via the so-called ROC-FS curve. The curve shows the trade-off achieved by the algorithm, between assigning high (resp. low) ranks to relevant (resp. irrelevant) features.

When examining the performance of feature ranking algorithms on real-world data, their performance has been mostly regarded from two aspects. The first aspect is the of consistency of the produced ranked lists. In the literature, this is usually referred to as the stability of feature rankings and was discussed in more detail in Section 2.4. The second aspect concerns their usefulness as filter methods. More specifically, they are evaluated by the performance of the predictive models induced after filtering on the produced rankings has been applied.

The general intuition behind the latter evaluations is to first apply the feature ranking algorithm on the data and induce a feature ranking. Next, for a certain value of k, the top-k features are selected and the remaining features are filtered out. At the end, a predictive model is constructed just from the selected features and its performance is evaluated.

Guyon et al. (2002) compare the performance of the predictive model at different values for k. Then, the feature ranking algorithm is evaluated by taking into account just the value k of the best performing model. Similar to this comparison are the ones by Furlanello et al. (2003) and Paoli et al. (2005). All of the predictive model evaluations are used to construct an average testing error curve (ATE), which is used for determining an optimal value for k. Feature ranking, as provided by random forests (Verikas et al., 2011), pursues the same intuition for evaluation. A variation of this type of evaluation is provided in the empirical study of feature ranking ensembles by Saeys et al. (2008). There, for a fixed k, the performance of the model is combined with the stability estimate and a single feature ranking quality index is produced.

2.6 Discussion

In this chapter, we presented the background and related work of this thesis. First, in Section 2.1, we presented the different general definitions of feature ranking. They were all placed in the context of feature selection. Building on these definitions we will from now on consider that the main goal of feature ranking is to solve the all-relevant feature selection problem, but and additionally provide a correct ordering of the relevant features.

As the process of feature ranking in the supervised learning context is based on the relevance of a feature w.r.t. a target, in Section 2.2, we presented and discussed various definitions of feature relevance. The axiomatic definitions provide means to distinguish relevant from irrelevant features, while the various feature relevance indices provided means to quantify this relevance. However, as noted by (Duch, 2006) these feature relevance indices treat each feature from a dataset as independent. This means that the very important concept of feature interactions (Jakulin and Bratko, 2004) is being overlooked. We propose to remedy this insufficiency in Chapter 3 by devising a relevance index grounded in information theory and based on feature interactions.

We also present an overview of different feature ranking methods in Section 2.3 and different ways to measure feature ranking stability in Section 2.4.This provides the necessary understanding of the different feature ranking methods and the way of estimating their stability. This is later used for the analyses and applications presented in both Chapters 5 and 6.

Existing approaches for evaluating feature rankings are presented in Section 2.5. This overview provides the necessary related work for the main contribution of this thesis, namely the proposed method for evaluating feature rankings. The method is described in detail in Chapter 4. This is a method for evaluating feature rankings in real-world domains, but it is also applicable for synthetic data. It is based upon the basic intuition of evaluating feature ranking methods as filter methods, more specifically, on the stepwise evaluation of predictive

models and constructing error curves.

The original contribution of our work is that we formalise and extend the basic intuition of the filter evaluation for feature rankings. The formalisation results in a general algorithmic procedure that can be used to evaluate the feature rankings. The purpose of this algorithmic procedure is to estimate the distribution of relevant features within a feature ranking. This is achieved by extending the construction of error curves to include not just the top-k, but also the bottom-k ranked features. The output of the evaluation method is an numeric indicator that can be used to qualitatively and quantitatively answer the question of whether a feature ranking R_A is better than feature a ranking R_B.

3 Ground Truth Relevance and Ranking

In this chapter, we define the notion of a ground truth feature ranking. We begin by providing basic notations and definitions in Section 3.1. We also introduce basic concepts from probability and information theory in Section 3.2 and discuss the notion of feature interactions in Section 3.3. We use these definitions in Section 3.4 to define and quantify the ground truth relevance of features. At the end, in Section 3.5, we propose a way of generating synthetic datasets with known ground truth ranking. This is important for performing experimental comparisons of feature rankings, as it provides a controlled setting for the experiments.

3.1 Basics

In this section, we define some basic terminology and notations that will be used further in the thesis. We first define the basic input to any machine learning algorithm, the dataset. We then provide a definition for a feature relevance function. The feature relevance function is then used to define the basic unit of analysis of our evaluation method, namely, a feature ranking. Finally, we define a feature ranking algorithm that takes as input a dataset and, based on a certain definition of feature relevance, outputs a feature ranking.

Dataset: The basic form of input to a machine learning algorithm is called a dataset, $\mathscr{D}$. A dataset consist of data examples (instances), where each instance e_k contains specific values for the n features of the dataset: $e_k = (f_{k1}, f_{k2}, \ldots f_{kn})$. If we denote with E_i the domain of a feature F_i, then $\mathrm{E} = E_1 \times E_2 \times \ldots E_n$ represents the complete instance space. Each instance e_k is a point in this instance space and it should be noted that a dataset is always just a sample of the whole instance space.

In a supervised learning setting, one of the features is of particular interest. We will call this feature the "target feature" and denote it by F_t. We will denote the initial unordered set of features as $\mathscr{F} = \{F_1, F_2, \ldots, F_n\}$. We are interested in the ranking of features from $\mathscr{F}$ with respect to F_t.

Feature relevance: By using a relevance function $\mathrm{rel}_i = \mathrm{rel}_{\mathscr{F}}(F_i, F_t)$, each feature from $\mathscr{F}$ gets a relevance value assigned to it. The relevance value is assigned with respect to the target feature F_t. We define the relevance function as a mapping $rel : \mathscr{F} \to \mathscr{F} \times \mathbb{R}^+$.

Feature rankings: The previously determined feature relevance can be further used to infer a rank $rank(F_i)$ for each feature F_i. The rank imposes an ordering of the features from $\mathscr{F}$, thus providing a *feature ranking*. We denote the ordered feature set by $\mathrm{R} = (F_{r1}, \ldots, F_{rj}, \ldots, F_{rk})$, where $rank(F_{r1}) \leq \cdots \leq rank(F_{rj}) \leq \cdots \leq rank(F_{rk})$. In the remainder of this thesis, a feature ranking of this form will be considered as the basic object of evaluation of the method we propose.

Feature ranking algorithm: Finally, we define the notion of a feature ranking algorithm $r(\cdot)$. Such an algorithm takes as input a dataset $\mathscr{D}$ and, based on a certain definition of relevance rel_i, provides a feature ranking R:

$$r(\mathscr{D}) \to \mathrm{R}.$$

3.2 Feature Independence, Entropy and Information

In this section, we introduce some basic concepts from probability and information theory, which are further used for defining and quantifying feature interactions. A feature F_i can be considered as a random variable, for which we observed specific values f_i from the domain E_i, contained in the dataset. Based on these values, we can estimate the feature's probability distribution, which we denote by $P(F_i)$.

The expected information of a feature F_i can be determined based on the observed values by using *Shannon's entropy* (Shannon, 1948), defined as:

$$H(F_i) = -\sum_{f_i \in E_i} P(F_i = f_i) \log_2 P(F_i = f_i) = -\sum_{f_i} P(f_i) \log_2 P(f_i) \tag{10}$$

Or put differently, if we observe k values of a feature, the entropy is indicative of the average uncertainty of the values of that feature.

Let us now consider two arbitrary features F_i and F_j. If they are independent, i.e., not related to each other, then their joint probability distribution $P(F_i, F_j)$ equals

$$P(F_i, F_j) = P(F_i)P(F_j)$$

where $P(F_i)$ and $P(F_j)$, are the corresponding probability distributions of F_i and F_j.

If there is a dependence between F_i and F_j, then it can be expressed via the conditional probability:

$$P(F_i|F_j) = \frac{P(F_i, F_j)}{P(F_j)} = \frac{P(F_j|F_i)P(F_i)}{P(F_j)}$$

Based on this, we can also define conditional entropy as:

$$H(F_i|F_j) = -\sum_{f_i, f_j} P(f_i, f_j) \log_2 P(f_i|f_j) = -\sum_{f_i, f_j} P(f_i, f_j) \log_2 \frac{P(f_i, f_j)}{P(f_j)} = H(F_i, F_j) - H(F_j)$$

Intuitively, conditional entropy quantifies the uncertainty of F_i that remains after we gain knowledge of F_j.

The amount of information that two features F_i and F_j have about each other is measured with their *mutual information*, calculated as:

$$\mathrm{I}(F_i; F_j) = \sum_{f_i, f_j} P(f_i, f_j) \log_2 \frac{P(f_i, f_j)}{P(f_i)P(f_j)} = H(F_i) + H(F_j) - H(F_i, F_j).$$

Intuitively, mutual information measures the information that the two variables (features F_i and F_j) share.

These simplified definitions, considering just two features F_i and F_j, are meant to provide a background on the topic of feature interactions that is discussed in more detail in Section 3.3.

3.3 Interactions of Features

In the previous section, when laying out the basics from probability and information theory, we considered only two features, a arbitrary features from a dataset F_i and F_j. Instead of an arbitrary feature F_j, we now consider the target feature F_t. Even if a feature F_i is not informative about the target F_t directly, it may be informative when considered in conjunction with a set of other features, via feature interactions.

Feature interactions can be of different degree, i.e., involve a different number of features from $\mathscr{F}$. The simplest type of interaction is the association between an individual feature and the target F_t. This is called a two-way interaction (Jakulin and Bratko, 2004), as it

includes just an individual feature F_i and the target F_t. In general, the interactions can be of an arbitrary degree between 2 and $n+1$, where n is the number of features in $\mathscr{F}$.

Before formally defining feature interactions, we will first intuitively explain what is a feature interaction. In the most general sense, a feature interaction is defined as an *irreducible whole* (Jakulin and Bratko, 2004). Probabilistically speaking, if we consider an arbitrary feature set $\mathscr{F}_S = \{F_1, F_2, \ldots, F_k\}$, then the features in it are interacting if the joint probability model $P(\mathscr{F}_S)$ can't be reduced to a simpler model by factorisation to its marginal distributions (Jakulin and Bratko, 2004).

This probabilistic interpretation of feature interactions stems directly from the properties of the joint probability distribution of independent features. Namely, let us assume that a single feature from $\mathscr{F}_S$ is independent from all other features of $\mathscr{F}_S$ and denote it by F_{ind}. Then, for the joint probability distribution $P(\mathscr{F}_S)$, we have:

$$P(\mathscr{F}_S) = P(\mathscr{F}_S \backslash F_{ind}) P(F_{ind}).$$

This means that the joint probability $P(\mathscr{F}_S)$ is no longer an irreducible whole as it can be described as a product of two separate probability distributions, namely $P(\mathscr{F}_S \backslash F_{ind})$ and $P(F_{ind})$.

From here, quantifying a feature interaction simply becomes a task of comparing the joint probability distribution with a factorised distribution. In other words we compare two models: a model that assumes feature interactions and a model that assumes features are not interacting (Jakulin and Bratko, 2004). In order to quantify a feature interaction $\mathrm{int}(F_1; F_2; \ldots; F_k)$ between the features of $\mathscr{F}_S$, one needs to define the two models (interaction and non-interaction) and a loss function $\mathscr{L}(\cdot)$ used to compare them, namely:

$$\mathrm{int}(F_1; F_2; \ldots; F_k) = \mathscr{L}(P(\mathscr{F}_S), \hat{P}(\mathscr{F}_S)), \tag{11}$$

where $P(\mathscr{F}_S)$ is the interaction model and $\hat{P}(\mathscr{F}_S)$ is the non-interaction model.

One instantiation of the general function for measuring feature interactions $int(F_1; F_2; \ldots; F_k)$, is *interaction information*. It is an entropy based measure proposed by Jakulin and Bratko (2004), generalised from the formulae of McGill (1954). It can be used to quantify n-way interactions and is defined as:

$$\mathrm{I}(F_1; F_2; \ldots; F_k) = -\sum_{S \subseteq \mathscr{F}_S} (-1)^{|\mathscr{F}_S \backslash S|} H(S) \tag{12}$$

It is shown in (Jakulin and Bratko, 2004) that this interaction information is the *Kullback-Leibler (KL) divergence* (Kullback and Leibler, 1951) between the joint probability model and the *Kirkwood superposition approximation* (KSA) (Matsuda, 2000) of the joint probability model. In other words, in this instantiation of the interaction function, the non-interaction model is the KSA of the joint probability distribution and the loss function is the KL divergence:

$$\mathrm{int}(F_1; F_2; \ldots; F_k) = \mathrm{I}(F_1; F_2; \ldots; F_k) = D_{KL}(P(\mathscr{F}_S) || \hat{P}_{KSA}(\mathscr{F}_S)). \tag{13}$$

The KL divergence is a measure used to compare two probability distributions, for example $P(F_i)$ and $Q(F_i)$ for an arbitrary feature F_i. It is defined as:

$$D_{KL}(P(F_i) || Q(F_i)) = \sum_{f_i} P(f_i) \log_2 \frac{P(f_i)}{Q(f_i)}. \tag{14}$$

Note that this measure is not a metric and is not necessarily symmetric. The KSA is a factorised approximation of a joint probability distribution, defined as:

$$\hat{P}_{ksa}(\mathscr{F}_S) = \prod_{S \subset \mathscr{F}_S} P(S)^{(-1)^{1+|\mathscr{F}_S \backslash S|}}. \tag{15}$$

3.4 Ground Truth Relevance of Features

In light of the previous discussion on feature interactions, we now address a question of central importance for the process of feature ranking: "What is a good feature ranking?" . In the broadest sense, a feature F_i is relevant and therefore higher ranked, if it contains more information about the target feature F_t.

A feature F_i can be informative about a target feature F_t in different ways. The most straightforward way is the direct correlation of F_i to the target F_t. However, depending on the context, i.e., on the other features present in the dataset, the feature F_i can also be informative in conjunction with some of them. This context-dependent aspect of individual features is exactly what feature interactions represent.

More formally, a feature F_i for a given dataset $\mathscr{D}$, consisting of features $\mathscr{F}$, can be informative about the target F_t, either individually or in conjunction with any of the possible subsets of features $\mathscr{F}_S \subseteq \mathscr{F} \setminus F_i$. Based on this rationale, we define the output of a feature relevance function, as follows:

$$\mathrm{rel}_{\mathscr{F}}(F_i, F_t) = \underset{\mathscr{F}_S \subseteq \mathscr{F} \setminus F_i}{\mathrm{agg}} \{\mathrm{int}(\{\mathscr{F}_S \cup F_i\}; F_t)\} \tag{16}$$

Essentially, feature relevance aggregates the values ($\mathrm{agg}\{\cdot\}$) of the interaction function $\mathrm{int}(\cdot)$, for a given feature in the context of different feature subsets $\mathscr{F} \setminus F_i$. The interaction function measures the interaction magnitude between a feature set $\mathscr{F}_S \cup \{F_i\}$ and the target F_t. The aggregation operator summarises the interaction magnitudes over all the different subsets $\mathscr{F}_S \cup \{F_i\}$. The basic intuition is that the relevance value of a feature F_i is related to the magnitude of the interaction information of F_i w.r.t. the target F_t. However, the aggregation provides larger relevance values if the feature is involved in multiple sets of features $\mathscr{F}_S \subseteq \mathscr{F} \setminus F_i$, which are interacting with the target. In light of the discussion about feature relevance from Section 2.2, this definition of the relevance function would yield larger relevance values for strongly relevant features than for weakly relevant ones.

With the help of Equation 12 and considering that the interaction is strictly with regard to the target F_t, we can calculate the interaction information for an arbitrary feature set $\mathscr{F}_S$ as:

$$\mathrm{I}(\mathscr{F}_S; F_t) = H(\mathscr{F}_S) + H(F_t) - H(\mathscr{F}_S, F_t). \tag{17}$$

Moreover, in our context, for calculating the relevance of a feature F_i, we are interested in the information that F_i has about the target F_t in conjunction with $\mathscr{F}_S$. Thus, the previous equation becomes:

$$\mathrm{I}(\mathscr{F}_S, F_i; F_t) = H(\mathscr{F}_S, F_i) + H(F_t) - H(\mathscr{F}_S, F_i, F_t) \tag{18}$$

It should be noted that this equation provides the information that the feature subset $\mathscr{F}_S \cup \{F_i\}$ has about the target F_t, although the information might originate from any subset $S \subset \mathscr{F}_S \cup \{F_i\}$. This means that the value provided by $\mathrm{I}(\mathscr{F}_S, F_i; F_t)$ can originate from the feature F_i, from the feature subset $\mathscr{F}_S$, or any combination of F_i with subsets of $\mathscr{F}_S$. Therefore, when defining the final feature relevance function, we need to take this consideration into account.

For a given dataset $\mathscr{D}$, consisting of features $\mathscr{F}$ and considering Equations 16 and 18, we can define one specific instantiation of the feature relevance function as:

$$\mathrm{rel}(\mathrm{F_i}, \mathrm{F_t}) = \sum_{\mathscr{F}_S \subseteq \mathscr{F} \setminus F_i} [\mathrm{I}(\mathscr{F}_S, F_i; F_t) - \sum_{S \subset \mathscr{F}_S} (-1)^{|\mathscr{F}_S \setminus S| - 1} \mathrm{I}(S, F_i; F_t)] = \sum_{\mathscr{F}_S \subseteq \mathscr{F} \setminus F_i} \sum_{S \subseteq \mathscr{F}_S} (-1)^{|\mathscr{F}_S \setminus S|} \mathrm{I}(S, F_i; F_t) \tag{19}$$

where

$$\mathrm{int}(\mathscr{F}_S, F_i; F_t) = \sum_{S \subseteq \mathscr{F}_S} (-1)^{|\mathscr{F}_S \setminus S|} \mathrm{I}(S, F_i; F_t)$$

The intuition behind the $\mathbf{int}(\cdot)$ function is that first, the interaction magnitude $\mathrm{I}(\mathscr{F}_S, F_i; F_t)$ of the feature set $\mathscr{F}_S \cup \{F_i\}$ is determined. Then, in order to be certain that this information originates directly from $\mathscr{F}_S \cup \{F_i\}$, the sum of the interaction magnitudes $\mathrm{I}(S; F_t)$ of all subsets $S \subset \mathscr{F}_S \cup \{F_i\}$ is calculated and subtracted from $\mathrm{I}(\mathscr{F}_S, F_i; F_t)$. We note, that this definition of feature relevance is very similar to the one provided by Štrumbelj et al. (2009) under the name of interaction contributions, where it is used in the context of an instance explanation method called IME (Interactions-based Method for Explanation).

In the case of redundant features present in $\mathscr{F}_S$, the value of $\mathbf{int}(\mathscr{F}_S, F_i; F_t)$ can be negative. There are two ways to approach the possible negative values. First, they can directly be included in the sum when calculating the feature relevance. Namely, the rationale is that the redundancy of features should be penalised and that redundant features should have lower overall relevance than non-redundant ones. The other way to approach this is to define $\mathbf{int}(\cdot)$ as:

$$\mathbf{int}(\mathscr{F}_S, F_i; F_t) = 0; \text{ if } \mathrm{I}(\mathscr{F}_S, F_i; F_t) < \sum_{S \subset \mathscr{F}_S} (-1)^{|\mathscr{F}_S \setminus S|-1} \mathrm{I}(S, F_i; F_t)$$

.

In this way, $\mathbf{int}(\cdot)$ can only be positive or equal to zero. The intuition here is that the relevance value should only reward the informativeness of a feature F_i and not penalise redundancy.

For our synthetic data experiments in Chapter 5, we use this last definition of $\mathbf{int}(\cdot)$ for generating the underlying ground truth feature relevance. Considering Equation 19, inferring the ground truth feature ranking R_{GT} would require determining the interaction information of all possible subsets of $\mathscr{F}$. Assuming a dataset $\mathscr{D}$ consisting of n features, the number of calculations necessary for determining the ground truth ranking would equal to:

$$|\mathscr{P}(\mathscr{F})| - 1 = 2^{|\mathscr{F}|} - 1 = 2^n - 1$$

where $\mathscr{P}(\mathscr{F})$ is the powerset of $\mathscr{F}$. The number of calculations is equal to the cardinality of the powerset reduced by one, because the interaction information of the empty set should not taken into account. This means that calculating the ground truth relevance of features entails heavy computational costs that are intractable, especially when working in high-dimensional domains. This corresponds to the conclusions by Nilsson et al. (2007), where finding all of the relevant features in a dataset (solving the all-relevant FS problem) is proven to be an intractable task.

3.5 Generating Synthetic Datasets with Known Ground Truth Ranking

After defining the ground truth relevance of features, we discuss the reverse process, namely, that of generating synthetic datasets for a given ground truth ranking. This is important mostly for comparative studies evaluating ranking methods, where experiments are performed which require a known ground truth ranking. The ground truth relevance function, in theory, provides a unique mapping from the features to relevance values. However, the opposite is not true. Namely, a set of ground truth relevance values and the corresponding ranking can be valid for an infinite number of datasets.

From the standpoint of the definition in Section 3.4 and Equation 16, the ground truth relevance/ranking depends on the feature interaction structure present in the synthetic data. Therefore, when generating the synthetic data, it is important to clearly specify the interaction structure. By specifying the feature interaction structure, the ground truth relevance is implicitly specified, and the relevance values $\mathbf{rel}_{GT,i} = \mathbf{rel}_{GT}(F_i, F_t)$, for each feature F_i, are calculated.

We can specify the feature interaction structure as a collection of feature interaction sets $\{\mathscr{F}_{int,1}, \ldots, \mathscr{F}_{int,p}\}$. Each feature interaction set, $\mathscr{F}_{int}$, consists of a set of features,

$\{F_i, \ldots, F_j\}$. Its relation to the target F_t is described by the interaction function $F_t = f(\mathscr{F}_{int})$ and the probability with the target $P(\mathscr{F}_{int}, F_t)$.

A feature F_i can be a member of several feature interacting sets. For inducing the feature relevance and from that the ground truth ranking, an instantiation of Equation 16 is used. Although Equation 16 requires all possible subsets of $\mathscr{F}$ to be considered, their number in this case is limited and is quite small in practical applications. Namely, only the specified feature interacting sets that a feature F_i belongs to are used to calculate the relevance value. The contribution of the other feature subsets is assumed to be zero.

For calculating the relevance values, knowing the actual interaction function, $F_t = f(\mathscr{F}_{int})$, is not necessary. However, it needs to be precisely specified for the process of generating the actual feature values in the resulting dataset.

4 Evaluation Method for Feature Rankings

In this chapter, we present the evaluation methodology for feature rankings, which is the main subject of this thesis. We first provide, in Section 4.1, a discussion on the main object of our investigation, namely, the feature ranking. We next provide the basic intuition behind the evaluation method for feature rankings in Section 4.2. The details of the evaluation methodology are presented in Section 4.3. There, we discuss the construction of the so-called *error curves*, which are at the core of our evaluation method. The guidelines for visualisation and interpretation of the error curves are given in Section 4.4, while their quantitative (numeric) comparison is discussed in Section 4.5. We end this chapter with Section 4.6, where we discuss the construction of an expected error curve, which can be used as a baseline for comparing various feature ranking approaches.

4.1 Feature Rankings and Feature Ranking Methods

The input to the evaluation method we are proposing is a feature ranking R. According to the definition in Section 3.1, the feature ranking is defined as an ordered feature set:

$$\mathrm{R} = (F_{r1}, \ldots, F_{rj}, \ldots, F_{rk})$$

where $rank(F_{r1}) \leq \cdots \leq rank(F_{rj}) \leq \cdots \leq rank(F_{rk})$.

The feature ranking itself will typically be the output from a feature ranking method r, applied to given data $\mathscr{D}$:

$$r(\mathscr{D}) \rightarrow \mathrm{R}.$$

In general, the purpose of feature ranking algorithms (Nilsson et al., 2007) is to solve the all-relevant feature selection problem. However, we argue that delineating relevant from irrelevant features would be the minimum requirement for a feature ranking algorithm. It should also provide an estimation of the relative importance of one feature w.r.t. the other features, from which a proper feature ranking can be induced.

An ideal feature ranking algorithm should produce the ground truth ranking R_{GT}. Although the ground truth ranking exists, in reality, the ranking methods provide only an approximation of it. More precisely, a ranking method provides a permutation R of the ground truth ranking, R_{GT}, that can be otherwise viewed as a noisy version of the ground truth. The amount of noise in the ranking R with respect to ground truth ranking, is related to the amount of randomness in it. This can be measured by the number of random permutations of the ground truth ranks of features and the proximity of those permutations.

A good feature ranking method would produce a ranking R that is well ordered. This means that the more relevant features would have a higher rank, i.e., all of them are concentrated at the beginning of the feature ranking. In contrast, if the feature ranking produces a random ranking, R_{rand}, the relevant features would be uniformly distributed in the ranking. Estimating and comparing this distribution of relevant features is the intuition on which we base our evaluation approach.

Table 1: Comparison between the expected number of relevant features $E[n_{rel,k}]$ for the top-k (bottom-k) ranking subsets of the GT ranking R_{GT} and a random ranking R_{rand}.

(a) top-k features

	top-k features		
	R_{GT}		R_{rand}
k	$\leq n_{rel}$	$> n_{rel}$	all k
$\mathbf{E[n_{rel,k}]}$	k	n_{rel}	$k \cdot n_{rel}/n_{tot}$

(b) bottom-k features

	bottom-k features		
	R_{GT}		R_{rand}
k	$\leq n_{irr}$	$> n_{irr}$	all k
$\mathbf{E[n_{rel,k}]}$	0	$k - n_{irr}$	$k \cdot n_{rel}/n_{tot}$

4.2 Expected Number of Relevant Features

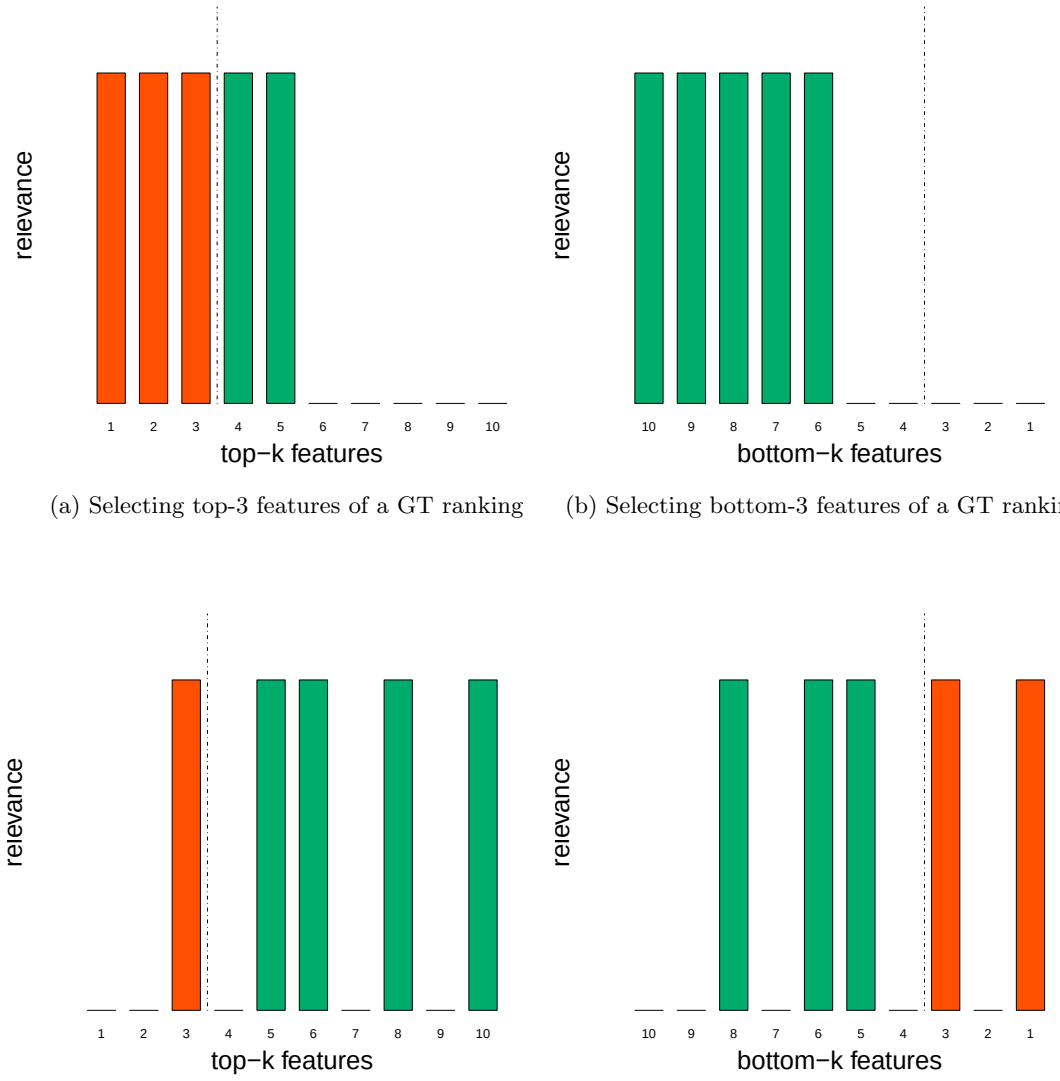
Before proceeding to the detailed description of the proposed evaluation method for feature rankings in Section 4.3, we first explain the rationale behind it and provide the basic mathematical intuition. We start with an illustrative example of this setting given in Figure 1. We consider a toy dataset $\mathscr{D}$ consisting of 10 features in total. Half of the features are relevant features $n_{rel} = 5$ and half are irrelevant features $n_{irr} = 5$. The distribution of the relevant features in the GT ranking is given in Figures 1a and 1b, while the distribution of the relevant features in the random ranking can be seen in Figures 1c and 1d. The difference between the distribution of the relevant features within the two rankings is clearly visible: for the GT ranking, all of the relevant features are concentrated at the beginning of the ranking, while for the random ranking they are uniformly distributed.

We can compare how the relevant features are distributed within the ranking starting either from the top or from the bottom of the ranking. If we first consider a specific point $k = 3$, from the top of the ranking, by comparing Figure 1a and Figure 1c, we will notice that the top-3 features of the GT ranking contain more relevant features ($n_{rel,k} = 3$) than the top-3 features of the random ranking ($n_{rel,k} = 1$).

The opposite is true if we start from the bottom of the rankings and compare the bottom-3 features in Figure 1b and Figure 1d. The number of relevant features present in the bottom-3 of the GT feature ranking ($n_{rel,k} = 0$) is smaller than the number of relevant features present in the bottom-3 of the random ranking ($n_{rel,k} = 2$). From this illustrative example, it is also visible that the relation between the top-k (bottom-k) features of the GT and the random ranking can be extended to an arbitrary k.

In order to systematically investigate this claim, we now consider a more general setting of a dataset $\mathscr{D}$, which consists of a total of n_{tot} features. A portion of these features n_{rel} can be considered relevant, while the remaining $n_{irr} = n_{tot} - n_{rel}$, represent irrelevant features. We also consider the comparison of two feature rankings produced from this arbitrary dataset, namely: the ground truth ranking R_{GT} and a random R_{rand}. The same rule applies, that R_{GT} has all the relevant features at the top of the ranking, while R_{rand} has a uniform distribution of the relevant features. Our goal is to determine and compare the expected number of relevant features $E[n_{rel,k}]$, which is present in the top-k (bottom-k) feature subsets provided by the GT ranking R_{GT} and a random ranking R_{rand}.

We first discuss the expected number of relevant features $E[n_{rel,k}]$ of the ground truth ranking R_{GT}. If we consider the top-k features of R_{GT} and take into account that R_{GT} has all of the relevant features at the top of the ranking, then we have two intervals of k to consider. First, the interval where k is smaller or equal to the number of the relevant features n_{rel}.



(a) Selecting top-3 features of a GT ranking (b) Selecting bottom-3 features of a GT ranking

(c) Selecting top-3 features of a random ranking (d) Selecting bottom-3 features of a random ranking

Figure 1: Comparison of the number of relevant features $n_{rel,k}$ selected by considering the top-k and the bottom-k features of the GT ranking R_{GT} and a random ranking R_{rand}. The bars indicate the relevance of a feature. For a fixed value of $k = 3$, the orange bars indicate the relevant features that have been included by considering the top-3 (bottom-3) features from the rankings R_{GT} and R_{rand}. The green bars indicate the features that have not been included in the top-3 (bottom-3) feature subsets.

In this case, all of the top-k features of R_{GT} would be relevant. Formally, for $k \leq n_{rel}$, we have $E[n_{rel,k}] = k$. The other interval is when k is larger than the number of the relevant features n_{rel}. Then, the expected number of relevant features in the top-k subset of R_{GT} is always equal to the number of relevant features n_{rel} in the dataset, i.e. if $k > n_{rel}$ then $E[n_{rel,k}] = n_{rel}$. For the bottom-k features of the GT ranking R_{GT}, we again consider two intervals. Considering that all of the relevant features are at the top of the GT ranking, while $k \leq n_{irr}$ the expected number of relevant bottom-k features $E[n_{rel,k}]$ is 0. If $k > n_{irr}$ then the expected number of relevant features would be $E[n_{rel,k}] = k - n_{irr}$.

Now, if we consider the random ranking R_{rand}, the expected number of relevant features $E[n_{rel,k}]$ in the top-k (bottom-k) feature set behaves like the expected value of a random

variable following a hypergeometric distribution. Namely, if we consider selecting the top-k (bottom-k) features from a random ranking, as drawing a sample of size k without replacement from n_{tot} features, and treat selecting a relevant feature as a positive outcome, then the expected value of selecting a relevant feature is $E[n_{rel,k}] = k \cdot n_{rel}/n_{tot}$. This discussion about the expected number of relevant features in the top-k (bottom-k) subsets, given the feature rankings $\mathbf{R}_{GT}$ and $\mathbf{R}_{rand}$, is summarised in Table 1.

If we compare the results from Table 1 for the GT ranking $\mathbf{R}_{GT}$ and the random ranking $\mathbf{R}_{rand}$, we can conclude that the previous claim from the illustrative example from Figure 1 can be indeed generalised for an arbitrary k. Namely, if we first compare the results from Table 1a for the top-k features, the expected number of relevant features present in the top-k subset from $\mathbf{R}_{GT}$ is either k or n_{rel}. This expected number is always larger than the expected number of relevant features present in the top-k subset from $\mathbf{R}_{rand}$ because $k \geq k \cdot n_{rel}/n_{tot}$ and $n_{rel} \geq n_{rel} \cdot k/n_{tot}$.

If we now compare the results from Table 1b, the expected number of relevant features present in the bottom-k subset from $\mathbf{R}_{GT}$ is either 0 or $k - n_{irr}$. These values are always smaller than the expected number of relevant features present in the bottom-k subset from $\mathbf{R}_{rand}$ because $0 < k \cdot n_{rel}/n_{tot}$ and $k - n_{irr} < k \cdot n_{rel}/n_{tot} = k - n_{irr} \cdot k/n_{tot}$.

In the discussion so far, we compared the two extremes of feature rankings, namely the ground truth ranking and a random ranking. For an arbitrary feature ranking algorithm, we would expect it to produce feature rankings $\mathbf{R}$, which are better than a random ranking and worse than or equal to a GT ranking. Therefore, the same expectation applies for the expected values of relevant features present in the top-k (bottom-k) feature subset of $\mathbf{R}$. A desired property of the feature ranking is to have the expected number of relevant features closer to that of the GT ranking and further away from that of a random ranking.

4.3 Stepwise Construction of Error Curves

In this section, we discuss in detail our method for evaluating feature rankings. It is based on the previous discussion about the expected number of relevant features present in top-k (bottom-k) feature subsets of a given ranking $\mathbf{R}$. However, in real case scenarios we do not know which features are relevant and our goal is to estimate how they are distributed across a feature ranking, but without explicitly discerning interaction structures present in the dataset.

For an arbitrary feature set $\mathscr{F}_S$, we can assess if it contains relevant features, by constructing and evaluating predictive models $\mathscr{M}(\mathscr{F}_S, F_t)$. This evaluation of the predictive models, provides a cumulative measure of the information contained by the features present in the subsets and it is quantified by an error measure $Err(\mathscr{M}(\mathscr{F}_S, F_t))$.

The output of our evaluation method will be based on error estimates of predictive models, built for specific feature subsets induced by the feature ranking. The question is, how to generate the feature subsets from the feature ranking, so that the error estimates can provide insight into the correctness of the feature ranking and constitute an evaluation thereof.

The construction of the feature subsets should be guided by the feature ranking $\mathbf{R}$. For a given feature $F_{ri} \in \mathbf{R}$, we can only say if it is more or less relevant than other feature(s) F_{rj}. Starting from the top ranked feature F_{r1} and going towards the bottom ranked feature F_{rn}, the feature relevance should decrease. Following this, one way to construct the feature subsets is to start with the highest ranked feature F_{r1} and to construct larger subsets by stepwise adding of lower-ranked features from the ranking $\mathbf{R}$. For this method of feature subset construction, we use the term *forward feature addition* (FFA).

More formally, forward feature addition (FFA) constructs the initial feature set with the highest ranked feature $\mathbf{R}_{11} = \{F_{r1}\}$, from the feature ranking $\mathbf{R}$. Each next set $\mathbf{R}_{1(i+1)} = \{F_{r1}, \ldots, F_{r(i+1)}\}$, is constructed by adding the next lower-ranked feature, namely $\mathbf{R}_{1(i+1)} =$

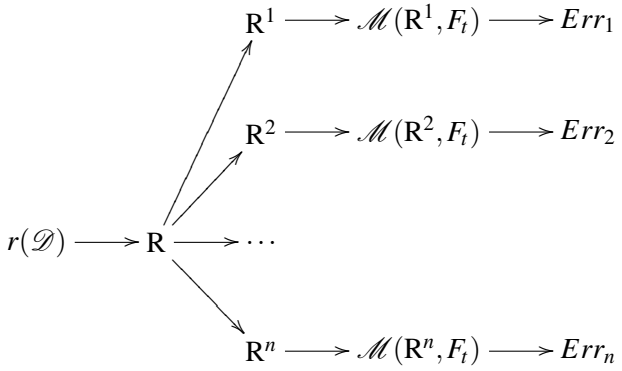


Figure 2: Graphical representation of the feature ranking evaluation method. A previously induced feature ranking R is used to construct feature subsets R^j of different cardinality. The feature sets are constructed either from the top-k or the bottom-k ranked features. Each feature set is used to learn a predictive model $\mathscr{M}(\mathrm{R}^j, F_t)$, which is evaluated and yield an error estimate Err_j.

$\mathrm{R}_{1i} \cup \{F_{r(i+1)}\}$, for $i = 1 \ldots n$. The process proceeds until R_{1n} contains all of the features from $\mathscr{F}$.

Also, the reverse direction for constructing the feature subsets is valid. Namely, starting with the lowest ranked feature F_{rn} and progressively adding higher-ranked features all the way to F_{r1}. This method for feature subset construction is named *reverse feature addition* (RFA).

More specifically, the initial feature set constructed by RFA contains only the lowest ranked feature, namely $\mathrm{R}_{nn} = \{Frn\}$. The following set $\mathrm{R}_{in} = \{F_{ri}, \ldots, F_{rn}\}$, is constructed by adding the next higher-ranked feature, namely $\mathrm{R}_{(i-1)n} = \mathrm{R}_{in} \cup \{F_{r(i-1)}\}$, for $i = n \ldots 1$. Similarly like for FFA, the process of RFA continues until $\mathrm{R}_{in} = \mathscr{F}$.

For each constructed feature subset R_S, which is based on the feature ranking R, we build a predictive model $\mathscr{M}(\mathrm{R}_S, F_t)$. For each constructed model, we evaluate its prediction error. Considering that the feature subset are generated in a stepwise fashion by two methods for feature construction, this results into two error curves. We name them FFA and RFA curves, each constructed by the corresponding FFA/RFA feature construction method. Examples can be seen in Figure 3a and Figure 3b.

The value for each point of the FFA curve can be defined as follows: $FFA(\mathrm{R}_{1i}) = Err(\mathscr{M}(\mathrm{R}_{1i}, F_t))$, where $\mathrm{R}_{1i} = \{F_{r1}, \ldots, F_{ri}\}$ and $i = 1 \ldots n$.

In a similar manner, the RFA curve can be defined as: $RFA(\mathrm{R}_{in}) = Err(\mathscr{M}(\mathrm{R}_{in}, F_t))$, where $\mathrm{R}_{in} = \{F_{ri}, \ldots, F_{rn}\}$ and $i = n \ldots 1$.

The whole process of FFA/RFA curve construction is summarised in algorithmic form in Table 5. In the algorithm, FFA/RFA curves referred under the common name of error curves. The only point of difference between them is in the function for generating the next feature to be added to the existing feature set, namely the $feature(\mathrm{R}, i)$ function.

The computational complexity $\mathscr{O}$ of the proposed algorithm, for constructing a single curve (FFA or RFA), depends linearly to the number of features. More specifically it is $\mathscr{O}(n \cdot (M + T))$, where n is the number of features, M is the cost of constructing the predictive model and T is the cost of evaluating the model. It should be noted that M and T are dependent on the specific learning method used for inducing the model and on the procedure used for evaluating it.

4.4 Visualisation and Interpretation of Error Curves

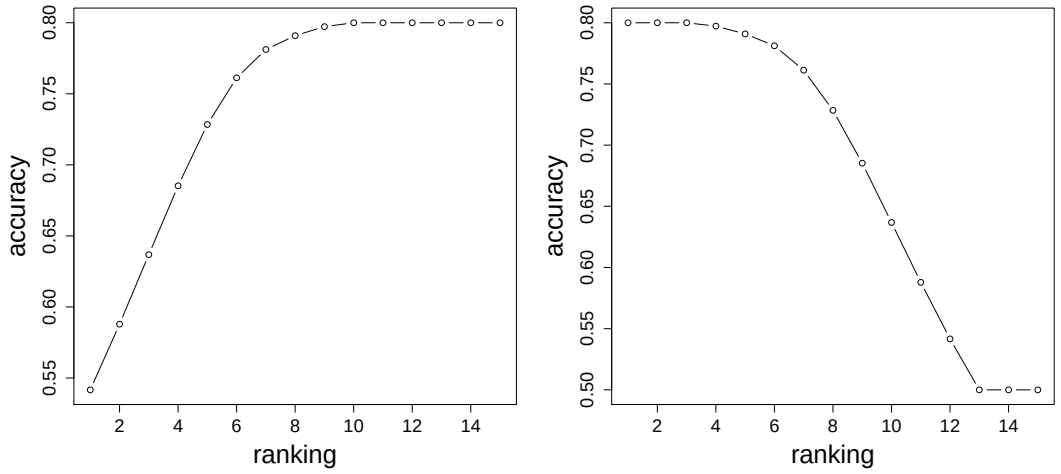
In this section, we discuss the visualisation and interpretation of FFA and RFA curves. We present a sample FFA and RFA curves in Figure 3. The y-axis for both curves, is the same

Algorithm 5 The algorithm for generating FFA and RFA curves.

Input: Feature Ranking, $\mathrm{R} = \{F_{r1}, \ldots, F_{rn}\}$; Target Feature, F_t
Output: Error Curve, Err, where $|Err| = n$

$\mathrm{R}_S \Leftarrow \emptyset$
for $i = 1$ to n **do**
  $\mathrm{R}_S \Leftarrow \mathrm{R}_S \cup feature(\mathrm{R}, i)$
  $Err[i] = Err(\mathscr{M}(\mathrm{R}_S, F_t))$
end for
  return Err

for FFA curve:
$feature(\mathrm{R}, i) = \{F_{ri}\}$
for RFA curve:
$feature(\mathrm{R}, i) = \{F_{r(n-i+1)}\}$



(a) An example of a FFA curve

(b) An example of a RFA curve

Figure 3: Examples of error curves

and depicts the error estimate of a feature subset. The x-axis represents the feature ranking and each point, i, has a different interpretation for the FFA and the RFA curve.

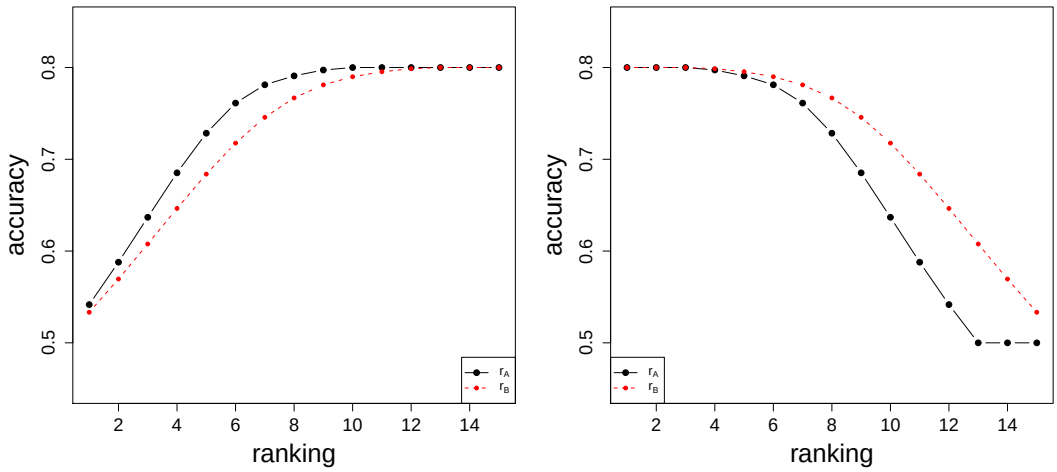
The FFA curve is always constructed from the top-k features, i.e., from the beginning of the ranking. Therefore, the error estimate at each point is made for the feature set $\mathrm{R}_{1i} = \{F_{r1}, \ldots, F_{ri}\}$. Or put differently, the FFA curve in Figure 3a, is constructed from left-to-right, as the top-ranked features are at the beginning of the ranking.

In contrast, the RFA curve is always constructed from the bottom-k features. The error estimate at each point, i, is for the feature set $\mathrm{R}_{in} = \{F_{ri}, \ldots, F_{rn}\}$. With respect to this, the RFA curve in Figure 3b, is constructed from right-to-left, starting with the end of the ranking and going towards the beginning of the x-axis.

If we first consider the FFA curve in Figure 3a, we can observe that as the number k, of features from the feature ranking R increases, the accuracy of the predictive models also increases. This can be interpreted as follows: By adding more and more of the top-k ranked features, the feature subsets constructed contain more relevant features, reflected in the improvement of the error measure.

For the RFA curve in Figure 3b, we can notice a slight difference at the beginning of the curve, which considering the previous discussion, is located at the end of the x-axis. Namely, the accuracy of the models constructed with the bottom ranked features is minimal, which means the ranking is correct in the sense that it puts irrelevant features at the bottom of the ranking. As the number of bottom-k features increases, some relevant features are included and the accuracy of the models slowly increases.

In summary, at each point k, the FFA curve gives us the expected error of the predictive



(a) Comparison of FFA curves

(b) Comparison of RFA curves

Figure 4: Comparison of FFA and RFA curves

models constructed with the top-k ranked features, while the RFA curve gives us the expected error of the bottom-k ranked features. In Section 4.5, we will discuss how to compare error curves constructed from different ranking methods.

4.5 Comparing of FFA and RFA Curves

In the previous section, we provided a basic intuition on how FFA and RFA curves are related to the distribution of relevant features through the feature ranking. However, the real utility of FFA/RFA curves can only be seen if we consider them in a relative, or more precisely, comparative context.

Let us consider two arbitrary feature ranking methods r_A and r_B, which produce feature rankings R_A and R_B, correspondingly. For these two methods rankings, we present the corresponding FFA/RFA curves in Figure 4. On the left-hand side of Figure 4, there are two FFA curves, while on the right-hand side two RFA curves are plotted.

If we first inspect the FFA curves visually, we find that the values of the FFA curve of the ranking method r_A, are (most of the time) above the curve of the other ranking method r_B. This can be interpreted in the following way: for arbitrary k, when considering top-k features of the feature rankings R_A and R_B, there are more relevant features included in the top-k features of ranking R_A than ranking R_B. This implies that ranking algorithm r_A produces a better ranking, as compared to ranking algorithm r_B.

A similar discussion applies for the RFA curve. When one considers the bottom-k features of a given feature ranking, most of the time, feature ranking R_A includes less relevant features than feature ranking R_B, i.e., the predictive models constructed are less accurate. Here, because the opposite logic of the FFA curve applies, one can also conclude that feature ranking algorithm r_A produces a better feature ranking than feature ranking algorithm r_B.

However, this is a mere visual inspection of the curves, which can only provide a qualitative intuition of which ranking method is better. For quantification purposes, it would be necessary to have a function which provides a cumulative assessment of the differences between the error estimates at different points of the curves. In the most general case, this would be an aggregation function of the weighted difference between two curves. For the

FFA curve we would have:

$$FFA_{diff} = \underset{i=1...n}{\text{agg}} \{w_i \times diff(FFA_{r_A}(i), FFA_{r_B}(i))\}. \quad (20)$$

While, for the RFA curve we would have:

$$RFA_{diff} = \underset{i=1...n}{\text{agg}} \{w_i \times diff(RFA_{r_A}(i), RFA_{r_B}(i))\}. \quad (21)$$

There are several sensible choices for instantiations of the aggregation function. The choice depends on the specific task at hand. Considering that we are comparing feature rankings, two aspects are important. The first is the position of most of the relevant features in the ranking. The second relates to the position of the "most" relevant features. In a comparative sense, the first aspect relates to the position of the FFA/RFA curves differences, while the second relates to the magnitude of these differences.

Differences between the FFA/RFA curves of two ranking methods at the beginning of the curves are more important than the differences at the end of the curves. Namely, if two FFA curves are different at the beginning, this means that one of the ranking method is not putting the most relevant features at the top of the ranking. Correspondingly, for the RFA curves, differences at the beginning of the curve (at the bottom of the ranking), means that one of the feature ranking method is giving low ranks to features which are relevant.

The second aspect is related more to the magnitude of the differences between the FFA/RFA curves than to their position. The intuition is that if "more" relevant features are misranked, then this is worse than "fewer" relevant features being misranked.

From a technical perspective, in order to emphasise the aspect of position, the weighting function from Equations 20 and 21, should be a function of the position, i, namely $w_i = f(i)$. In the same manner, in order to emphasise the aspect of magnitude, the weighting function should depend on the size of the difference, namely $w_i = f(diff_i)$. In addition, it is also possible to construct a weighting function that incorporates both the position and magnitude aspect, $w_i = f(i, diff_i)$.

For illustrative purposes, we define four instantiations of Equation 20 and Equation 21, which we use to calculate the difference between the FFA/RFA curves from Figure 4. We consider the following weighting functions, namely:

- $w_i = 1$, equal weight for all differences
- $w_i = f(i) = 1/|\mathrm{R}_i|$, weight inverse to feature subset size
- $w_i = f(diff_i) = diff_i$, weight proportional to the difference magnitude
- $w_i = f(i, diff_i) = diff_i/|\mathrm{R}_i|$, weight which includes both position and magnitude

The aggregation function used for summarising the differences (in all of the four instantiations) is the average, namely:

$$\underset{i=1...n}{\text{agg}}(\mathrm{w_i}, \mathrm{diff_i}) = \frac{\sum_{i=1}^{n} w_i \times diff_i}{\sum_{i=1}^{n} w_i} \quad (22)$$

The obtained values are given in Table 2. They are calculated for the FFA/RFA examples in Figure 4a and Figure 4b. The difference is calculated for r_A with respect to r_B. As seen in Table 2, the values for the FFA curves are positive, which can be interpreted as "r_A is better than r_B". While the values for the RFA curves are negative, the interpretation is the same: "ranking method r_A is better than ranking method r_B".

Table 2: Different quantitative comparisons of error curves

	$w_i = 1$	$w_i = 1/\|\mathrm{R}_i\|$	$w_i = diff_i$	$w_i = diff_i/\|\mathrm{R}_i\|$
FFA	0.018	0.019	0.032	0.03
RFA	-0.042	-0.054	-0.08	-0.077

4.6 Expected FFA and RFA Curves

In Section 3.4 we defined how one can obtain the ideal ranking given the data. Since the calculation of the ground truth ranking R_{GT} is in most cases intractable, it is not possible to compare a single ranking FFA/RFA curve to the one of the ground truth ranking.

The opposite to the ground truth ranking, is the random ranking. More precisely, there is an expected value that can be calculated for each point k of the FFA or RFA curves. The intuition behind the expected FFA/RFA curves is as follows: If we can not say how good a single ranking R is, maybe we can say how close to random it is.

The expected value of the FFA/RFA curves is dependent solely on the properties of the dataset under consideration and is the same for either the FFA or the RFA curve. Namely, at each point k, the expected value of the error measure, Err, is the average of the error estimations of all possible subsets $\mathscr{F}^k \in \mathscr{F}$, where for the cardinality of each subset we have $|\mathscr{F}_S^k| = k$. More precisely:

$$E[Err(M(\mathscr{F}^k, F_t))] = \frac{1}{\binom{n}{k}} \sum_{\mathscr{F}_S^k \subseteq \mathscr{F}, |\mathscr{F}_S^k| = k} Err(M(\mathscr{F}_S^k, F_t)) \tag{23}$$

According to Equation 23, calculating the expected value of the ranking is also intractable, especially for high-dimensional spaces as we have to consider an exponentially high number of feature subsets. However, for practical purposes this problem can be circumvented by sampling the space of possible feature subsets $\mathscr{P}_i(\mathscr{F})$, for each i.

Suppose we have somehow calculated or approximated the expected FFA/RFA curve. If we have a ranking algorithm r that produces a good (mostly correct) ranking, its FFA curve would be above the expected FFA curve. For the RFA curve, the oposite would apply and the algorithm's curve would be below the values of the expected RFA curve.

5 Experiments for Establishing the Evaluation Methodology

In this chapter, we present and discuss a large part of the experimental work performed within this thesis. The experiments are centred around the evaluation method presented in Chapter 4. Their aim is to demonstrate the usefulness of the proposed feature ranking evaluation method. In general, this is done by a comparative analysis of different error curves (FFA and RFA), constructed for different feature rankings of varying quality.

In order to provide a controlled environment for the experiments, we use synthetic datasets with known ground truth ranking. The process of generating the data and the ground truth ranking is described in Section 5.1. We provide proof-of-concept results in Section 5.2: these demonstrate the usefulness of the feature ranking evaluation method. In Section 5.3, we analyse the behaviour of the error curves of feature rankings produced by different feature ranking methods.

In all of our experiments, we use a single learning method to produce the predictive models used for generating the FFA and RFA curves. The choice of the learning method has been supported by an empirical study presented in Appendix A.

5.1 Generating Synthetic Data

The synthetic datasets used for our experiments provide a controlled environment in which specific feature ranking evaluation experiments can be performed. More precisely, the main advantage of having synthetic data is the possibility of clearly defining the ground truth (GT) ranking. In order to generate the data, as discussed in Section 3.5, we need to define the desired feature interaction structure of the dataset.

We define the structure by first defining the type of feature interactions sets, $\mathscr{F}_{int}$, that we are going to consider. First, for simplicity, we take both the features, F_i, and the target feature, F_t, to be binary. We also constrain the feature interaction sets, such that a feature, F_i, can belong to one and only one feature interaction set $\mathscr{F}_{int}$.

We then consider only feature interaction sets of cardinality one, $|\mathscr{F}_{int}| = 1$, and cardinality two, $|\mathscr{F}_{int}| = 2$. The feature sets with cardinality one are single features, F_i, that are in and individual interaction with the target, F_t. The interaction function for these feature sets is $f(\mathscr{F}_{int}) = F_i$. Feature sets with cardinality two define interactions between two features, F_i and F_j, and the target F_t. The interaction function in this case is the XOR function, namely: $f(\mathscr{F}_{int}) = XOR(F_i, F_j)$.

Each of the interacting feature sets have different conditional probability values $P(\mathscr{F}_{int}|F_t)_{F_t=f(\mathscr{F}_{int})}$ assigned to them. We considered conditional probability values of 0.8, 0.7, 0.6 and 0.5. The feature sets with conditional probability values of 0.5 are in fact random variables which are independent of the target F_t, an can be considered as irrelevant features.

With combinations of these feature interaction sets, three datasets were generated. The first dataset comprises only of individually correlated features and is named “single”. The second dataset contains features related via an XOR relation, as well as irrelevant features. It

n	$\lvert\mathscr{F}_{int}\rvert$	$f(\mathscr{F}_{int})$	$P(\mathscr{F}_{int}\lvert F_t)$
3	1	F_i	0.8
3	1	F_i	0.7
3	1	F_i	0.6
91	1	F_i	0.5

(a) "single" dataset

n	$\lvert\mathscr{F}_{int}\rvert$	$f(\mathscr{F}_{int})$	$P(\mathscr{F}_{int}\lvert F_t)$
3	2	$XOR(F_i,F_j)$	0.8
3	2	$XOR(F_i,F_j)$	0.7
3	2	$XOR(F_i,F_j)$	0.6
82	1	F_i	0.5

(b) "pair" dataset

n	$\lvert\mathscr{F}_{int}\rvert$	$f(\mathscr{F}_{int})$	$P(\mathscr{F}_{int}\lvert F_t)$
3	2	$XOR(F_i,F_j)$	0.8
3	2	$XOR(F_i,F_j)$	0.7
3	2	$XOR(F_i,F_j)$	0.6
3	1	F_i	0.8
3	1	F_i	0.7
3	1	F_i	0.6
73	1	F_i	0.5

(c) "combined" dataset

Table 3: Synthetic datasets statistics according to the feature interaction sets contained in each dataset. The "single" dataset contains relevant features that are just individually correlated to the target. The "pair" dataset contains features that are related to the target via an XOR relation. The "combined" dataset is a combination of the previous two datasets

is named "pair". The third dataset is a combination of XOR related feature interactions and individually correlated features, named "combined". In order to simulate the redundancy of features, which occurs in real datasets, each dataset contains several identically defined feature sets.

The complete statistic of the generated datasets and their feature interaction sets, are summarised in Table 3. All of the datasets consisted of 1000 instances and 100 features in total. From the 100 features, the "single" dataset has 9 relevant features, the "pair" dataset contains 18 relevant features and the "combined" dataset contains 27 relevant features.

For each dataset, we define the ground truth ranking, R_{GT}, from the feature relevance values rel_i. The relevance values are calculated directly from the specified feature interaction structure, by using Equation 19:

$$\mathrm{rel}(\mathrm{F_i},\mathrm{F_t}) = \sum_{\mathscr{F}_S \subseteq \mathscr{F} \setminus F_i} [\mathrm{I}(\mathscr{F}_S, F_i; F_t) - \sum_{S \subset \{\mathscr{F}_S \cup \{F_i\}\}} \mathrm{I}(S; F_t)]$$

Recall the constraint that a feature F_i can belong to only one feature interaction set $\mathscr{F}_{int}$. Considering this constraint the equation for calculating the ground truth relevance can be simplified to:

$$\mathrm{rel}(\mathrm{F_i},\mathrm{F_t}) = \mathrm{I}(\mathscr{F}_{int}, F_i; F_t)$$

In addition, we consider that all members of a feature interacting set $\mathscr{F}_{int}$ contribute equally to the information of the entire set. Therefore when calculating the relevance of a feature $F_i \in \mathscr{F}_{int}$, we weight the information magnitude inverse of the size of the interaction set and obtain:

$$\mathrm{rel}(\mathrm{F_i},\mathrm{F_t}) = \frac{\mathrm{I}(\mathscr{F}_{int}, F_i; F_t)}{|\mathscr{F}_{int} \cup \{F_i\}|}$$

The intuition is that features that contain more information about the target F_t individually, should be rewarded with a higher relevance, than features that are informative about the target but only in conjunction with other features.

5.2 Evaluation by Randomising the Ground Truth Ranking

The goal of the experiments presented in this section is to demonstrate the usefulness of our feature ranking evaluation method. As mentioned in the discussion in Section 4.1, feature ranking methods provide an approximation of the ground truth ranking that can be viewed as a noisy ground truth. A noisier ranking is more distant from the ground truth ranking and therefore of a worse quality.

An evaluation method should be sensitive to different amounts of noise and should provide a corresponding quality estimate of the feature ranking. For that purpose, we design experiments that are able to demonstrate that our evaluation method is sensitive to the addition of noise to the ground truth ranking. In Section 5.2.1, we describe in detail the setup used when performing the experiments. The results are discussed in Section 5.2.2 and Section 5.2.3.

5.2.1 Experimental Setup

In order to investigate whether our evaluation method is sensitive to noise added to the ground truth ranking, we first generate the noisy feature rankings and then construct FFA/RFA curves from them. Accordingly, we structure the description of the experimental setup into two parts: we first explain how to generate the noisy feature rankings and then explain how the error estimates for the FFA/RFA curves are calculated.

The noisy ground truth rankings are produced by using a noisy feature rankings generator, based on the function:

$$\mathrm{R}_\theta = noise(\mathrm{R}_{GT}, \theta)$$

where θ is the proportion of noise added to the ground truth ranking.

The noise is introduced by selecting a proportion, θ, of the features, which are randomly selected. For these features, random relevance values are assigned, while the remaining features preserve their ground truth relevance. By considering these partially changed relevance values a new, noisy feature ranking, R_θ, is defined.

Now that we have described how the noisy rankings are generated, the question is how to properly construct the FFA/RFA curves. The aim is to construct curves, which would relate to specific amount of noise, θ. As the random relevance values can be distributed differently throughout the ranking, different FFA/RFA curves can be constructed for the same amount of noise. Therefore, for a specific amount of noise, θ, the expected error curve, $E[FFA]_\theta$ and $E[RFA]_\theta$, should be calculated.

In theory, calculating all of the possible FFA/RFA curves for a given amount of noise is required, if we want to calculate the expected FFA/RFA curves. However, in practise, we estimate the expected error values by sampling the space of possible FFA/RFA curves for a given θ. This is done by first generating n different noisy feature rankings and then constructing n FFA/RFA curves based on them. The expected FFA/RFA curve is estimated by averaging the values of the individual curves, namely:

$$E[FFA]_\theta = \frac{1}{n}\sum_{i=1}^{n} FFA_{\theta,i}$$

$$E[RFA]_\theta = \frac{1}{n}\sum_{i=1}^{n} RFA_{\theta,i}$$

for a specified n and θ. This is schematically represented in Figure 5.

Each individual FFA/RFA curve is constructed by following the algorithmic description from Table 5 in Section 4.3. For estimating the error values, SVMs with a polynomial (quadratic) kernel were used and a 10-fold cross validation was performed, on the dataset

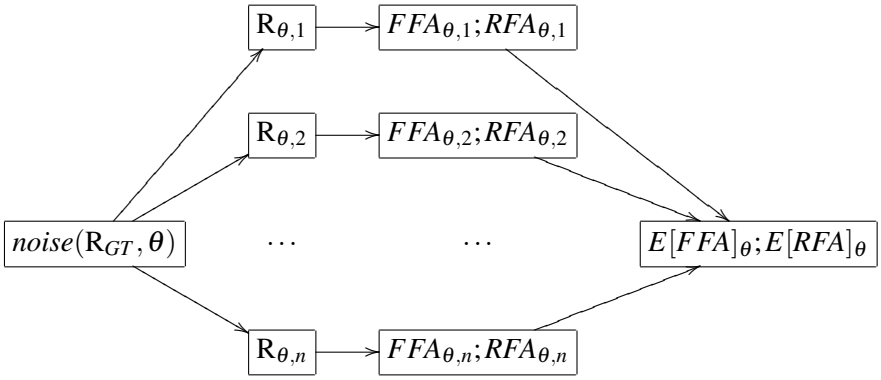


Figure 5: Graphical representation of the procedure for error curve estimation for a given level of noise θ. The noise generator produces n noisy rankings from the original ground truth ranking. Each ranking is evaluated, yielding an error curve and the error estimates from the different curves are averaged, thus providing the error estimate for the final curve.

under consideration. The epsilon parameter of the SVMs was set to 1.0E-12, while the complexity parameter was set to 0.1.

For our experiments, we consider several different amounts of noise, namely: 5%, 10%, 15%, 20%, 30% and 50%, as well as the completely random ranking (100% of noise). Each noisy error curve was produced by summarising the errors of 100 noisy rankings produced for a given θ. We additionally constructed error curves based on the ground truth ranking. In summary, for the value of θ we had:

$$\theta = \{0; 0.05; 0.1; 0.15; 0.2; 0.3; 0.5; 1\}$$

The experiments were performed on the three synthetic datasets described in Section 5.1, each with its corresponding ground truth ranking, R_{GT}.

5.2.2 Comparison of FFA and RFA Curves

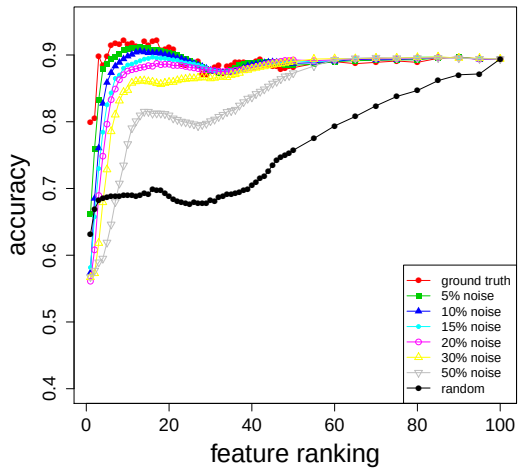
The results of our experiments are plotted as graphs containing error curves. Each graph (e.g., in Figure 6a) refers to a single dataset and contains plots of either FFA or RFA curves. The FFA/RFA curves plotted on each graph are for rankings with different noise levels θ, as well as for the ground truth R_{GT} and random rankings.

The graphs are organised in two columns of three graphs, each as in Figure 6. The left column of figures contains plots of FFA curves, while the right column contains RFA curves. Each row of figures refers to a single dataset and the corresponding ground truth ranking, from which the noisy rankings are generated.

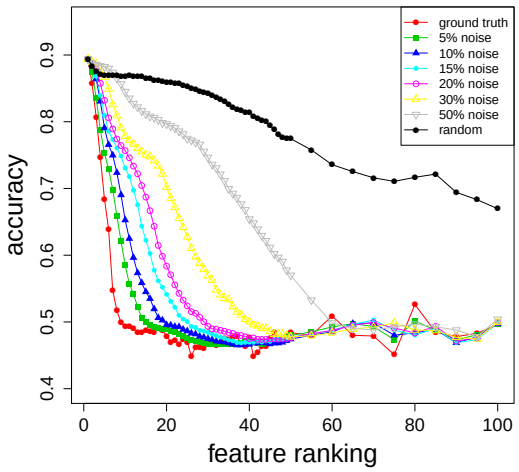
First, we consider the results for the "single" dataset in Figure 6a and Figure 6b. In both figures, the FFA and the RFA curves seem to be sensitive to the addition of noise. To begin with, the FFA/RFA curves of all the noisy rankings are located between the ground truth ranking FFA/RFA curve and the random ranking FFA/RFA curve. As noise is added to the ground truth ranking, the FFA/RFA estimates are slowly moving away from the ground truth FFA/RFA curve, or the other way around, they are slowly moving towards the random ranking FFA/RFA curve.

Next, we consider the results for the dataset containing only features related to the target via the second degree XOR relation (Figure 6c and Figure 6d). If we first consider the FFA curves, in Figure 6c, we can see that they are sensitive to the addition of noise to the ground truth ranking, in a similar way as for the "single" dataset. The RFA curves in Figure 6d, seem to be less sensitive to noise as the differences between the noisy RFA curves are quite small. However, if we consider the size of the region between the ground truth and the random RFA curves, the differences are quite proportional to it.

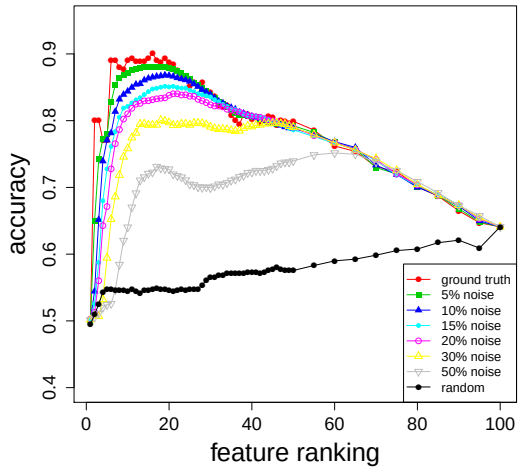
We now consider the last set of results for the current section, namely the combination of individually associated and XOR interacting features. If we consider Figure 6e and Figure 6f, the results are consistent with the previously discussed ones. As noise is added to the ground truth ranking the FFA/RFA curves are becoming increasingly similar to the random rankings curve. Both types of curves are sensitive to the level of added noise.



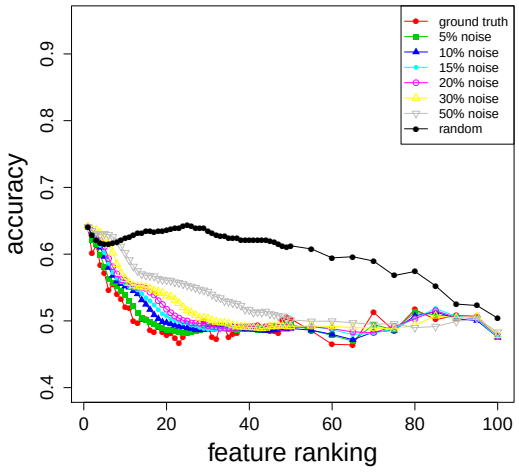
(a) FFA curves for the "single" dataset



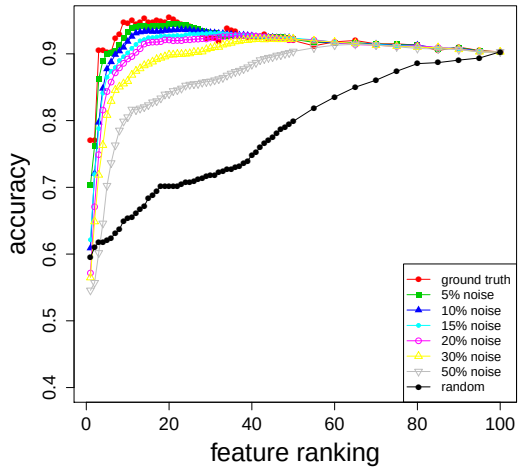
(b) RFA curves for the "single" dataset



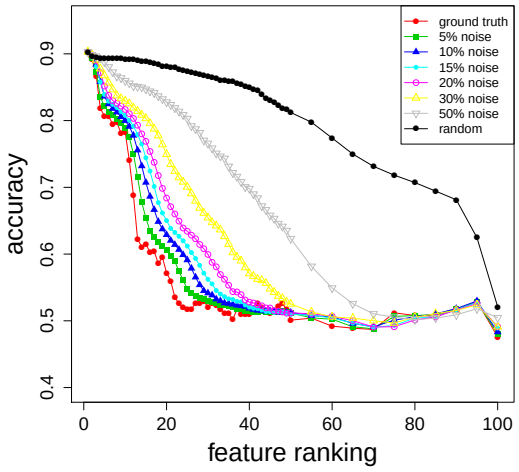
(c) FFA curves for the "pair" dataset



(d) RFA curves for the "pair" dataset



(e) FFA curves for the "combined" dataset



(f) RFA curves for the "combined" dataset

Figure 6: Plots comparing the FFA (left column) and RFA (right column) curves for different synthetic datasets (rows). Each figure contains error curves for the GT ranking, rankings with different noise levels θ and the random ranking.

5.2.3 Numerical Analysis of Error Curves

The visual inspection of the error curves in Figure 6 shows that error curves can be used to compare feature rankings with different levels of noise w.r.t. the ground truth ranking. In this section, we analyse the sensitivity of the error curves to the different levels of noise. Our aim is to see whether the changes in the error curves accurately reflect the amount of noise added to the ground truth ranking.

For this kind of quantitative analysis, we first need to summarise the differences of the noisy rankings error curves w.r.t. the ground truth error curve. This can be done as discussed in Section 4.5. Also, some kind of baseline is required for comparison. As the ground truth ranking is known, the distance between the ground truth ranking and the noisy rankings can serve as a baseline. By comparing the error curve differences and the ranking distance for different values of θ, we can determine how sensitive the error curves are to the added noise in the ground truth ranking.

For calculating a numeric summary of the difference between two FFA/RFA curves, following the discussion in Section 4.5, we use an instantiations of Equation 22. If we have two error curves, derived from feature ranking methods r_A and r_B, we calculate the weighted average of the differences between the curves at each point i of the curves. For the FFA curve we use:

$$FFA_{diff}(r_A, r_B) = \frac{\sum_{i=1}^{n} w(i) \cdot (FFA_{r_A}(i) - FFA_{r_B}(i))}{\sum_{i=1}^{n} w_i}$$

and for the RFA we use:

$$RFA_{diff}(r_A, r_B) = \frac{\sum_{i=1}^{n} w(i) \cdot (RFA_{r_A}(i) - RFA_{r_B}(i))}{\sum_{i=1}^{n} w_i}$$

.

We combine both values into a single value by calculating the error curve average (ECA):

$$ECA_{diff}(r_A, r_B) = \frac{FFA_{diff}(r_A, r_B) - RFA_{diff}(r_A, r_B)}{2} \tag{24}$$

Note that the minus sign in the equation is due to the inverse interpretation of negative values for the RFA curve. Namely, if r_A is better than r_B, then the differences of the RFA curves should be negative. This places the overall interpretation of the ECA_{diff} value on the positive scale. Namely, if r_A is better than r_B, then the overall score should be positive.

Different weighting functions are considered, such as the ones discussed in Section 4.5, namely:

- $w_i = 1$, equal weight for all differences
- $w_i = f(i) = 1/|\mathrm{R}_i|$, weight inverse to the feature subset size
- $w_i = f(diff_i) = diff_i$, weight proportional to the difference magnitude
- $w_i = f(i, diff_i) = diff_i/|\mathrm{R}_i|$, weight which includes both position and magnitude

For calculating the baseline values, i.e., the distance between the ground truth ranking R_{GT} and the noisy rankings $\mathrm{R}_{\theta,i}$, we use the average Spearman rank correlation coefficient, calculated as

$$dist(\mathrm{R}_{GT}, \mathrm{R}_{\theta}) = 1 - \bar{\rho}_{GT,\theta} = 1 - \frac{1}{k} \sum_{i=1}^{k} \rho(\mathrm{R}_{GT}, \mathrm{R}_{\theta,i})$$

where k is the number of different noisy rankings considered for a given θ. The value of ρ is calculated as:

$$\rho = \frac{\sum_{i=1}^{n} (rank(F_{GT,i}) - \overline{rank(F_{GT})})(rank(F_{\theta,i}) - \overline{rank(F_{\theta})})}{\sqrt{\sum_{i=1}^{n} (rank(F_{GT,i}) - \overline{rank(F_{GT})})^2 \sum_{i=1}^{n} (rank(F_{\theta,i}) - \overline{rank(F_{\theta})})^2}}$$

	$\theta = 0.05$	$\theta = 0.1$	$\theta = 0.15$	$\theta = 0.2$	$\theta = 0.3$	$\theta = 0.5$	$\theta = 1$	
1- rank corr.	0.219	0.34	0.449	0.51	0.628	0.81	1.056	**corr.**
w= 1	0.012	0.023	0.036	0.048	0.076	0.145	0.242	0.937
w= 1/rank	0.037	0.067	0.083	0.1	0.128	0.174	0.194	0.992
w= diff	0.059	0.098	0.120	0.141	0.169	0.216	0.266	0.998
w= diff/rank	0.095	0.147	0.164	0.186	0.211	0.252	0.262	0.955

(a) "single" dataset comparisons

	$\theta = 0.05$	$\theta = 0.1$	$\theta = 0.15$	$\theta = 0.2$	$\theta = 0.3$	$\theta = 0.5$	$\theta = 1$	
1- rank corr.	0.126	0.239	0.327	0.397	0.519	0.726	1.091	**corr.**
w= 1	0.006	0.013	0.020	0.026	0.043	0.085	0.182	0.974
w= 1/rank	0.016	0.035	0.046	0.054	0.074	0.105	0.147	0.998
w= diff	0.030	0.056	0.065	0.075	0.094	0.126	0.207	0.995
w= diff/rank	0.058	0.099	0.107	0.115	0.133	0.159	0.209	0.986

(b) "pair" dataset comparisons

	$\theta = 0.05$	$\theta = 0.1$	$\theta = 0.15$	$\theta = 0.2$	$\theta = 0.3$	$\theta = 0.5$	$\theta = 1$	
1- rank corr.	0.1	0.171	0.252	0.32	0.432	0.652	1.048	**corr.**
w= 1	0.009	0.02	0.027	0.037	0.061	0.117	0.223	0.992
w= 1/rank	0.018	0.042	0.047	0.064	0.084	0.132	0.178	0.991
w= diff	0.029	0.061	0.070	0.09	0.115	0.174	0.263	0.998
w= diff/rank	0.044	0.091	0.095	0.121	0.142	0.199	0.254	0.982

(c) "combined" datasets comparisons

Table 4: Comparison of different ECA values obtained by different weighting functions w. The ECA values are compared with the distance between the noisy rankings R_θ and the GT ranking R_{GT}. The final column of each table "corr." is the value of the correlation coefficient calculated between the ranking distance (first row) and each of the ECA differences rows.

where n is the number of features.

We summarise the results in Table 4a, Table 4b and Table 4c. Each table contains values calculated with respect to the ground truth ranking. The first row of the table refers to the distance between the ground truth ranking R_{GT} and the noisy rankings R_θ calculated with the Spearman rank correlation coefficient. The other rows are the ECA differences between the FFA/RFA curves of the GT ranking and the FFA/RFA curves of the noisy rankings. Each row containing the ECA differences refers to different weighting functions. All columns, except the last one, refer to different levels of noise, θ. The final column, gives the correlation coefficient calculated between the Spearman rank correlation coefficient (row one) and the curve distances (rows 2 to 5), for the different levels of noise θ.

The final column gives an indication of how well the ECA differences relate to the distance between the ground truth ranking and the noisy rankings. As it can be seen in the tables, for every dataset, the curve distances correlate very well to the rank distances. The highest correlation can be observed, when the weight of the ECA difference is a function of the rank or the magnitude of difference (rows 3 and 4).

From this quantitative analysis, it can be concluded that the ECA difference, derived from the error curves, has the same sensitivity to noise as the actual distance between the ground truth and the noisy rankings. This implies that our method in practical scenarios, can be used not just to qualitatively distinguish between different rankings, but also to quantify the difference between them. As for the specific weights used for calculating the ECA differences, it can be concluded that any of the considered weights for curve distances, can be used to properly compare the error curves.

5.2.4 Conclusions

In this section, we performed experiments designed to illustrate the use and demonstrate the usefulness of our evaluation method. The purpose of our evaluation method is to distinguish between feature rankings of different quality. By adding noise to a known ground truth ranking, we provide feature rankings with worsening quality that are more distant from the ground truth.

The visual inspection of the FFA and RFA curves in Figure 6 shows that as more noise is added to the ground truth, the error curves become more distant from the ground truth error curves. In addition, they slowly become similar to the expected error curve, i.e., the expected error curve of a completely random ranking. This clearly demonstrates that our evaluation method is capable of distinguishing between feature rankings of different quality and can provide an answer to the practical question of "whether ranking R_A is better than R_B".

Additionally, the ECA differences derived from the error curves, given in Table 4a, Table 4b and Table 4c, show that our method is also sensitive to the specific amount of noise, θ. Irrespective of the weighting function used for calculating the numeric scores, the values of the ECA differences are highly correlated to the values of the actual distance between the ground truth ranking and the noisy rankings. This means that the ECA values derived from the error curves can be used to quantify the difference between different rankings.

5.3 Evaluation of Different Feature Ranking Methods

In the previous section, we analysed rankings of different quality (with different amounts of noise) by comparing them to the ground truth ranking. In a real-world setting, the ground truth ranking is unknown and the feature rankings are induced directly from the data. Therefore, in this section we analyse feature rankings, produced by different feature ranking methods, induced from the synthetic data described in Section 5.1. The whole experimental procedure, including the feature ranking methods considered, is described in Section 5.3.1.

Our analysis of the results starts in Section 5.3.2 by investigating each feature ranking method individually. We examine the correctness of the produced feature rankings in comparison to the ground truth ranking, as well as their stability, and establish a connection between these and the constructed error curves. We then proceed, in Section 5.3.4, to a comparative assessment of the feature ranking methods. We visually compare the FFA and RFA curves and also perform a quantitative analysis of the error curves. The summary conclusions are given in Section 5.3.5.

5.3.1 Experimental Setup

In this section, we present in detail how the experiments were conducted. To fully describe the experimental setup, we first define what feature ranking methods were used and on which data they were employed. We then present the specific experimental procedure used to generate the rankings and their evaluation.

For our experiments, we consider four feature ranking methods:

- **Information gain**, calculating the information gain of each feature F_i as: $IG(F_t, F_i) = H(F_t) - H(F_t|F_i)$. This does not require any specific parameter setting.

- **SVM-RFE** is the redundant feature elimination (RFE) procedure that employs an SVM to evaluate the feature weights at each iteration. A linear SVM was employed, as proposed by Guyon et al. (2002). The epsilon parameter of the SVM was set to 1.0E-12 , while the complexity was set to 0.1.

- **ReliefF** algorithm as proposed by Robnik-Šikonja and Kononenko (2003). The number of neighbours was set to 10 and all of the instances were used for estimating the relevance values.
- **Random forests**, which can be used for estimating feature relevance as described by Breiman (2001). A forest of 100 trees was used, constructed by randomly choosing a $\log_2$ of the number of features.

To generate the error curves, SVMs with polynomial (quadratic) kernel, were employed as classifiers. The epsilon parameter was set to 1.0E-12, while the complexity parameter was set to 0.1. For estimating the stability of each feature ranking algorithm the stability indicator described by Jurman et al. (2008) was used.

This stability indicator employs the Canberra distance for feature rankings when calculating the stability value. It was used for several reasons. The Canberra distance is very easy to calculate and puts a larger weight on the stability of top-ranked features. This is convenient, as most of our features in the dataset are irrelevant and we are mostly concerned with the stability of the relevant or top-ranked features. The adaptation provided by Jurman et al. (2008) also allows for the comparison of partially ranked lists with the Canberra distance. In addition, the stability index is normalised in such a way that it is possible to directly compare two indices calculated for different sizes k of partially ranked lists.

All of the feature ranking methods were employed on the three datasets, "single", "pair" and "combined", described in Section 5.1.

5.3.2 Individual Analysis of Feature Ranking Methods

In this section, we discuss each of the feature ranking methods described in Section 5.3.1 individually. We first investigate what kind of feature ranking each feature ranking method produces: we analyse the feature rankings in term of how well ordered they are and also in terms of how stable they are. We then discuss the FFA and RFA curves for these feature rankings.

For our analysis, we consider three types of graphs. The first type of graph (e.g., Figure 7c) is an error curve graph. It contains the plots of both the FFA and RFA curves. The second type of graph is the stability estimate graph (e.g., Figure 7e). The y-axis refers to the value of the stability indicator: the higher the value, the less stable the ranking method. Each point, k, of the x-axis represents the size of the considered feature subsets, consisting of top ranked k features. The third type of graph represents the distribution of the ground truth relevance (e.g., Figure 7a). The y- axis is the ground truth relevance value, as estimated in Section 5.1. Each point, i, represents the i-th ranked feature, F_{ri}, as determined by the feature ranking method. Each graph refers to a single feature ranking method and to a single dataset.

Overall, the graphs are organised in a two column format (e.g., Figure 7). A column contains three graphs, one of each of the three types and refers to a single dataset and ranking method. The graphs are ordered as follows, top to bottom: GT distribution graph, FFA/RFA graph, and stability graph. This allows for a quick visual overview of the results discussed in the following section.

The following discussion of the results is organised according to the dataset under consideration. As each of the datasets has a specific feature interaction structure and therefore a different ground truth ranking, it is logical to discuss the properties of each feature ranking method in this way.

Analysis of the "single" dataset

In Figure 7 and Figure 8, we present the results for the "single" dataset. If we first examine how the algorithms rank the relevant features (Figures 7a, 7b, 8a and 8b), we can conclude

they can all separate the relevant from the irrelevant features. Ranking by using the information gain, ReliefF and SVM-RFE produces a correct ordering of the relevant features. However, as it can be seen in Figure 7b, when using random forests the relevant features are obviously top-ranked, but their ordering seems to be mixed.

The stability evaluation in Figures 7e, 7f, 8e and 8f shows that all of the algorithms are stable in the region of the relevant features, except for random forests in Figure 7f, which has an instability peak exactly in this region. This means that random forests are in this case capable of detecting all the relevant features, but are highly unstable in the estimation of their ordering.

Finally, we consider how the feature ranking and the stability of the algorithms is reflected in the accuracy estimates provided by the FFA/RFA curves. As expected, since there are no differences between the ranking of the relevant features for information gain, ReliefF and SVM-RFE, all of them have very similar, almost identical FFA/RFA curves, as evident in Figures 7c, 7d,8c and 8d. The instability of the random forests in the region of the relevant features is reflected in the FFA/RFA curves, as it can be seen in Figure 7d. For the FFA curve, the accuracy estimate starts at a considerably lower value as compared to the FFA curves of the other algorithms. It then has steadily increasing values, until it reaches a maximum, which is similar to the curves of the other ranking methods.

Analysis of the "pair" dataset

We now consider the dataset where the features are related to the target via the XOR relation, i.e., via an interaction of degree two. In this case, the different algorithms have very different behaviours as compared to the simpler "single" dataset, which contains only features individually correlated to the class.

If we first consider the ordering of the relevant features in the final ranking in Figures 9a, 9b, 10a and 10b, it can be seen that only ReliefF produces a final ranking which is mostly correct. The other methods produce rankings which seems to have the relevant features randomly distributed over the feature ranking.

Also, when examining the stability estimates in Figures 9e, 9f, 10e and 10f, all methods have the same stability pattern, with the exception of ReliefF. Namely, as the number of features considered rises, the instability also increases until it reaches a certain "saturation" level, with random forests being the most unstable method. In contrast, ReliefF has a different stability pattern. It is quite stable in the region where the relevant features are concentrated and later gets unstable when determining the relevance of the irrelevant features.

Finally, we examine the FFA/RFA curves in Figures 9c, 9d, 10c and 10d. Both the FFA and RFA curves of information gain and SVM-RFE behave like a random (expected) curve, i.e., they have almost linearly increasing accuracy values. Although random forests are unstable and it seems that the relevant features are randomly distributed in the final ranking, the FFA curve is different from the ones of information gain and SVM-RFE. The curve shows that most of the relevant features are concentrated at the beginning of the ranking, although the ranking is not entirely correct overall. The FFA/RFA curves of ReliefF, in Figure 10c show a proper ranking. Namely, the FFA curve sharply increases its values at the beginning, quickly reaching a maximum value which is indicative of very relevant features being present at the top of the ranking. The RFA curve shows the same, as there is an increase in the estimated accuracy of the predictive models only when the top-ranked features are included.

Analysis of the "combined" dataset

Finally, we consider the dataset which contains features which are individually associated to the target as well as features that are related to it via an XOR relation of parity two. From

the previous results, when we considered both of the feature types separately, we would expect that some of the ranking methods would be able to detect just the individually associated features while for the others they would provide random relevance estimates.

This partially random behaviour of some ranking algorithms is exactly what can be seen if we consider the ranking of the relevant features in Figures 11a, 11b, 12a and 12b. Information gain and SVM-RFE are correctly ranking the features individually associated with the target, while the XOR features are randomly distributed in the ranking. Random forests separate relevant from irrelevant features, but the ordering of the relevant features is mixed. As expected from the previously discussed results ReliefF provides the most correct ranking.

Stability-wise, as seen in Figures 11e, 11f, 12e and 12f, all of the methods behave as in the "single" dataset case, with the visible instability peak for the random forests.

The above is reflected in a visible difference between the FFA/RFA curves of the different ranking methods, especially when considering the RFA curves. In Figures 11c and 12d, the RFA curves have a similar behaviour, namely a linearly increasing accuracy in the region where the relevant features are randomly distributed and a sharp increase in accuracy in the region which includes the properly ranked relevant features. If we compare this to the RFA curves of random forests and ReliefF in Figures 11d and 12c, we will notice that they only have a sharp increase in accuracy when the top-ranked features are included. There is also a difference the FFA curves in the same figures. Namely, although the FFA curves of information gain and SVM-RFE reach a point of high accuracy, this accuracy value is not as high as the one achieved by random forests and ReliefF.

5.3.3 Conclusions

In this section. we analysed four different feature ranking methods. Each feature ranking method produced feature rankings with different properties, ranging from well-ordered through partially random to completely random rankings. The type of feature ranking produced was clearly dependent on the feature interaction structure of the dataset under consideration.

Completely random rankings were produced by information gain and SVM-RFE when considering the "pair" dataset. They are characterised by high instability (Figure 9e and Figure 10f) and a uniform distribution of the relevant features (Figure 9a and Figure 10b). Their corresponding FFA and RFA curves are very similar to the expected error curve, which is further discussed in Section 5.3.4 and can be comparatively seen in Figure 13c and Figure 13d.

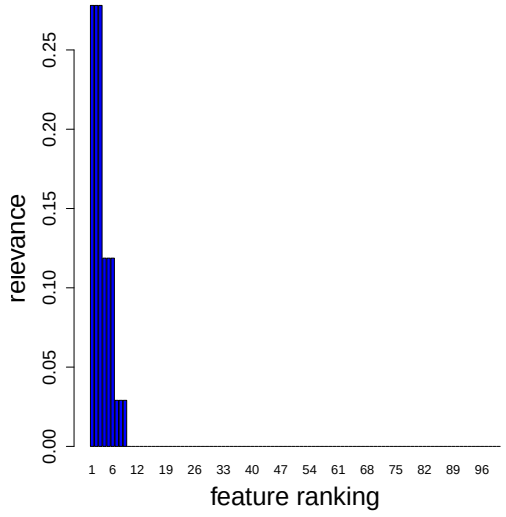
The partially random rankings exhibit some interesting properties. All of them are characterised by a uniform noise distribution, but only in certain areas of the feature ranking. These areas are determined by the type of features present in the ranking.

Random forests have the tendency to produce feature rankings where the relevant and non-relevant features are separated. However, uniform noise is present in the relevant region of the feature ranking. This is also visible in the specific stability patterns, in Figure 7f and Figure 9f, where there is an instability peak located in the middle of the relevance region.

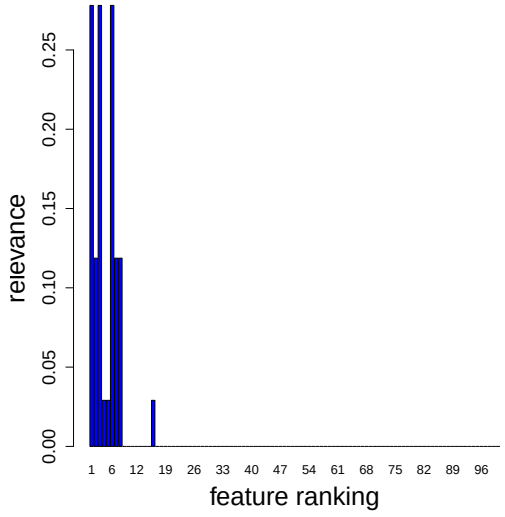
Information gain and SVM-RFE also produce partially random rankings when considering the "combined" dataset. Unlike random forests, they fail to separate all of the relevant from the non-relevant features. They are only able to properly rank the features that are individually associated with the target. The remaining relevant features, related to the target via an XOR relation, are randomly ranked. In this case, it can be said that the noise is uniformly distributed in the region after the individually associated features. All of these regions of randomness are also clearly reflected in the FFA and RFA curves.

Finally, the well ordered rankings are the ones that seem closest to the ground truth ranking. From our experiments, it seems that ReliefF is the only method that produces more

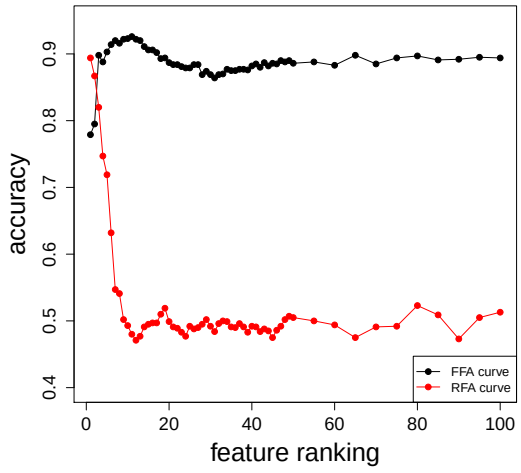
or less accurate rankings, independent of the dataset under consideration. Also, the rankings produced by ReliefF are fairly stable, even in the case of the "pair" dataset, where all other ranking methods produce highly unstable rankings. In addition to ReliefF, information gain and SVM-RFE produce well ordered rankings, but only in the case of the "single" dataset.



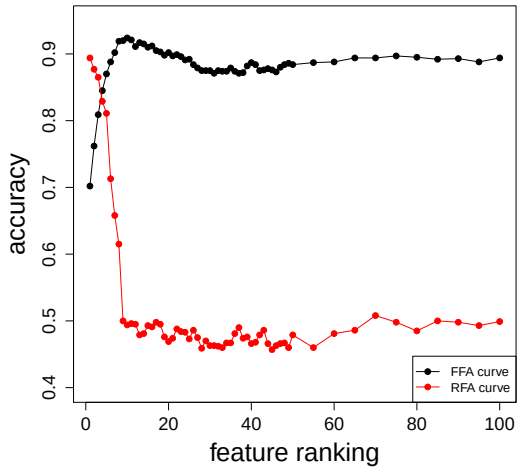
(a) **IG,** rel. features distribution



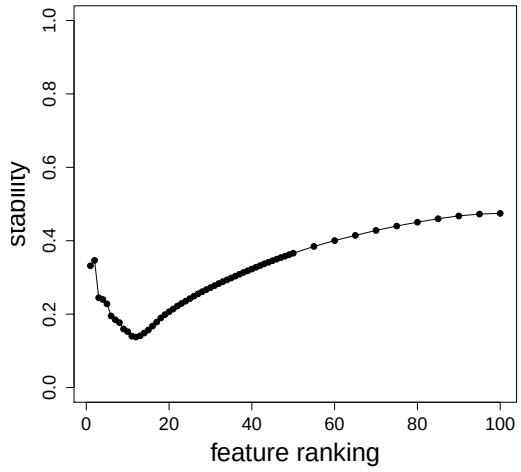
(b) **RFs,** rel. features distribution



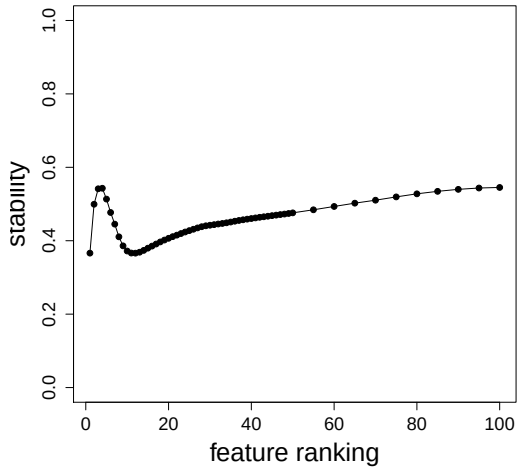
(c) **IG,** FFA and RFA curves



(d) **RFs,** FFA and RFA curves

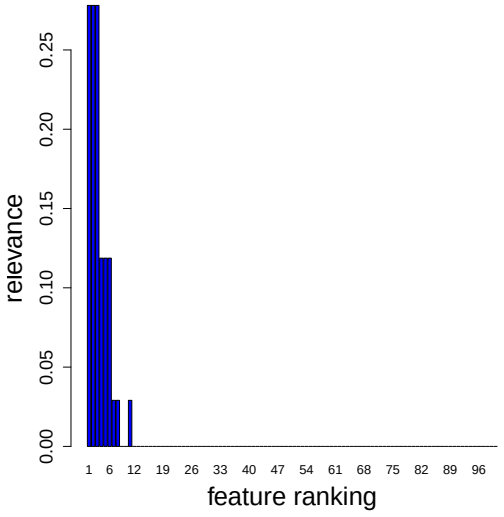


(e) **IG,** stability estimate

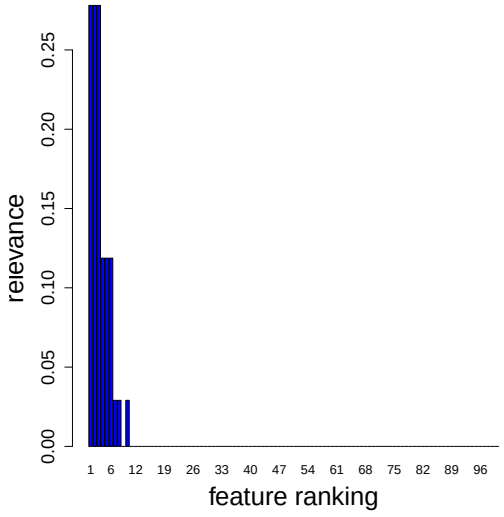


(f) **RFs,** stability estimate

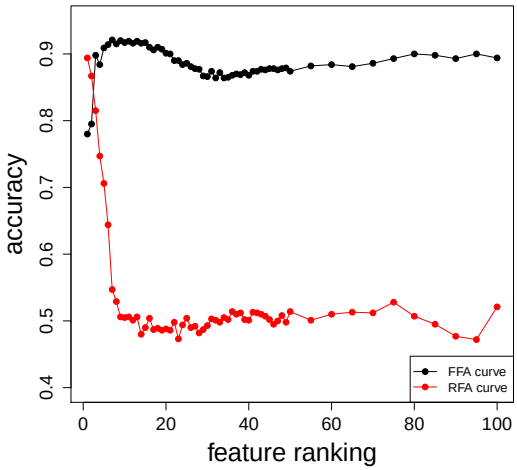
Figure 7: The properties of the rankings produced by information gain (IG) and random forests (RFs), on the "single" dataset. The graphs in the first row represent the distribution of the relevant features within the rankings. The second row graphs, plot the FFA and RFA curves. The third row graphs, depict the stability estimates for the ranking methods.



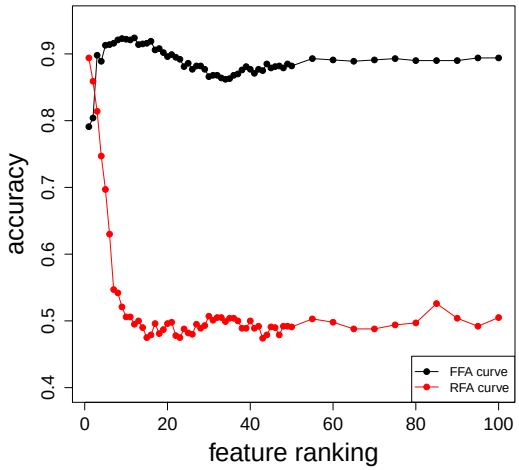
(a) **ReliefF,** rel. features distribution



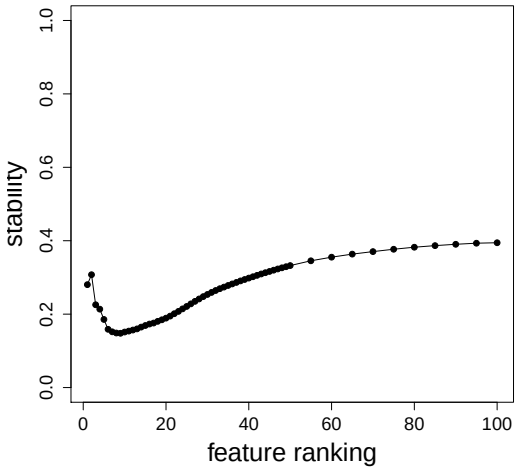
(b) **SVM-RFE,** rel. features distribution



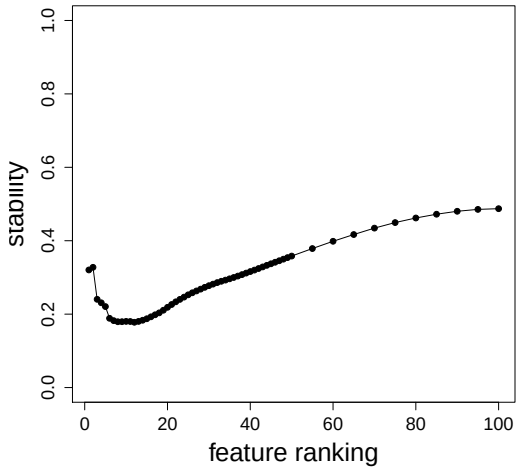
(c) **ReliefF,** FFA and RFA curves



(d) **SVM-RFE,** FFA and RFA curves

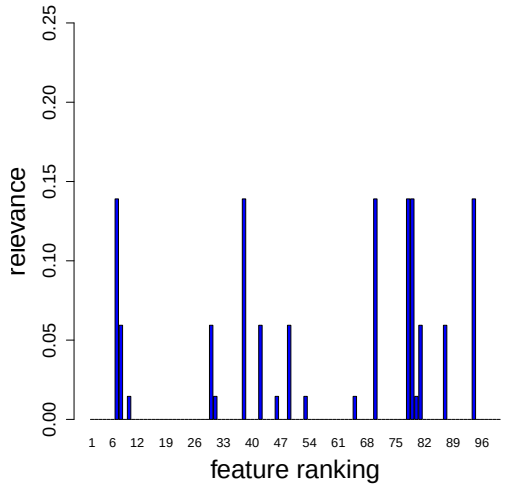


(e) **ReliefF,** stability estimate

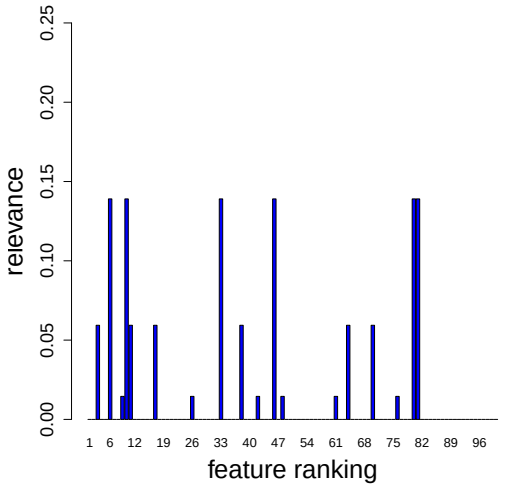


(f) **SVM-RFE,** stability estimate

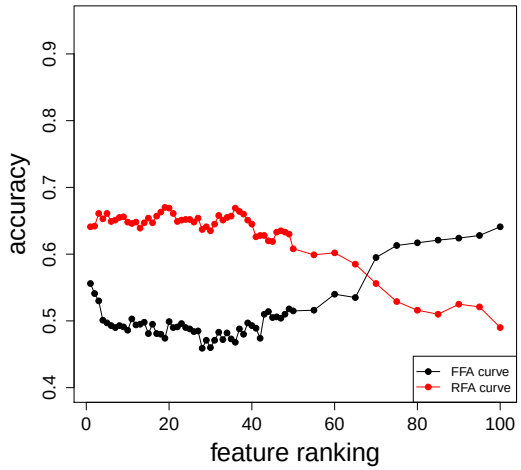
Figure 8: The properties of the rankings produced by ReliefF and SVM-RFE, on the "single" dataset. The graphs in the first row represent the distribution of the relevant features within the rankings. The second row graphs, plot the FFA and RFA curves. The third row graphs, depict the stability estimates for the ranking methods.



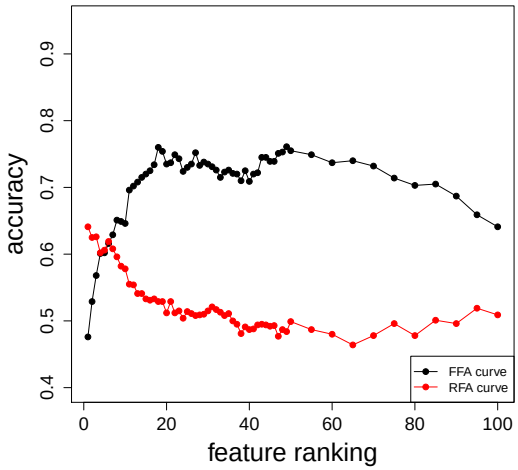
(a) **IG,** rel. features distribution



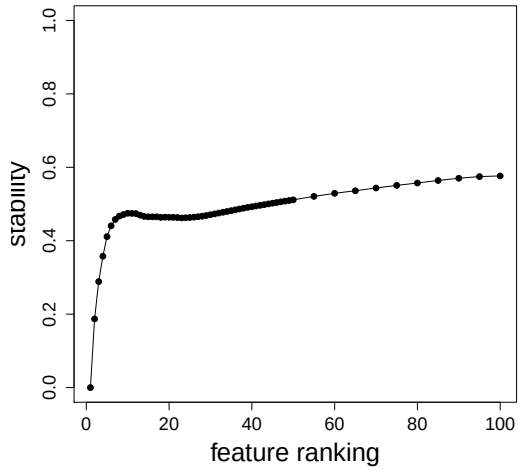
(b) **RFs,** rel. features distribution



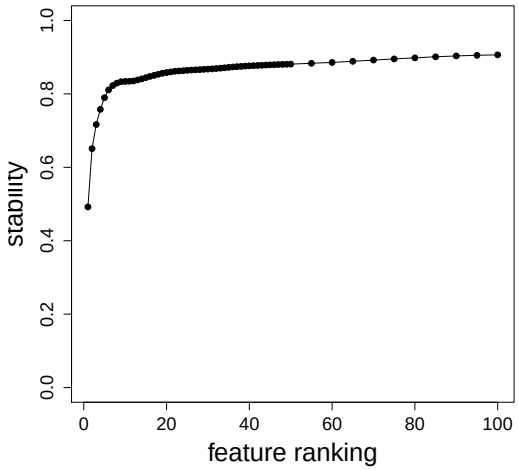
(c) **IG,** FFA and RFA curves



(d) **RFs,** FFA and RFA curves

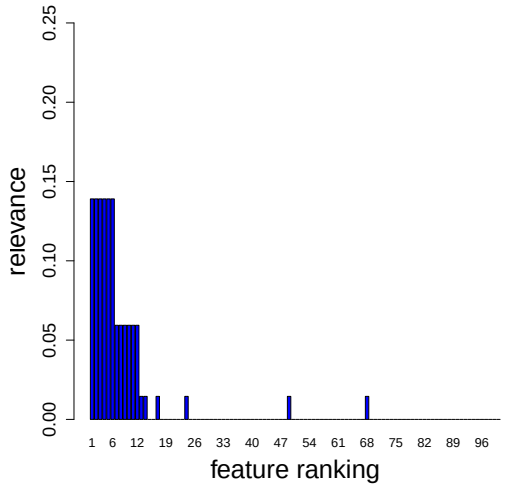


(e) **IG,** stability estimate

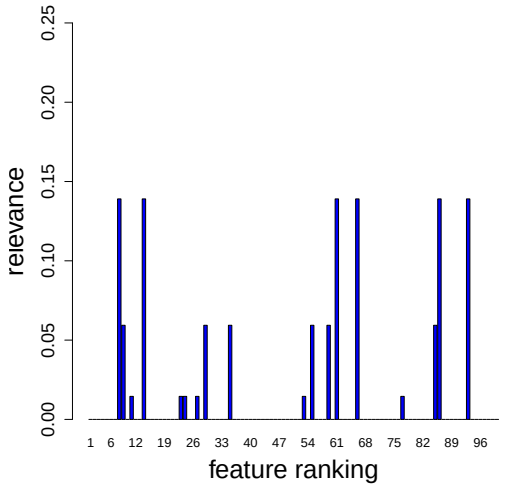


(f) **RFs,** stability estimate

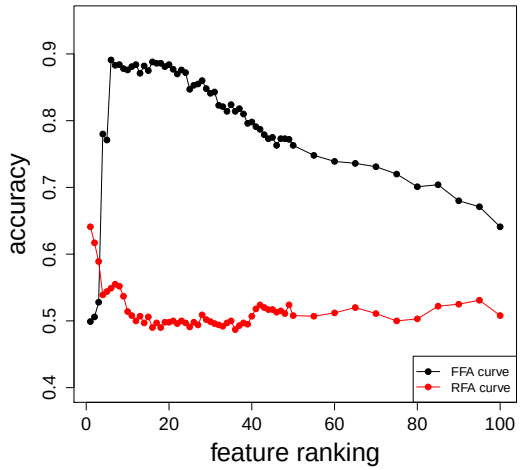
Figure 9: The properties of the rankings produced by information gain (IG) and random forests (RFs), on the "pair" dataset.. The graphs in the first row represent the distribution of the relevant features within the rankings. The second row graphs, plot the FFA and RFA curves. The third row graphs, depict the stability estimates for the ranking methods.



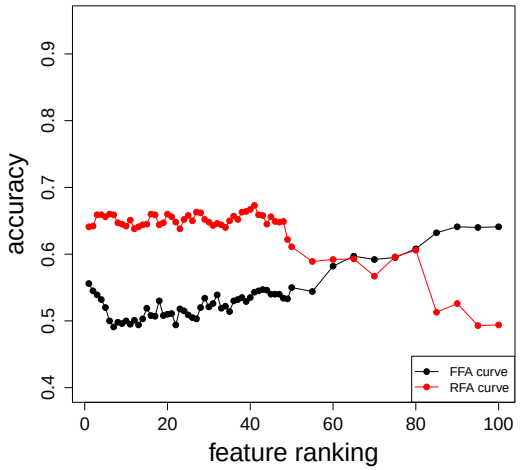
(a) **ReliefF,** rel. features distribution



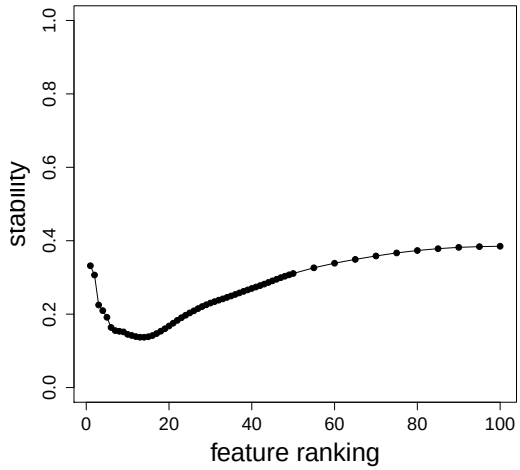
(b) **SVM-RFE,** rel. features distribution



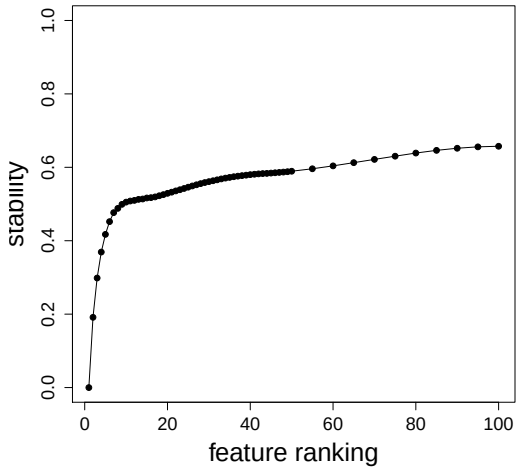
(c) **ReliefF,** FFA and RFA curves



(d) **SVM-RFE,** FFA and RFA curves

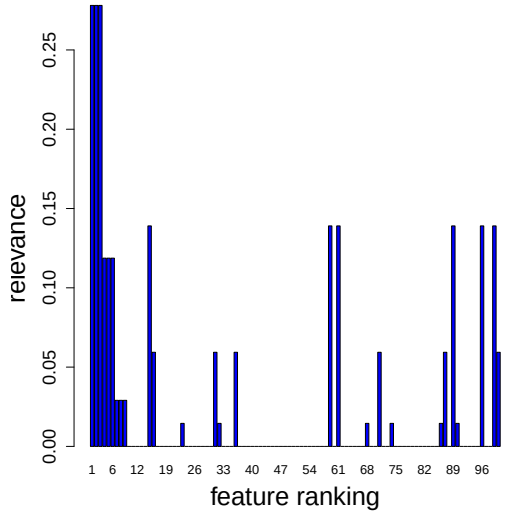


(e) **ReliefF,** stability estimate

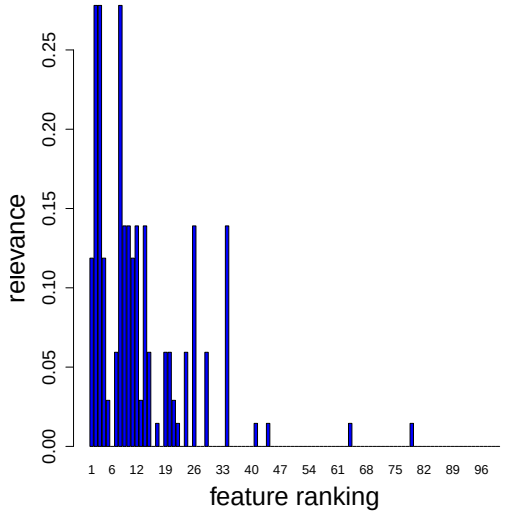


(f) **SVM-RFE,** stability estimate

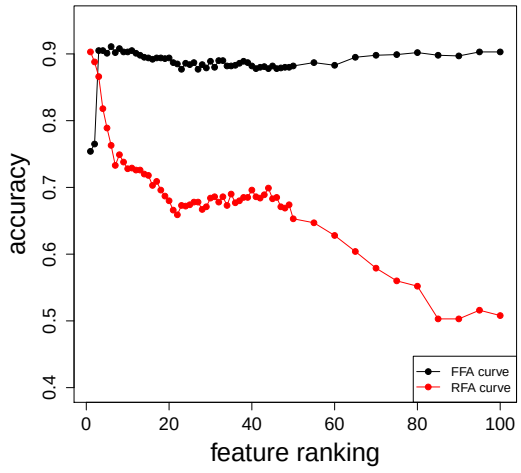
Figure 10: The properties of the rankings produced by ReliefF and SVM-RFE, on the "pair" dataset. . The graphs in the first row represent the distribution of the relevant features within the rankings. The second row graphs, plot the FFA and RFA curves. The third row graphs, depict the stability estimates for the ranking methods.



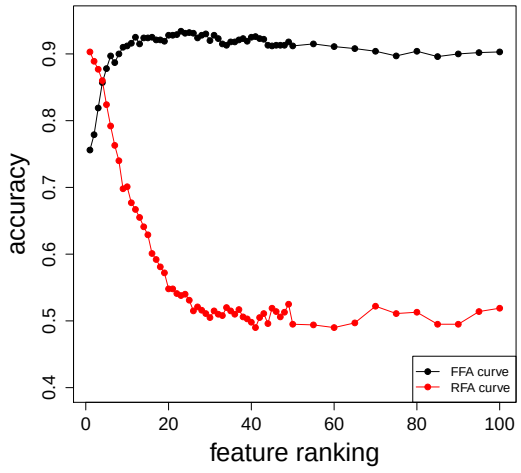
(a) **IG,** rel. features distribution



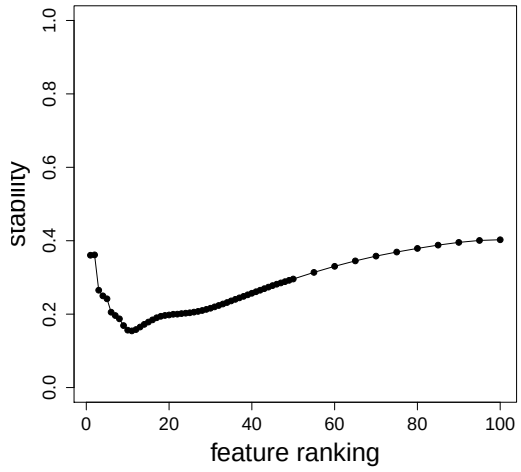
(b) **RFs,** rel. features distribution



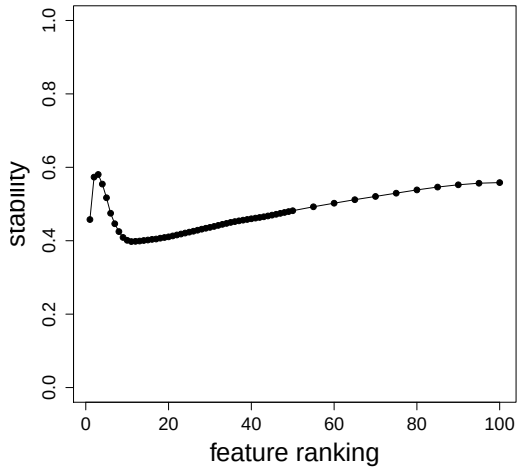
(c) **IG,** FFA and RFA curves



(d) **RFs,** FFA and RFA curves

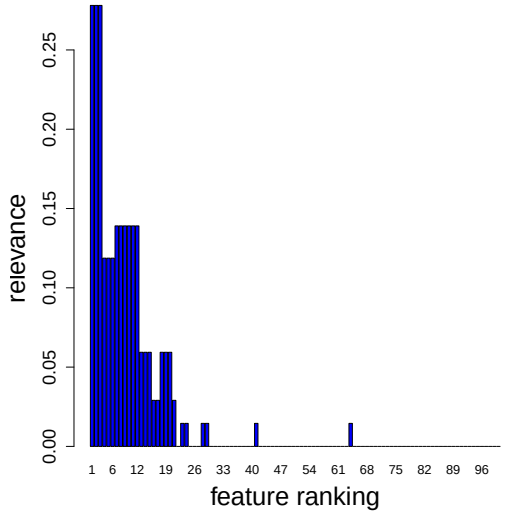


(e) **IG,** stability estimate

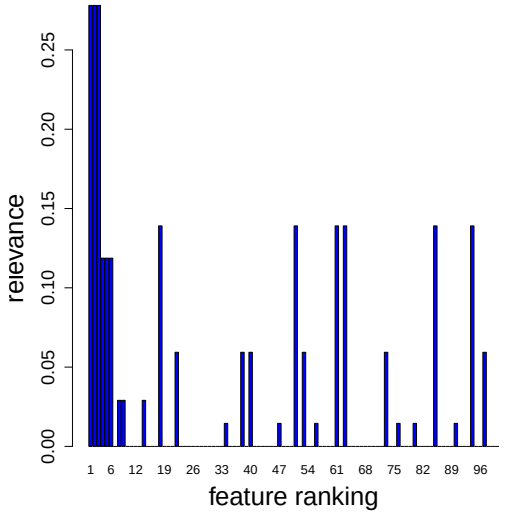


(f) **RFs,** stability estimate

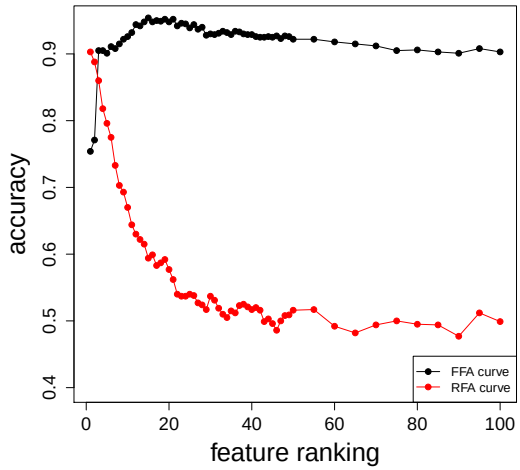
Figure 11: The properties of the rankings produced by information gain (IG) and random forests (RFs), on the "combined" dataset. The graphs in the first row represent the distribution of the relevant features within the rankings. The second row graphs, plot the FFA and RFA curves. The third row graphs, depict the stability estimates for the ranking methods.



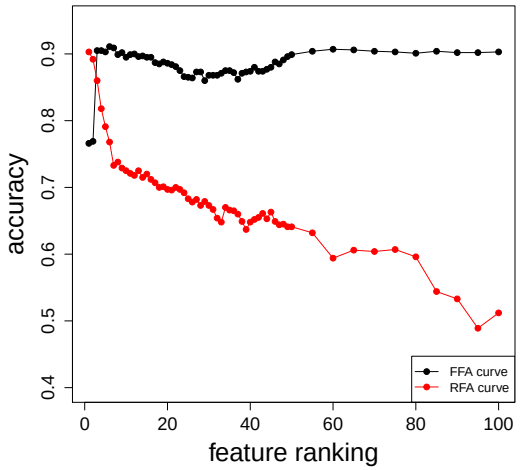
(a) **ReliefF,** rel. features distribution



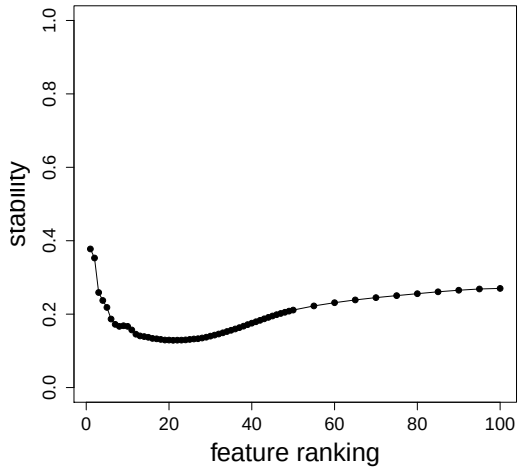
(b) **SVM-RFE,** rel. features distribution



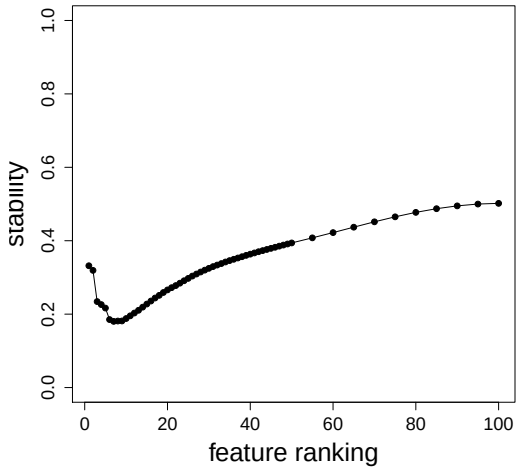
(c) **ReliefF,** FFA and RFA curves



(d) **SVM-RFE,** FFA and RFA curves



(e) **ReliefF,** stability estimate



(f) **SVM-RFE,** stability estimate

Figure 12: The properties of the rankings produced by ReliefF and SVM-RFE, on the "combined" dataset.. The graphs in the first row represent the distribution of the relevant features within the rankings. The second row graphs, plot the FFA and RFA curves. The third row graphs, depict the stability estimates for the ranking methods.

5.3.4 Comparative Analysis of Feature Ranking Methods

In this section, we present a comparative analysis of the FFA and RFA curves of the four different feature ranking methods, on the three datasets considered above. We first visually compare the error curves of the different ranking methods in Figure 13. We discuss their similarities and differences in context of the previous discussion in Section 5.3.2.

We then proceed to a quantitative analysis based on the error curves. We calculate the error curve average (ECA) difference between the error curves of all ranking methods, in the same manner as in Section 5.2.3. We also calculate the ECA differences between the GT ranking with different noise levels (Section 5.2.3) and perform a hierarchical clustering of the different rankings. We present the dendrograms in Figure 14, where we analyse the similarity of the different feature ranking methods.

Comparison of FFA and RFA Curves for Different Ranking Methods

We present the comparison of the FFA and RFA curves of the four different feature ranking methods in Figure 13. Each graph contains the plots of FFA or RFA curves of the different ranking methods. In addition to this, on each graph, the error curves of the ground truth ranking and the expected error curve (random ranking) are plotted. Each graph refers to a single dataset. Overall the graphs are organised in two columns and three rows. The left column contains the plots of the FFA curves, while the right column contains the plots of the RFA curves. Each row refers to one dataset.

We first examine the error curves for the "single" dataset in Figure 13a and Figure 13b. In light of the previous discussion of the "single" dataset in Section 5.3.4, all of the methods produce a more or less correct ranking and their corresponding error curves are almost identical to the error curve of the ground truth. There is only a slight barely noticeable, deviation of the error curves of random forests, which is due to the instability of random forests in the relevant features region.

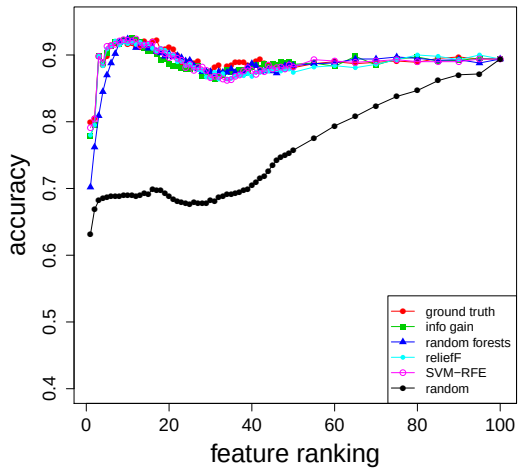
Next, we consider the results for the dataset containing only features related to the target via an XOR relation of second degree. The results are presented in Figure 13c and Figure 13d. As previously discussed in Section 5.3.4, information gain and the SVM-RFE method produce more or less random rankings. This is clearly visible when directly comparing their error curves to the error curves of the random rankings: their FFA curves are below the random curve, while their RFA curves are above the random curve.

ReliefF is the algorithm that provided the most correct ranking of the features, with FFA and RFA curves comparable to those of the ground truth ranking. Random forests are unstable, but still manage to identify some of the relevant features. Their results are similar to the ground truth ranking with 30-50% of added noise.

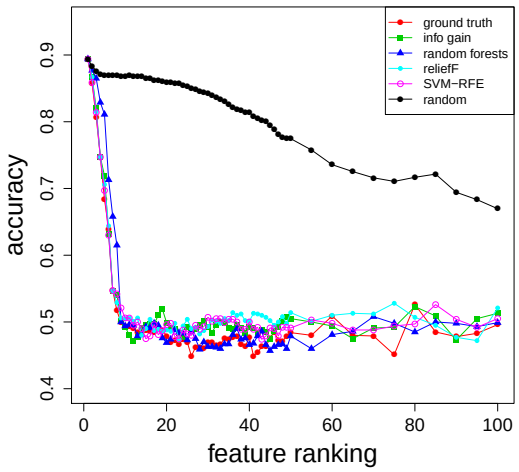
The last set of results discussed in this section are those for the "combined" dataset. Comparatively, the ranking methods exhibit different behaviours as visible in the FFA/RFA curves in Figure 13e and Figure 13f. It is again obvious that ReliefF captures the relevant features and provides mostly a correct ranking, as the error curves are quite similar to those of the ground truth ranking. Random forests also manage to capture the relevant features, but due to their instability, the FFA curve estimates never get as high as the maximum of the accuracy estimates of the ReliefF ranking. This means that the random forest ranking is on average worse than the ranking provided by ReliefF. For the information gain and the SVM-RFE algorithm, the FFA curves indicate that the methods manage to identify a portion of the relevant features and put them on the top of the ranking. However, this portion is smaller than the one identified by Random Forests.

This partially random behaviour of information gain and SVM-RFE is most obvious when considering their RFA curves. There are two distinct regions of the curves, the first being a linearly increasing region and the second with a sharp accuracy increase due to the presence of the individually associated features. Note that the reason that the random

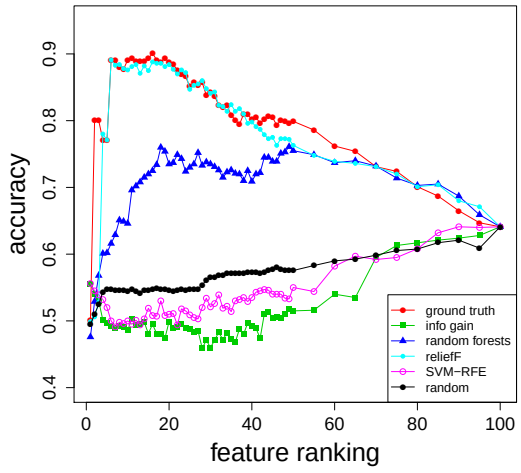
region does not overlap with the RFA curve of the random ranking is due to the fact that the individually associated features are not included in this region, as they are correctly ranked and are at the top of the ranking, while for the random ranking, all of the relevant features are distributed uniformly across the ranking.



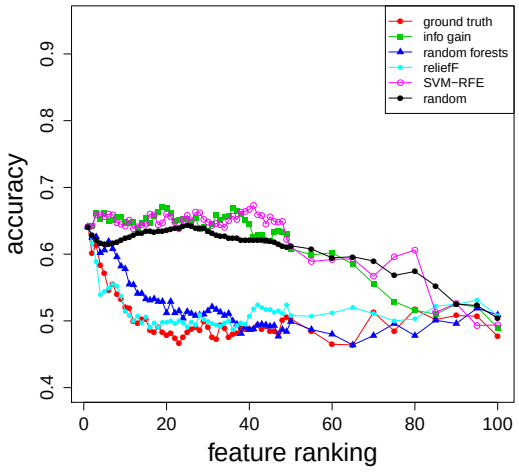
(a) “single” dataset, FFA curves



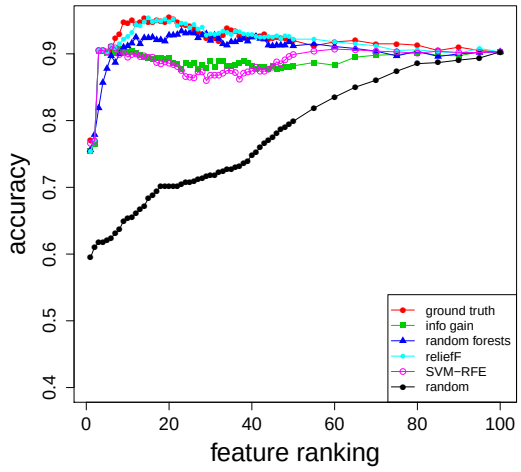
(b) “single” dataset, RFA curves



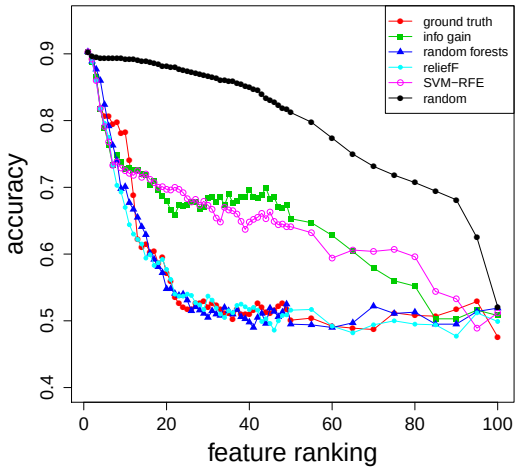
(c) “pair” dataset, FFA curves



(d) “pair” dataset, RFA curves



(e) “combined” dataset, FFA curves



(f) “combined” dataset, RFA curves

Figure 13: Figures representing the comparison of the FFA (left column) and RFA (right column) curves of different feature ranking algorithms. Each figure additionally contains plots of the GT and the random FFA/RFA curves.

Numerical Comparison of FFA and RFA Curves

After the visual inspection of the error curves, we also perform a quantitative analysis, similar to the one in Section 5.2.3. We calculate the ECA differences between the error curves of the different feature ranking methods, by using Equation 24. In addition, we calculate the ECA differences between the error curves of the different ranking methods and the different noisy rankings, described in Section 5.2.1. For the calculation of all ECA differences we use an average weighted inverse of the ranking subset size, or a weighting function $w_i = f(i) = 1/|\mathrm{R}_i|$.

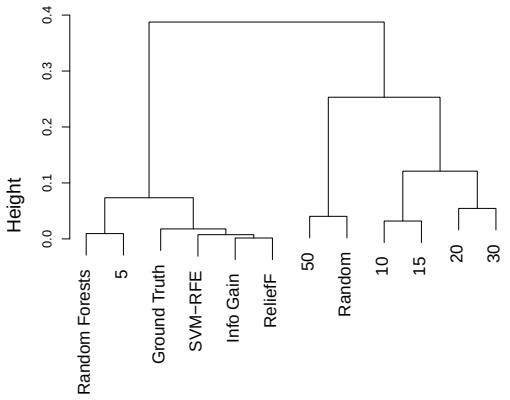
We summarise these calculated ECA differences, by performing hierarchical clustering of the rankings based on their pairwise ECA differences. We present the obtained dendrograms in Figure 14. Each dendrogram refers to a different dataset.

If we first consider the "single" dataset (Figure 14a), we obtain two major clusters. The first one contains all of the feature ranking algorithms, the original ground truth ranking and the ground truth ranking with low level of added noise (5%). Random forests are grouped together with the noisy ground truth ranking with 5% noise as the ranking of the relevant features is slightly unstable. The second cluster groups together the rest of the noisy rankings with 10%, 15%, 20%, 30% and 50% of noise, as well as the random ranking.

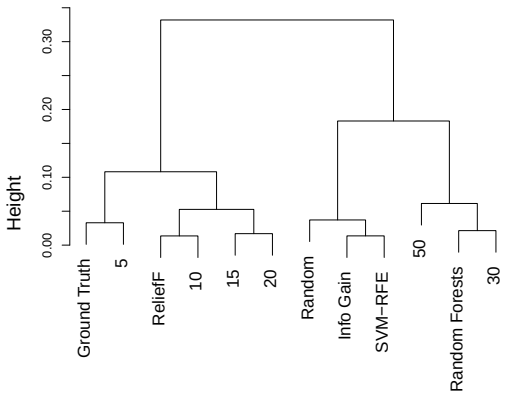
In Figure 14b, we consider the rankings for the dataset that contains just XOR related features. Both information gain and SVM-RFE perform as bad as a random ranking algorithm, i.e., they cluster together with the random ranking. Random forests perform better, but are clustered together with the 30% and 50% noise rankings. ReliefF is clustered together with 10% noise ranking, while closest to the ground truth ranking is the ranking with 5% noise.

This is consistent with the results from Section 5.3. Namely, info gain and SVM-RFE have a completely random behaviour as they are unable to detect the XOR relation. Random forests are able to detect some of the XOR interactions, but are highly unstable and are not the best choice for the "pair" dataset. ReliefF detects most of the relevant features, although their ordering in the final ReliefF ranking is not completely correct, which is why it is clustered together with the 10% noise ranking.

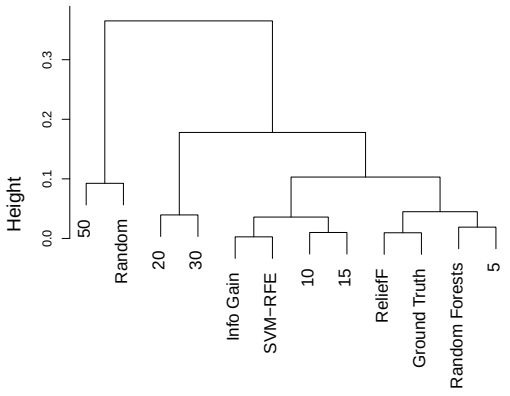
Finally, the last dendrogram in Figure 14c, it contains results for the"combined"dataset. There are four distinct clusters to be considered. First, the random ranking is a clustered together with the 50% noise ranking. It is not similar to the produced rankings of any of the feature ranking algorithms, since all of them manage to identify at least a part of the relevant features. Then the 20% and 30% rankings are grouped together. The information gain and SVM-RFE rankings are together with the 10% and 15% noise rankings, which is expected due to their inability to detect the relevance of the XOR related feature, or put differently, due to their partially random behaviour. Finally, the ReliefF ranking, the Random Forests ranking, the 5% noise ranking as well as the ground truth ranking are all grouped together.



(a) Dendrogram of rankings for the "single" dataset



(b) Dendrogram of rankings for the "pair" dataset



(c) Dendrogram of rankings for the "combined" dataset

Figure 14: Hierarchical Clustering Dendrograms of the rankings produced by the different algorithms based on the ECA differences between them

5.3.5 Conclusions

From a practical perspective, the purpose of the comparative analysis of the different feature ranking methods is to determine which of them is better and how much better. This analysis was performed by using our evaluation method for producing error curves. The outcome of the comparative analysis is also supported by the previous discussion in Section 5.3.2, where the individual feature ranking methods are analysed in terms of the type of the feature rankings they produce.

Both the visual inspection of the error curves (Figure 13) and the dendrograms of the quantitative comparison (Figure 14) suggest the following ordering of the feature ranking algorithms:

$$\textbf{ReliefF} > \textbf{Random Forests} > \textbf{Information Gain}, \textbf{SVM} - \textbf{RFE}$$

where ">" stands for "better than".

Namely, the ReliefF rankings have the best error curves, i.e., the curves that are closest to the ground truth ranking. Therefore it is the best feature ranking method among the ones considered. This is in agreement with the analysis of the individual methods in Section 5.3.2, where it can be seen that ReliefF always produces a well-ordered ranking.

Second come random forests. They are not as good as ReliefF, but according to the error curves, they outperform information gain and SVM-RFE. This is consistent with the type of rankings they produce. Namely, they usually manage to delineate relevant from irrelevant features, or at least distribute the relevant features closer to the top of the ranking. However, due to their instability, they produce noisy rankings in the relevant region. This is reflected in their error curves, which are more distant from the ground truth ranking than those of ReliefF.

The last place is shared by the information gain and the SVM-RFE feature ranking method. According to the error curves, they only have good performance when considering a dataset with features that are associated with the target only individually and have no pairwise or higher order interactions. In all other cases, the error curves indicate that they provide a worse, more noisy ranking then the other methods. This is again consistent with the individual analysis of the ranking methods in Section 5.3.2. Their rankings are completely or partially random. However, their partially random behaviour is different from the one of the Random Forests rankings. They fail to fully delineate relevant from irrelevant features and only provide correct ordering of the individually associated features. For all other features (the XOR related features), they provide random rankings. Therefore, they can be considered as worse ranking methods than Random Forests and ReliefF.

5.4 Summary

The experimental work presented in this section investigated the behaviour of our feature ranking evaluation method. We studied the behaviour under a variety of experimental conditions. All of the experiments were performed in a controlled environment, by generating synthetic data.

Different ground truth rankings were considered, for datasets with different feature interaction structures. Our first experiments were based on uniformly applying noise to the know ground truth ranking. For each different noise level, corresponding error curves (FFA and RFA) were generated. The analysis of the error curve plots shows that the FFA and RFA curves can be used to distinguish between feature rankings of different quality. The quantitative analysis based on the error curves shows that the error curve average (ECA) differences, are sensitive to the added noise in the same way as the actual distances between the feature rankings. Therefore, the ECA differences can be used to properly quantify the difference in quality of feature rankings.

Our further experimental work included the analysis of four different feature ranking methods, namely: information gain, random forests, ReliefF and SVM-RFE. We first assessed the individual feature rankings produced by the feature ranking methods. The analysis provided insight into the type of rankings the different feature rankings produced in terms of their stability and in terms of how well ordered they were. This was then used to establish a connection with the constructed FFA and RFA curves. We then performed a comparative analysis of the feature ranking methods with the help of FFA and RFA curves. The results showed that the overall best rankings, i.e., the rankings closest to the ground truth, were produced by ReliefF. Random forests often delineated relevant from irrelevant features, but produced a noisy ordering of the relevant features. For that reason they were a worse method for feature ranking than ReliefF, but were much better than information gain and SVM-RFE. Both, info gain and SVM-RFE were able to properly rank only the features individually associated with the target, while for the other types of features (XOR related) they provided a random ranking.

In summary the experiments presented in this chapter empirically demonstrate that our evaluation method, proposed in Chapter 4, is able to distinguish between feature rankings of different quality and to quantify the difference between them. This is true regardless of whether the noise is uniformly distributed through the whole ranking or it is located only in certain regions of the feature ranking. In addition using our evaluation methodology, we provided an empirical comparative study of different feature ranking methods.

6 Applied Experiments

In this chapter, we present results of analyses performed with our feature ranking evaluation method on many datasets from different domains. After the analysis of the evaluation method and the empirical study of different feature ranking methods in Chapter 5, we present here an experimental analysis that is more application oriented. The main purpose of the experiments is to demonstrate how our feature ranking evaluation method can be used for practical data analysis.

The experiments are divided in three parts. The first part of the experiments is presented in Section 6.1. They involve synthetic data and aim to investigate the effects of using feature ranking ensembles on the quality of the final feature ranking.

The second part of the experiments concerns different real-world datasets from various domains and is presented in Section 6.2. Our report on the obtained experimental results, includes the behaviour of different feature ranking methods on datasets with different structure and with varying number of features.

The last part of the experimental work, presented in Section 6.3, involves a compendium of gene expression datasets. These datasets are from a specific medical domain, which investigates embryonal tumors. This compendium of datasets provides the possibility of different application-driven experiments and is therefore treated separately from the experiments on other real world datasets from Section 6.2.

6.1 Feature Ranking Ensembles

In this section, we present experimental results involving feature ranking ensembles. The intuition for constructing ensembles of feature rankings is similar to the one for constructing ensembles of predictive models (Dietterich, 2000a). First, the original data are sampled and for each data subsample, a feature ranking is induced. All of the different feature rankings induced from the data samples are then combined into a single ranking.

The aim of our experiments is to answer the practical question of whether and under what conditions, there is an advantage in using feature ranking ensembles. For that purpose, we compare feature rankings induced by a feature ranking ensemble with the ones induced by the individual feature ranking methods. The qualitative comparison of the feature rankings is performed by using FFA and RFA curves.

The experimental design and the experimental conditions are described in Section 6.1.1. The results are presented in two parts, each relating to a different research question about the feature ranking ensembles. In Section 6.1.2, we first present a comparative analysis of different aggregation functions that can be used for combining the individual feature rankings. We then present an analysis of the sensitivity of feature ranking ensembles to the different number of data subsamples. Finally, the conclusions that summarise these different aspects of feature ranking ensembles are given in Section 6.1.3.

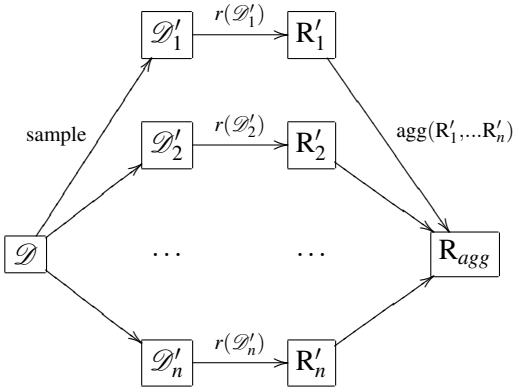


Figure 15: Graphical representation of the process of generating a feature ranking ensemble. First, the data $\mathscr{D}$ are sampled and subset samples $\mathscr{D}'_j$ are produced. From each sample, a feature ranking R'_j is generated by a base ranking method $r(\mathscr{D}'_j)$. At the end, the rankings are combined into a single aggregated ranking R_{agg}.

6.1.1 Experimental Setup

The various experimental scenarios considered in this section aim to establish the conditions under which it would be advantageous to use feature ranking ensembles. The different conditions are related to the various possible choices that can be made during the process of constructing a feature ranking ensemble. The whole process of constructing an aggregated feature ranking from the feature ranking ensemble is represented schematically in Figure 15.

Starting from the original data $\mathscr{D}$, by using a sampling procedure, k data subsets $\mathscr{D}'_i$ are produced. From each data subset $\mathscr{D}'_i$, a single feature ranking, R'_i, is induced by the feature ranking method of choice. Finally, the k rankings are combined into a single aggregated ranking, R_{agg}, by using an aggregation function, $\mathbf{agg}(\mathrm{R}'_1, \ldots \mathrm{R}'_k)$. When following this process of constructing feature ranking ensembles, there are several possible choices to consider.

First, the data subsamples can be created from the original dataset in different ways. In addition, a parameter that can be varied in this process is the number of data subsamples k. For generating the data subsets necessary for constructing the ensembles, from a single dataset $\mathscr{D}$, we use bootstrap sampling. In the process, k data samples are drawn by using sampling with repetition. Each data subsample $\mathscr{D}'_i$ has the same number of instances as the original dataset. Not all instances from the original data will be present in a subsample, and some will be repeated. We considered several values for the number of subsamples k, namely 50, 100, 200, 300.

Different feature ranking methods $r(\mathscr{D})$, can be considered for constructing the base feature rankings R'_i from the data subsamples $\mathscr{D}'_i$. For inducing the baseline feature rankings, we considered the four feature ranking methods used in the experiments in Chapter 5:

- **Information gain**, calculating the information gain of each feature F_i as: $IG(F_t, F_i) = H(F_t) - H(F_t|F_i)$. This does not require any specific parameter setting.

- **SVM-RFE** is the redundant feature elimination (RFE) procedure that employs an SVM to evaluate the feature weights at each iteration. A linear SVM was employed, as proposed by Guyon et al. (2002). The epsilon parameter of the SVM was set to 1.0E-12 , while the complexity was set to 0.1.

- **ReliefF** algorithm as proposed by Robnik-Šikonja and Kononenko (2003). The number of neighbours was set to 10 and all of the instances were used for estimating the relevance values.

- **Random forests**, which can be used for estimating feature relevance as described by

Breiman (2001). A forest of 100 trees was used, constructed by randomly choosing a $\log_2$ of the number of features.

The final choice to when constructing the feature ranking ensembles is how to combine the k base rankings $\mathbf{R}'_i$ into a single aggregated ranking $\mathbf{R}_{agg}$. For that purpose, we employ a weighted aggregation function that determines the aggregated rank of each feature F_i as follows:

$$rank(F_i)_{agg} = \underset{j=1\ldots k}{\text{agg}} \{w_j \cdot rank'_j(F_i)\}.$$

For our experiments we considered four simple aggregation functions that calculate the aggregated rank of each feature F_i as:

- the **mean** value of the base ranks: $rank(F_i)_{agg} = \text{mean}_{j=1\ldots k}\{rank'_j(F_i)\}$
- the **median** value of the baseline ranks: $rank(F_i)_{agg} = \text{median}_{j=1\ldots k}\{rank'_j(F_i)\}$
- the **maximal** value of the baseline ranks: $rank(F_i)_{agg} = \max_{j=1\ldots k}\{rank'_j(F_i)\}$
- the **minimal** value of the baseline ranks: $rank(F_i)_{agg} = \min_{j=1\ldots k}\{rank'_j(F_i)\}$

As input to the feature ranking ensemble process, we considered the three synthetic datasets described in Section 5.1, namely: "single", "pair" and "combined". These datasets are useful for feature ranking ensembles analysis, because they have different but known feature interaction structures. This allows for the feature ranking ensembles to be studied w.r.t. different feature interaction structures.

For each of the datasets and for each of the previously described feature ranking ensembles settings, we constructed feature ranking ensembles. The output from each experiment is a single aggregated feature ranking and its respective FFA and RFA curves. These curves are compared to the FFA and RFA curves of feature rankings induced from the whole original dataset. The feature rankings are induced by some of the four feature ranking methods, with the same settings as described in the previous text. The aggregated feature ranking that is compared with the single feature ranking is always induced with the same baseline feature ranking method as the single feature ranking.

To generate the error curves, SVMs were employed as classifiers, with polynomial (quadratic) kernel. The epsilon parameter was set to 1.0E-12, while the complexity was 0.1. For estimating the stability of each feature ranking algorithm the stability indicator described in (Jurman et al., 2008) was used.

6.1.2 Results

Here, we present the results of the different experimental settings for constructing feature ranking ensembles, described in the previous section. The results are graphs that include either FFA or RFA curves. On each graph, three types of error curves are compared, namely: the error curves of feature ranking ensembles, the error curves of individual feature ranking methods, as well as the expected error curves of random rankings that are generated as described in Section 5.2.

The graphs are organised to account for two different types of comparisons. The first type of comparison that we consider is that of different aggregation functions, used for combining the baseline rankings into a single aggregated ranking. The second type of comparison is a sensitivity analysis of the impact of the number of data subsamples k, on the quality of the produced rankings.

The full set of results, with all of the comparison are given integrally in Appendix B.1. In this section, we present just a selection of the results where the FFA and RFA curves of feature ranking ensembles, are different from those of an individual feature ranking method. For both discussed types of comparisons, the presented results are limited to using random

forests as a feature ranking method, as it was the only method for which feature ranking ensembles had influence on the quality of the induced ranking.

Comparison of Different Aggregation Functions

In Figure 16, we present the results of comparing different aggregation functions, when using random forests as a baseline ranking method. Additionally, we compare just the feature rankings when the number of data subsamples k is 300, i.e., the maximal considered value of k. The graphs are organised in two columns and three rows. The first column contains plots of FFA curves, while the second column contains RFA curves. Each row refers to a different dataset in the following order: "single" dataset, "pair" dataset and "combined" dataset.

From a quick overview of the graphs, a general observation that can be made is that feature ranking ensembles induce a better feature ranking as compared to the individual ranking, produced by constructing a single random forest on the whole dataset. To begin with, there is a slightly noticeable difference in the FFA/RFA curves when considering the "single" dataset, as seen in Figure 16a and Figure 16b. This slight difference is expected, as the detailed analysis in Section 5.3, shows that the ranking induced by an individual random forest is just slightly different than the ground truth ranking. Additionally, there is no visible difference in the error curves, when considering different aggregation functions.

This improvement in the feature ranking, is slightly bigger in the case of the "combined" dataset, according to the error curves in Figure 16e and Figure 16f. This is again consistent with the results from Section 5.3, which show that random forests successfully delineate relevant from irrelevant features, but are unstable in this relevance region. In this case it also applies that there are no noticeable differences in the error curves, when considering the different aggregation functions.

Finally, the biggest improvement can be seen when considering the results for the "pair" dataset, in Figure 16c and Figure 16d. The feature ranking ensembles induce an aggregated feature ranking, that drastically improves the quality of the feature ranking, as reflected in the FFA and RFA curves. Unlike in the previously discussed figures, there is a noticeable difference between the error curves of different aggregation functions. When using the mean or median aggregation function, the improvement is identical and it is biggest as compared to the other two methods. Using the min aggregation function is slightly worse, but still provides considerable improvement over the individual feature ranking. The smallest improvement of the feature ranking occurs when using the max aggregation function.

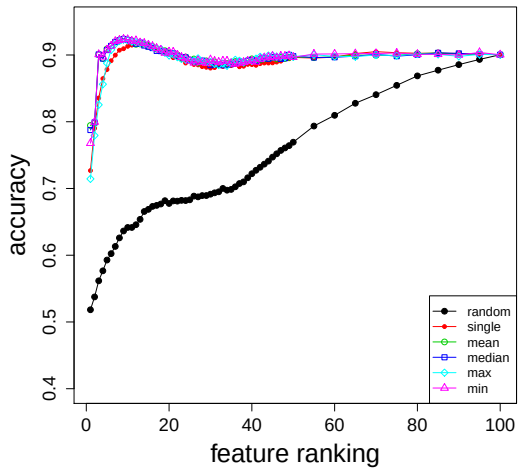
The explanation why feature ranking ensembles improve the quality of the feature ranking just for random forests, can be understood by considering the previous analysis of the individual feature ranking methods from Section 5.3 and the kind of feature rankings they provide. First, completely random rankings, as provided by info gain and SVM-RFE for the "pair" dataset, can not be improved as an aggregation of random baseline rankings again results in a random ranking. This also applies for partially random rankings produced by info gain and SVM-RFE for the "combined" dataset. The XOR interacting features from the data, have a random rank assigned in all subsamples and therefore the final aggregated rank assigned to these features is random. ReliefF for all datasets, as well as info gain and SVM-RFE for the "single" dataset, provide a well ordered ranking and there is no room for improvement by the feature ranking ensembles.

Random forests, according to the discussion in Section 5.3, are successful at mostly delineating the relevant from irrelevant features, but are unstable which results in an erroneous ranking of the relevant features. We hypothesise that feature ranking ensembles have a profoundly stabilising effect on random forests and the final aggregated ranking is therefore better w.r.t. the individual one.

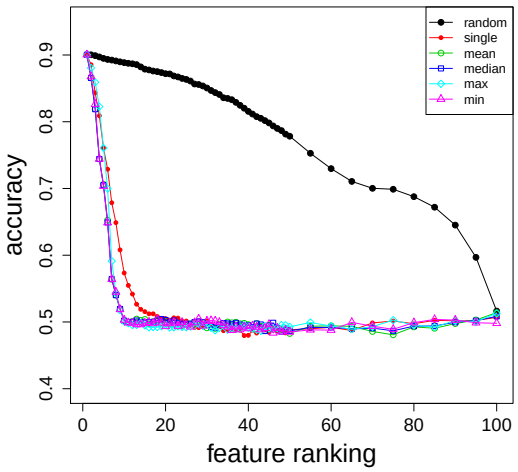
This can be seen in Figure 17, where we compare the different stability patterns of random forests. Each graph compares the stability of the feature rankings constructed by an

individual random forest with the stability of the aggregated ranking produced with different aggregation functions. It can be seen in the graphs that all of the aggregation functions reduce the instability of the produced feature rankings compared to the individual feature ranking. Additionally, when using the mean or median aggregation function, the stability pattern is also completely changed. Instead of the instability peak area in Figure 17a and Figure 17c, there is a region where the feature ranking is very stable. The most obvious stability improvement is in the case of the "pair" dataset seen in Figure 17b. Here, the stability pattern is completely inverse of that of the individual feature ranking.

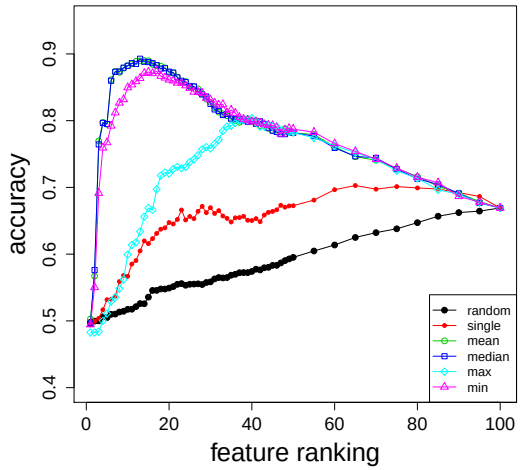
The min aggregation functions, provides a slightly more unstable ranking than the one provided by mean and median aggregation functions. In the case of the "pair" dataset, seen in Figure 17b, it provides a stability pattern with an instability peak in the area of the relevant features of the ranking. This is reflected in a slightly worse FFA and RFA curves, as seen in Figure 16c and Figure 16d. Finally, the stability pattern of the max aggregation function is divided in two parts, a region of complete stability, where the value of the stability indicator equals zero and a region of rising instability.



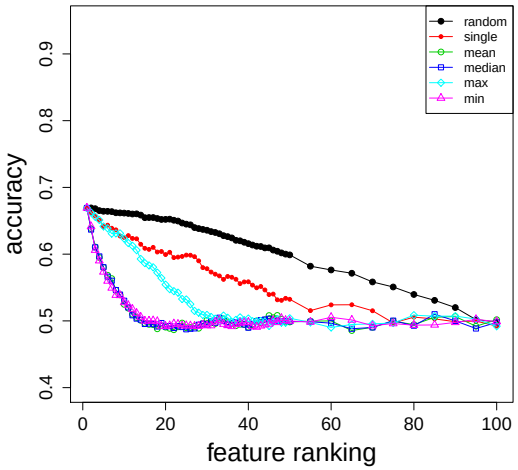
(a) FFA curves of the "single" data.



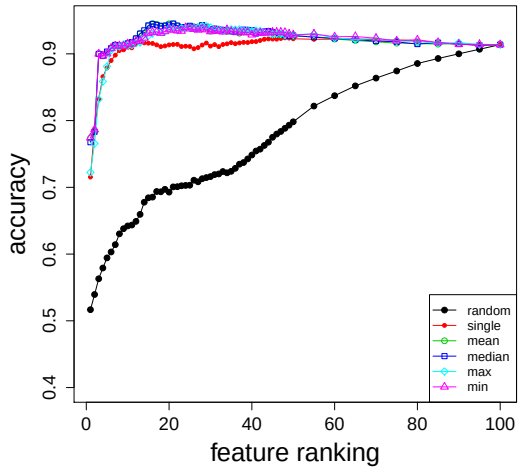
(b) RFA curves of the "single" dataset



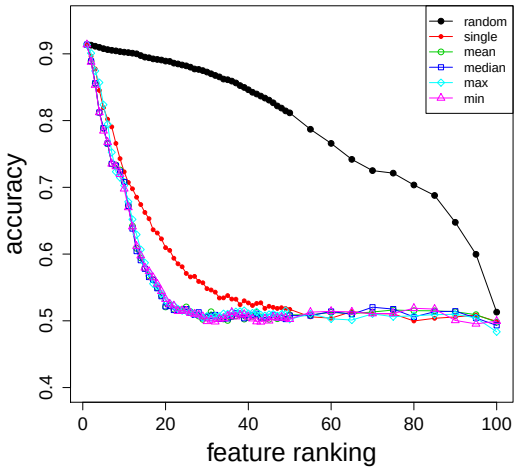
(c) FFA curves of the "pair" data.



(d) RFA curves of the "pair" data.

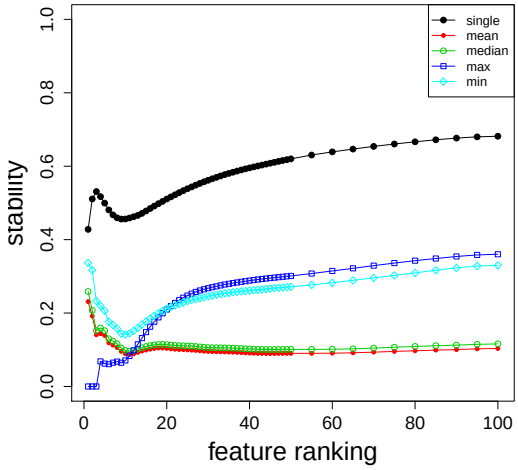


(e) FFA curves of the "combined" data.

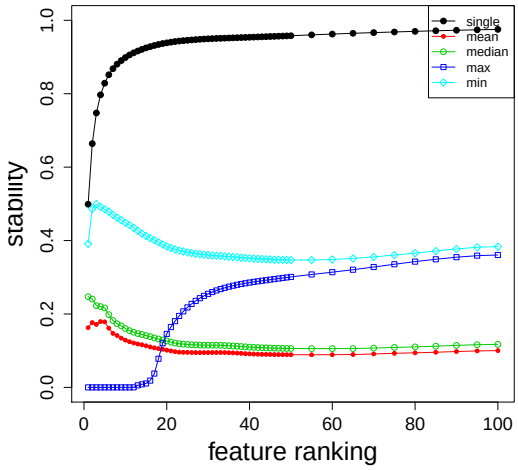


(f) RFA curves of the "combined" data.

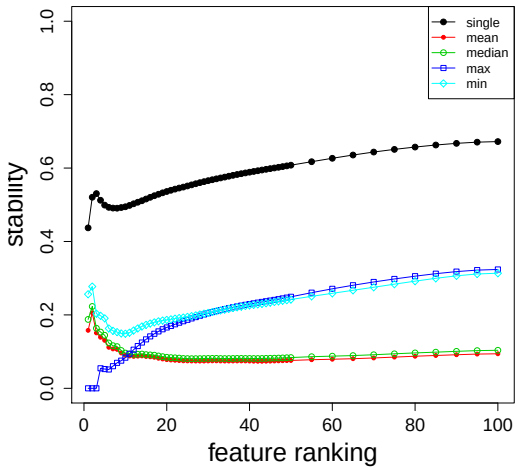
Figure 16: Figures representing the comparison of the FFA (left column) and RFA (right column) curves of different aggregation functions. The baseline ranking is random forests and the number of samples k is 300. Each figure also contains plots of FFA/RFA curves of an individual ranking and the random FFA/RFA curves.



(a) “single” dataset

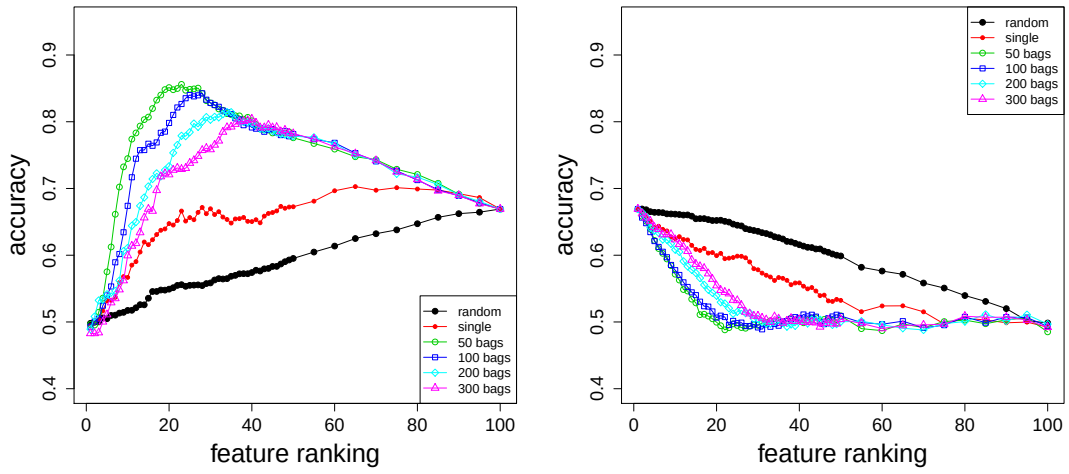


(b) “pair” dataset



(c) “combined” dataset

Figure 17: Stability comparison graphs of different aggregation functions, with random forests as baseline rankers.



(a) FFA curves comparison

(b) RFA curves comparison

Figure 18: Figures representing the comparison of the FFA (left) and RFA (right) curves obtained by varying the number of data samples k. The baseline rankers are random forests and they refer only to the "pair" dataset

Comparison of Different Number of Data Subsamples

In this section we consider the results from the sensitivity analysis, related to the number of data subsamples k. In general, it did not reveal any dependency between the number of data subsamples k and the improvement of the quality of the aggregated feature rankings, except in one specific case. This specific case occurs when considering random forests as base ranking methods, with the max aggregation function and only on the "pair" dataset.

In Figure 18, we present the FFA and RFA curves only for the previously decribed conditions, while the full set of results can be seen in Appendix B.1. By varying the value of k, both the FFA and RFA curves, behave opposite of the intuitively expected result. Namely, although feature ranking ensembles improve the quality of the feature ranking, they demonstrate that as the number of data subsamples k increases, the aggregated ranking quality is worsening.

The underlying reason for this is the instability of the random forests, as well as the fact that the max aggregator favours high rank values. Namely, this influences the quality of the aggregated ranking in the following way: for a given k, because of the instability of the random forests, there is a probability that a non-relevant feature will be randomly appointed a high(er) rank for at least one of the data subsamples. As the max aggregator prefers higher ranks, in the final aggregated ranking this non-relevant feature will be awarded a high rank value. As the number of data subsamples k increases, there is a greater probability that more and more non-relevant features will be awarded a higher rank in the final ranking.

6.1.3 Conclusions

In our experiments in this section we explored different experimental settings for constructing feature ranking ensembles. This included different baseline feature ranking methods, different functions for aggregating the baseline feature rankings and datasets with different interaction structures. Overall, from the results presented, it can be concluded that feature ranking ensembles have a limited effect on improving the quality of feature rankings.

The benefit is only apparent when using random forests as a baseline feature ranking method. For random forests, the improvement of the quality of the feature rankings, was

consistent for the feature ranking ensembles under all the other considered experimental settings. For all the other feature ranking methods, there was no visible effect when using feature ranking ensembles to produce an aggregated feature ranking.

The comparison of different aggregation functions, revealed that the mean and median provide identical results in all of the experiments. They can be considered as the optimal aggregation methods from the ones tested, because they provide a stable feature ranking and also a well ordered one, which is visible from the constructed FFA and RFA curves. The sensitivity analysis provided by varying the number of data subsamples k, reveals that a small number of k (50) is sufficient to produce a better aggregated ranking. In the case of the max aggregation function, a smaller number of k is even preferred to a bigger one, as it produces a better feature ranking.

In summary, feature ranking ensembles provide a benefit and should be preferred only when using methods that delineate relevant from irrelevant features, but are unstable. Their major advantage is that they provide a stabilising effect, which is beneficial even when using a small number of data subsamples, irrespective of the aggregation function used. From the feature ranking methods considered, only random forests provided feature rankings that can take advantage of this strength of feature ranking ensembles.

6.2 Experiments in Different Domains

So far, all of our analysis considered data that was artificially generated. In this section, with the help of our feature ranking evaluation method, we analyse datasets originating from various real-life domains. The purpose of the experiments is to examine the quality of the feature rankings produced by several feature ranking methods, on data with different characteristics.

The analysis is primarily a comparative one, performed solely with generating FFA and RFA curves and the numeric scores derived from them. The datasets we consider are quite diverse, with unknown interaction structure and therefore unknown ground truth ranking. However, for each datasets it is possible to generate the expected error curves of random rankings. The expected curves are used as a baseline for comparing the different feature ranking methods.

The detailed description of the datasets and their important characteristics are given in Section 6.2.1. The full experimental setup used for the analysis of the different datasets is given in Section 6.2.2. We report on the obtained results from our experiments in Section 6.2.3 and we present the conclusions in Section 6.2.4.

6.2.1 Datasets Description

For our experiments, 23 diverse datasets were selected. Most of them (20) originate from the UCI data repository (Newman and Merz, 1998) and are from various domains. From the remaining 3 datasets, one is from a medical study of acute abdominal pain in children (aapc) (Džeroski et al., 1997), while the remaining two ("water" and "diversity") are from an ecological study of water quality of rivers (Džeroski et al., 2000).

Besides covering different domains (including biology, medicine, ecology etc.) these datasets have a wide range of different properties. All of them are classification datasets with a single target class. The target class has a varying number of possible values (2–7). The features that were ranked w. r. t. the target class, were either discreet or/and numeric, ranging from 4 to 90 features. The datasets also have a considerable range of the number of instances from 150 up to 5000.

As the analysis in this section is a comparative study for different feature ranking methods, we refrain from further detailed description of each dataset. However, we summarise

Table 5: Statistics for datasets from various domains

Dataset	#Inst.	#Feat.	(D/N)	#Cl.
aapc	335	84	(83/1)	3
australian	690	14	(8/6)	2
balance	625	4	(0/4)	3
breast-cancer	286	9	(9/0)	2
breast-w	699	9	(9/0)	2
car	1728	6	(6/0)	4
chess	3196	36	(36/0)	2
diabetes	768	8	(0/8)	2
diversity	292	86	(0/86)	5
german	1000	20	(13/7)	2
heart	270	13	(6/7)	2
heart-c	303	13	(7/6)	5
heart-h	294	13	(7/6)	5
hepatitis	155	19	(13/6)	2
image	2310	19	(0/19)	7
ionosphere	351	34	(0/34)	2
iris	150	4	(0/4)	3
sonar	208	60	(0/60)	2
tic-tac-toe	958	9	(9/0)	2
vote	435	16	(16/0)	2
water	292	80	(0/80)	5
waveform	5000	21	(0/21)	3
wine	178	13	(0/13)	3

all of the important characteristic of the various datasets in Table 5, where their diversity can be clearly seen.

6.2.2 Experimental Setup

As mentioned in the previous text, the purpose of our experiments was to perform a comparative analysis of various feature ranking methods on a wide range of real-life datasets. The experimental setups for this kind of analysis was quite simplistic. Namely, for each of the described datasets in Section 6.2.1, four feature ranking methods were used for inducing ordering of features. For each feature ranking, both FFA and RFA curves were induced and compared between each other. Also, as a baseline of the comparison, expected FFA and RFA curves were used.

The four feature ranking methods that were examined were the same ones that were used in our previous work, namely:

- **Information gain**, calculating the information gain of each feature F_i as: $IG(F_t, F_i) = H(F_t) - H(F_t|F_i)$. This does not require any specific parameter setting.
- **SVM-RFE**, is the redundant feature elimination (RFE) procedure that employs an SVM to evaluate the feature weights at each iteration. A linear SVM was employed, as in (Guyon et al., 2002). The epsilon parameter of the SVM was set to 1.0E-12 , while the complexity was set to 0.1.
- **ReliefF** algorithm as given in (Robnik-Šikonja and Kononenko, 2003). The number of neighbours was set to 10 and all of the instances were used for estimating the relevance values.

- **Random forests**, which can be used for estimating feature relevance as described in (Breiman, 2001). A forest of 100 trees was used, constructed by randomly choosing $\log_2$ of the number of features.

For estimating the error values necessary for generating the FFA and RFA curves, SVMs with polynomial (quadratic) kernel were used and 10-fold cross validation was performed, on the dataset under consideration. The epsilon parameter of the SVMs was set to 1.0E-12, while the complexity was 0.1.

The expected error curves were induced by generating 100 random rankings for each dataset under consideration. For each random ranking, error curves were induced and the average of the error values was used as the expected error. This was done in a similar manner as described in Section 5.2.1.

6.2.3 Results

The output of the experiments described in the previous section are FFA and RFA curves of different feature ranking methods, generated for each separate dataset. For practical purposes the graphs that can be used for visual inspection of the results are given in Appendix B.2. In this section, we present only the numeric summary of the differences between the error curves, or more precise the error curves average (ECA) differences.

The ECA differences are calculated as described in Section 5.2.3, by using Equation 24 and a weighting function $w_i = 1$, i.e. a standard mean value. The results are summarised in Table 6. Each row of Table 6 refers to a single dataset, while each column corresponds to a single feature ranking method. The ECA values in the table are calculated w.r.t. the baseline error curve, namely, the expected error curve. This gives an indication of how much each feature ranking method is better than a random ranking generator, but also allows comparison between the quality of the feature rankings of the different methods.

The positive values of the ECA differences indicate that a feature ranking method performs better than a random rankings generator. The negative values, however, do not indicate that it performs worse than random, but that it provides a non-random ranking that is inverse of the correct one. A value close to zero, means the feature ranking method provides rankings that are more random. The missing values for the SVM-RFE method are due to the implementation of SVM-RFE in Weka (Hall et al., 2009), which did not support feature ranking when the number of classes of the target is greater than two.

If we examine the results in Table 6, an initial observation would be that random forests often have negative ECA values. The FFA and RFA curves of random forests, for these particular datasets seen in Appendix B.2, are below or over the expected FFA and RFA curves of random rankings. Upon closer inspection of their feature rankings (results not shown) they are inverse that those of the other feature ranking methods.

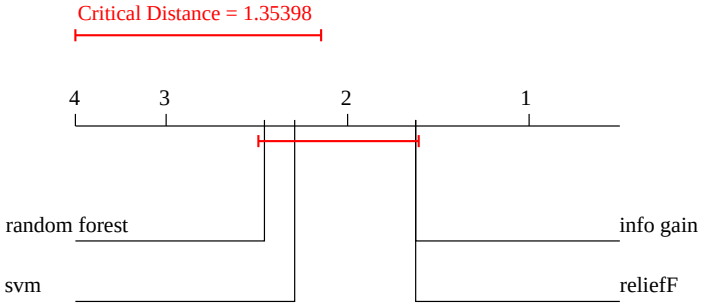
In order to summarise the results from Table 6 and to draw meaningful conclusions about the performance of different ranking methods across different domains, we use statistical tests. We adopt the recommendations by Demšar (2006) for the statistical evaluation of the results. We use the Friedman test (Friedman, 1940) for statistical significance with the correction from Iman and Davenport (1980). Afterwards, to check where the statistically significant differences appear (between which feature ranking methods), we use the Nemenyi post-hoc test (Nemenyi, 1963).

We present the results from the statistical analysis with critical distance diagrams (Demšar, 2006) in Figure 19b and Figure 19a. In the diagrams the feature ranking methods are ordered according to which one is better on average (across all datasets). A method is better if it is positioned closer to the value one on the axis.

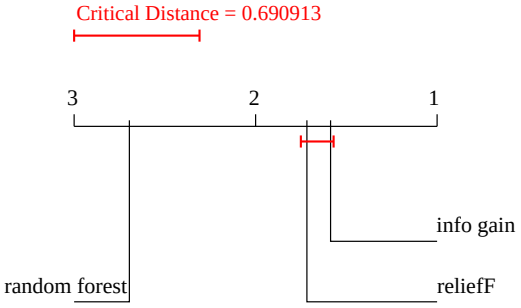
The lines drawn on the diagram, connect the feature ranking methods whose performance is not statistically significantly different. The length of the line connecting the methods,

Table 6: ECA differences calculated between the FFA/RFA curves of various feature ranking methods w.r.t. the curves of a random ranking. The ommited values, marked by "-" are where SVM-RFE could not produce results due to a multi-class target.

Dataset	**Info Gain**	**Random Forest**	**ReliefF**	**SVM-RFE**
aapc	0.09	0.07	0.09	-
australian	0.12	0.08	0.12	-
balance	-0.03	0.03	-0.02	0
breast-cancer	0.01	0	0.01	-
breast-w	0.01	0	0.01	-
car	0.02	-0.02	0.02	-
chess	0.15	-0.05	0.15	-
diabetes	0.05	-0.03	0.05	0.05
diversity	0.03	0.03	0.04	0.02
german	0.02	-0.01	0.01	-
heart	0.05	-0.03	0.04	-
heart-c	0.05	-0.04	0.04	-
heart-h	0.06	-0.01	0.05	-
hepatitis	0.01	0	0.01	-0.03
image	0.02	0	0.02	0.03
ionosphere	0.05	0.05	0.02	0.06
iris	0.05	0.05	0.04	0.04
sonar	0.02	0.03	0.03	0.02
tic-tac-toe	0.03	-0.02	0.03	-
vote	0.07	0.07	0.07	0.07
water	0.04	0.03	0.04	0.02
waveform	0.02	-0.05	0.03	0.03
wine	0.04	0.03	0.04	0.04



(a) Critical distance diagrams calculated for all four algorithms



(b) Critical distance diagrams calculated for three algorithms

Figure 19: Critical distance diagrams representing the statistical comparison of the ECA differences of the various ranking methods. The critical distance is calculated for a p value of 0.05 and is represented by a horizontal line. If the feature ranking methods are connected by a line, then their performance is not statistically significantly different.

is always smaller than the calculated critical distance. The value of the critical distance depends on the level of the desired statistical significance, which in our case is 0.05.

The first diagram (Figure 19a) is comparing all of the feature ranking methods, however only on part of the datasets that include results from SVM-RFE. As it can be seen, all of the methods are connected by a line. This means, that on the subset of data considered, on average all the methods performed equally well, i.e. produced rankings with similar quality.

However, if we exclude SVM-RFE from the comparison and perform the statistical tests on all of the 23 datasets, we obtain the diagram from Figure 19b. On this diagram it can be seen that info gain and reliefF, on average outperform random forests. Additionally, info gain and reliefF are connected by a line, which means that there is no statistically significant difference between their performance.

6.2.4 Conclusions

In this section, we performed a comparative analysis of different feature ranking methods, while considering a variety of real-life datasets. The data encompassed multiple domains and had a wide range of different properties. The analysis included comparison of error curves produced with our feature ranking evaluation method.

The results showed some interesting behaviour of random forests, which at certain cases provided ranking that was inverse of the other methods. This was visible on the error curves, as well as from the numeric summary of the error curves in form of ECA differences. Here, we simply report on this fact and do not explore it further.

The statistical analysis of the ECA differences between all of the methods, on a subset of the 23 datasets, showed no statistically significant difference in performance of the methods. However, when SVM-RFE was excluded from the analysis and all of the 23 datasets were used, it showed that random forests were outperformed by the other two methods. ReliefF and info gain, showed no statistically significant difference in performance. This means that across the different datasets considered, a simple method of feature ranking as info gain is sufficient for producing a well-ordered feature ranking.

6.3 Embryonal Tumors Expression Data Experiments

The experimental work presented in this section is application oriented and it involves data from a very specific medical domain. The purpose of the experiments is to illustrate how our feature ranking evaluation method can be used as an integrative part of a knowledge discovery scenario. The application area of the results is cancer research, or more precisely embyronal tumours research.

The term "embryonal tumours" encompasses several different tumour entities (types) that appear in young children between the age between 0-6 years. Each tumour entity has been individually researched and the knowledge and data generated so far, has opened up the possibility of a more integrative embryonal tumours analysis. In the work presented here, we analyse the embryonal tumours both individually and integratively. We consider the integrative analysis from the aspect of searching for common genes for all tumour entities that are part of a mechanism provoking tumour aggressiveness.

In Section 6.3.1 we describe the specific datasets used for the analysis, while in Section 6.3.2, we provide a more detailed account of the problem under consideration. We divide the presentation of the experimental work in two separate sections. In Section 6.3.3, we provide individual analysis of each different tumour entity. The second part of the experiments, given in Section 6.3.4, concerns with the combined analysis of the different tumour entities. The conclusions of the overall analysis are given in Section 6.3.5.

6.3.1 Datasets Description

As previously mentioned, the analysis performed in this section is from the domain of cancer research, specifically embryonal tumours (ET). They are childhood malignancies that account for 30 % of cancer cases in children. Embryonal tumours include different tumour entities, depending on the tissue or organ where they occur. They are divided on six types of entities, namely: neuroblastoma (NB), nephroblastoma or Wilms' tumour (WT), medulloblastoma (MB), retinoblastoma (RB), the Ewing sarcoma family of tumours (EWS) and rhabdoid tumours (RT).

The collection of ET data used for our experiments consisted of 10 gene expression microarray datasets. All of the ET entities were included in this collection, except for the retinoblastoma (RB) tumour. Each datasets consists of numeric values, which provide quantitative information of how active (expressed) is a certain gene. The data was made available for analysis, as part of a collective effort in the FP6, E.E.T.-Pipeline project (LifeSciHealth-2005-037260).

The individual datasets statistics are given in Table 7. The breakdown of contents of the ET data collection is the following:

- one dataset of ewings sarcoma, denoted as "ews12102" (Scotlandi et al., 2009)
- three datasets of medulloblastoma, denoted as: "mb10327" (Kool et al., 2008), " mb12992" (Fattet et al., 2009) and "mbDKFZ"(E.E.T.P., 2007-2009)
- two datasets of neuroblastoma, denoted as: "nbCol251" (Oberthuer et al., November 1, 2006) and "nbEssen" (Schramm et al., 2005)

Table 7: Embryonal tumors datasets statistics. Number of instances and entity type are presented.

Dataset	#Inst.	ET Entity
ews12102	37	EWS
mb10327	62	MB
mb12992	40	MB
mbDKFZ	47	MB
nbCol251	251	NB
nbEssen	70	NB
rtCurie	33	RT
wt10320	144	WT
wt11024	27	WT
wtETABM53	60	EWS

- one dataset of rhabdoid tumours, denoted as "rtCurie" (E.E.T.P., 2007-2009)
- three datasets of wilms' tumour, denoted as: "wt10320"(Huang et al., 2009), "wt11024" (Kort et al., 2008) and "wtETABM53" (Corbin et al., 2009).

6.3.2 Problem Description

Besides their site of origin, the ET entities have a lot of differences in terms of their clinical course and aggressiveness. However, there are genes whose expression, or more precisely over-expression, is linked to higher incidence of more aggressive tumour entities. Specifically, we focus on N-MYC and c-MYC genes, as indicators of tumour aggressiveness. As implicated in (Westermann et al., 2008), it is sufficient if only one of the genes is over-expressed for the tumour to develop aggressively.

Both for research and clinical purposes it is relevant to discern the mechanism through which MYC genes regulate the processes in the cell, leading to aggressive tumour development. The first step towards discovering this mechanism is to identify the key genes involved in this process. Taking into account the data available, we map this problem of discovering the key genes, into a problem of feature ranking.

As the datasets from the ET collection are microarray expression data, they measurements of the expression of various genes, which includes the c-MYC and N-MYC genes. Considering this, finding the genes involved in tumour aggressiveness, becomes equivalent to finding the genes whose expression is related to the expression of either the c-MYC or N-MYC genes. This relatedness goes beyond simple individual correlation, as genes can be tied in functional modules that are collectively influenced by the MYC genes.

From this description of the biological problem, it is easy to map it to a feature ranking problem. Namely, finding the set of all genes (features) that are related or relevant w.r.t. MYC genes, would be the exact output of feature ranking method that solves at least the minimal-optimal problem (Nilsson et al., 2007). As a target gene for the feature ranking we are interested in both the expression of the c-MYC and N-MYC genes.

More specifically, we are interested if any of them is over-expressed, as in (Westermann et al., 2008) it is shown that one of them needs to be over-expressed for the tumour to be aggressive. For that purpose, we derive an indicator of aggressiveness from the expression of both the c-MYC and N-MYC genes, which we use as a target for feature ranking. This derived indicator of MYC expression, represents the maximum of the expression of either the c-MYC or N-MYC gene, namely: $exp(\mathrm{MYC}) = \max\{exp(\mathrm{c-MYC}), exp(\mathrm{N-MYC})\}$.

In practise, the various feature ranking methods available and the various datasets provide different lists of important genes for tumour aggressiveness. By comparing the gene

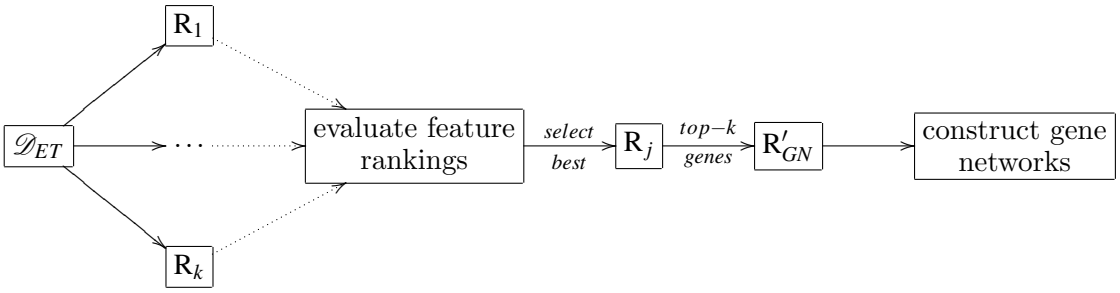


Figure 20: Generic knowledge discovery scenario that can be used in analysis of gene expression data. First, various ranked gene lists $\mathbf{R}_j$ are generated and each is evaluated by our evaluation method for feature rankings. After the best method is determined, the top-k genes are selected and further used for gene network construction

lists with our feature ranking evaluation method, we provide a way to determine which ordered set of genes is most reliable. This genes can be further used for different biological, knowledge discovery scenarios and as an illustrative example we use them to re-construct gene networks.

6.3.3 Individual Embryonal Tumor Datasets

In this section, we present an individual analysis of the different ET datasets. The purpose of the specific knowledge discovery scenario that we consider is two-fold. First, is discovering the key genes involved in tumour aggressiveness. Second, is to infer possible connection between the identified genes, as well as between them and MYC genes.

The whole knowledge discovery process is represented schematically in Figure 20. It basically consists of two consecutive steps. It begins by comparing the gene lists induced by different feature ranking methods, by constructing FFA and RFA curves. Once the best feature ranking method is determined, a subset of genes is selected that consists of the top-k genes from the ordered gene lists. This subset of genes, acts as an input to gene network reconstruction tool, which constructs gene networks by using a database of previous biological knowledge.

Experimental Setup

Here we give the specific details of the experimental setup used for the previously described knowledge discovery scenario from Figure 20. It should be noted that as the $exp(\mathbf{MYC})$ value is numeric and this influenced the further choice of the feature ranking methods used, as well as of the error measure used for constructing the FFA and RFA curves.

For the first part of the knowledge discovery process, we considered just two feature ranking methods for comparison- random forests and reliefF, as they both proved most successful in the synthetic data analysis. The specific setting we used for each of the ranking method are the following:

- **RReliefF** algorithm as given in (Robnik-Šikonja and Kononenko, 2003). This is algorithm is an adaptation of the reliefF algorithm, suited for regression, i.e., for numeric targets. The number of neighbours was set to 10 and all of the instances were used for estimating the relevance values.

- **Random forests**, which can be used for estimating feature relevance as described in (Breiman, 2001). A forest of 100 regression trees was used, constructed by randomly choosing 10% of the number of features.

Dataset	Random Forest	ReliefF
ews12102	0.07	-0.05
mb10327	-0.03	0.06
mb12992	-0.02	0.11
mbDKFZ	0.02	0.17
nbCol251	0.01	0.09
nbEssen	-0.05	0.2
rtCurie	0.05	0.18
wt10320	-0.04	-0.01
wt11024	0.09	0.1
wtETABM53	-0.03	-0.01

Table 8: ECA differences calculated between the FFA/RFA curves of various feature ranking methods w.r.t. the curves of a random ranking.

Additionaly, we generate expected error curves for each datasets, by generating 100 random rankings. For each random ranking, error curves were induced and the average of the error values was used as the expected error. This was done in a similar manner as described in Section 5.2.1.

The feature rankings were compared by constructing FFA and RFA curves. Since the target was numeric, instead of estimating the classification error, we estimated for each point of the curves, the correlation coefficient between the true and the predicted values. For generating the predicted values SVMs for regression with polynomial (quadratic) kernel were used and 10-fold cross validation was performed, on the dataset under consideration. The epsilon parameter of the SVMs was set to 1.0E-12, while the complexity was 0.1.

For constructing the gene networks, we consider a subset of top-50 ranked genes from each list. We choose this specific number of top genes based on the stability analysis, whose details are given in the following text. The top ranked genes are provided as input of the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) (Szklarczyk et al., 2011) version 9.0. This search tool utilises various databases of biological experiments, as well as text mining, to re-construct the gene networks and gives them in an appropriate visual format.

Results

Here, we present the results of the individual analysis of the ET datasets. Following the knowledge discovery scenario from Figure 20, we first present the results from the comparative analysis of the different feature ranking methods. For the comparison, we consider the FFA and RFA curves of the different methods, as well as their stability patterns. Then we present and compare the re-constructed gene networks for the different methods.

The summary of the comparison of the FFA and RFA curves is given by calculating the ECA differences, presented in Table 8. The full set of graphs of the FFA and RFA curves are given in Appendix B.3. It should be noted here that the ECA differences from Table 8, are calculated with respect to the expected FFA/RFA curves of the random rankings.

If we examine in detail the ECA differences from Table 8, we can conclude that most of the time, for 7 out of 10 datasets, RReliefF have a bigger ECA difference than random forests. This means that the FFA and RFA curves of ReliefF, are more distant from the expected curves than random forests and therefore the ranking provided by ReliefF is of better quality.

Before proceeding to the next step of re-constructing gene networks, we also examine the stability patterns of the feature ranking methods, given in Figure 21a and Figure 21b. On each graph, the stability of a single feature ranking method is considered and the different

stability curves refer to different datasets. The y-axis of the graphs, gives the estimated stability value, while the x-axis gives the feature subset size. It should be noted that bigger values mean that the ranking method is less stable.

If we examine the stability patterns, we can observe that they are quite different between random forests (Figure 21a) and RReliefF (Figure 21b), but seem to produce the same stability pattern across all of the different ET datasets. Random forests, in Figure 21a, have an interesting double-peak stability curves. RReliefF has a single instability peak at the beginning of the ranking and then the instability decreases and maintains more or less the same value.

For re-constructing gene networks involved in the mechanism of tumour aggressiveness, we require as input only a part of the ranked genes list. It would make most sense to take the top-k ranked genes, as they are determined by the ranking method as most relevant w.r.t. the MYC genes. Beside relevant, the top-k genes output by the ranking method also have to be a stable set. This insures that the top-k genes are relevant to a target, but also that the ranking method is certain in their level of relevance.

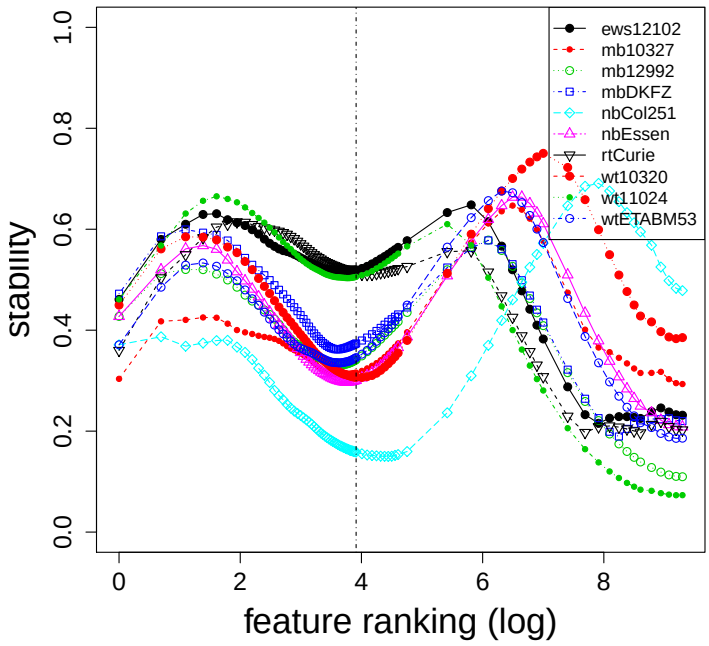
If we again consider the stability patterns from Figure 21a and Figure 21b, we can determine the best value for k, optimal for both ranking methods. For random forests, the optimal cutoff point would be the saddle point between the two instability peaks, marked by a vertical line in Figure 21a. For RReliefF this point would be located in the area after the instability peak, where the values of the curve are virtually constant. We round this number of k to 50 genes.

After evaluating which feature ranking method has an overall better performance and determining the cut-off point 50 top ranked genes, we present the results of gene network reconstruction. For each dataset and for both of the ranking methods, the networks are constructed from the top-50 genes by using STRING. For practical purposes, here we include just one representative example in Figure 64, while the full set of gene networks can be seen in Appendix B.4.

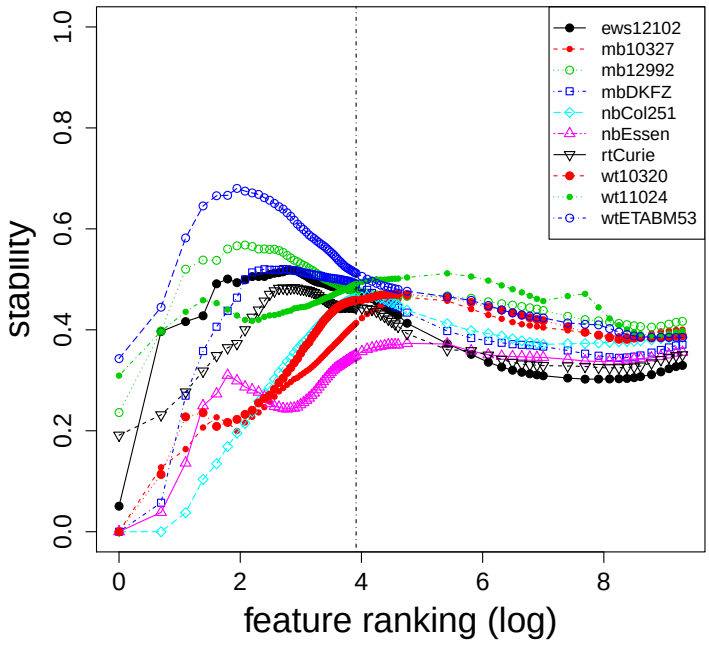
A gene network consists of edges that represent genes (or proteins) and vertices which signify that some kind of connection between two genes exists. This connection can be either proven in previous experiments or derived by text mining of medical articles. The genes that have no connections with any other of the input genes, are discarded from the final graph. When constructing the networks additionally the c-MYC and n-MYC genes were included, which makes sense considering the ranking is generated w.r.t. to the maximum of their expression.

If we examine the networks in Figure 64 it can be immediately noticed that the networks constructed from RReliefF and from random forests are very different. The one constructed from the ranked gene list of ReliefF, contains much more genes (edges) that are highly connected among themselves as compared to the network of random forests. This means that a lot of the genes provided by ReliefF as important, have already been shown in previous biological knowledge, to be influenced by the activity of c-MYC and n-MYC. This difference in the connectedness of gene networks is also visible in all the other re-constructed networks for the different datasets.

This gene networks comparison, in a way confirms the results obtained by comparing the FFA and RFA curves. If ReliefF is better than random forests, this means that selecting the top-k genes output by ReliefF should contain more relevant genes than the top-k genes of random forests. The more relevant genes are more likely to be implicated in previous biological knowledge as related to c-MYC and n-MYC activity.

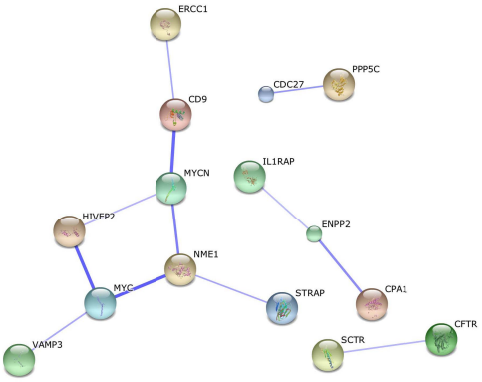


(a) Random Forests stability graph comparison for all ET datasets

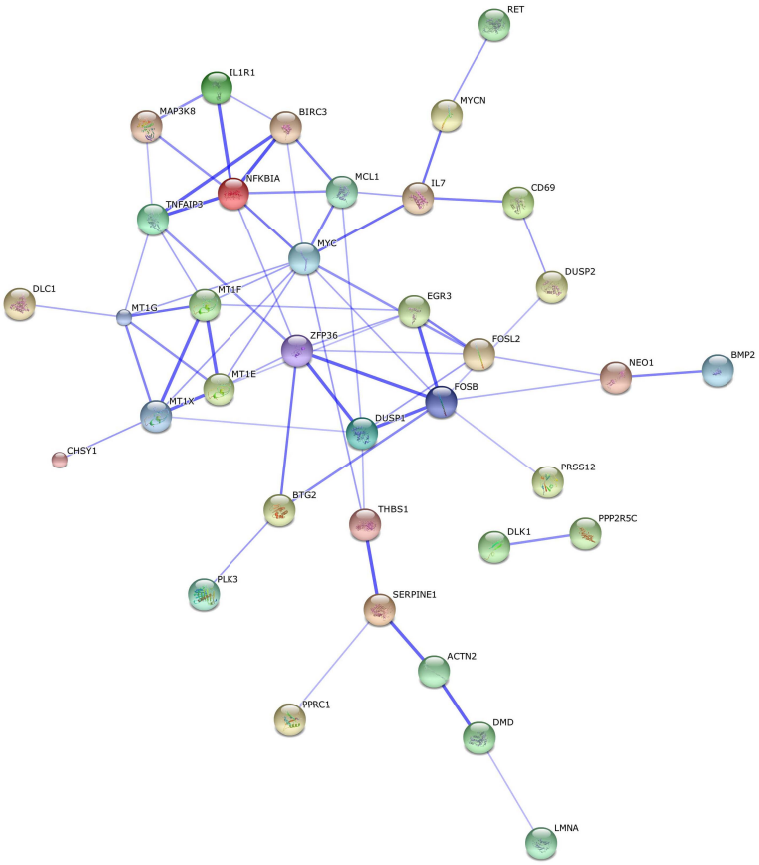


(b) ReliefF stability graph comparison for all ET datasets

Figure 21: Each figure represents a comparison of stability of a single ranking method for all different ET datasets. They are used to determine the cut-off point for selecting the top-k features, used for gene network construction. This is represented by a vertical line on each graph



(a) Gene network constructed from the top-50 genes provided by RFs



(b) Gene network constructed from the top-50 genes provided by ReliefF

Figure 22: Gene networks constructed from the top-50 genes of different feature ranking algorithms. The ranked gene list is induced for the "nbCol251" dataset. Nodes represent a gene/protein and vertices previously shown connections between genes. The genes that are not connected are omitted from the graph.

6.3.4 Aggregation of Embryonal Tumor Datasets

In this section, we present an integrated analysis of the feature rankings of the different tumour entities. In principle, we follow the same general knowledge discovery scenario as described in Section 6.3.3, graphically depicted in Figure 20. However, there is a difference in the biological question that is being investigated and in how the feature ranking comparison.

The biological question being investigated is if a common set of genes exists, valid for all the tumour entities that can be reasonably correlated to the aggressiveness of the tumour. This common set of genes is induced as a consensus or an aggregated ranking from all the individual gene lists. Therefore, our comparison in terms of feature rankings is a comparison between the aggregated gene list with the gene lists derived from the individual ET datasets.

Experimental Setup

Here, we present the details of our experimental setup for the analysis of the aggregated ET gene list. We first describe how the aggregated rankings are generated. We then explain in detail how the comparison with the individual ET rankings is executed.

The individual datasets feature rankings are generated by using random forests and reliefF, with the same settings as in Section 6.3.3, namely:

- **RReliefF** algorithm as given in (Robnik-Šikonja and Kononenko, 2003). This algorithm is an adaptation of the reliefF algorithm, suited for regression, i.e. for numeric targets. The number of neighbours was set to 10 and all of the instances were used for estimating the relevance values.
- **Random forests**, which can be used for estimating feature relevance as described in (Breiman, 2001). A forest of 100 regression trees was used, constructed by randomly choosing 10% of the number of features.

If we denote the set of feature rankings generated from the individual datasets as $\{\mathrm{R}_{1,ind},\ldots,\mathrm{R}_{10,ind}\}$, we calculate the aggregated rank of each feature $F_i \in \mathrm{R}_{agg}$, as:

$$rank(F_i)_{agg} = \underset{j=1\ldots10}{\mathrm{agg}} \{w_j \cdot rank_{ind,j}(F_i)\}$$

. The specific aggregation functions that we consider for calculating the aggregated rank of each feature F_i are:

- the **mean** value of the individual ranks: $rank(F_i)_{agg} = \mathrm{mean}_{j=1\ldots10}\{rank_{ind,j}(F_i)\}$
- the **median** value of the individual ranks: $rank(F_i)_{agg} = \mathrm{median}_{j=1\ldots10}\{rank_{ind,j}(F_i)\}$
- the **maximal** value of the individual ranks:$rank(F_i)_{agg} = \max_{j=1\ldots10}\{rank_{ind,j}(F_i)\}$
- the **minimal** value of the individual ranks: $rank(F_i)_{agg} = \min_{j=1\ldots10}\{rank_{ind,j}(F_i)\}$

For comparing the aggregated ranking R_{agg} with the set of individual rankings, $\{\mathrm{R}_{1,ind},\ldots,\mathrm{R}_{10,ind}\}$, we construct FFA and RFA curves. The purpose is to compare the quality of the aggregated (consensus) ranking across all ET entites, with the quality of the individual rankings, also across all ET entities. In order to do so, the estimation of the values of the FFA and RFA curves is conducted slightly different for the aggregated and for each of the individual rankings.

The estimation of the FFA and RFA values for each of the individual rankings is presented schematically in Figure 23. For each ET dataset $\mathscr{D}_i$, a corresponding feature ranking R_i is induced. Then this feature ranking is evaluated on all the remaining ET datasets, except the $\mathscr{D}_i$ dataset from which the feature ranking was induced. By averaging the values of

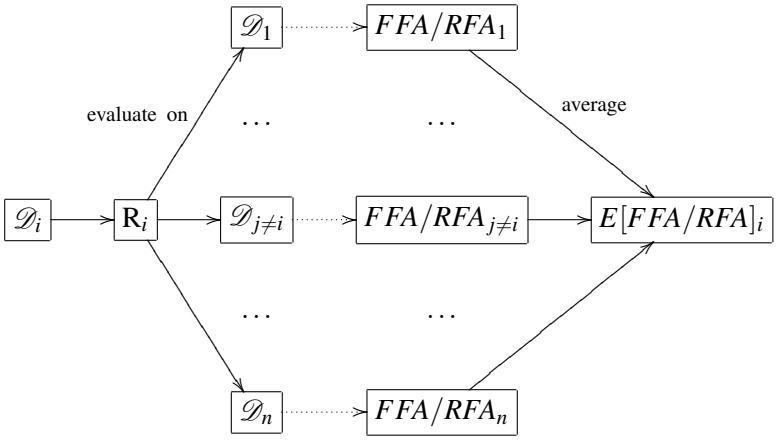


Figure 23: Graphical representation of the error estimation for a ranking induced from an individual dataset. Each ranking R_i is evaluated on all the remaining $\mathscr{D}_{j\neq i}$ datasets and the error is averaged.

these FFA and RFA curves, we obtain the estimated FFA and RFA curves for the individual ranking R_i across all ET datasets.

The estimation procedure of the FFA and RFA curves for the aggregated ranking R_{agg}, is slightly different, as evident in Figure 24. The general intuition behind this estimation is performing a "leave-one-dataset -out" evaluation. Averaging the FFA and RFA curves obtained for each left-out dataset provides the estimated FFA and RFA curves for the aggregated ranking.

If we consider this estimation step by step, it begins by inducing a feature ranking R_i, for each corresponding ET dataset $\mathscr{D}_i$. Then, an equal number of aggregated rankings $\mathrm{R}_{-i,agg}$, are calculated. Each aggregated ranking that is generated, does not include one of the individual feature rankings. Namely, the aggregated ranking denoted as $\mathrm{R}_{-i,agg}$, leaves out the individual ranking R_i induced from dataset $\mathscr{D}_i$.

In the next step, each of the aggregated rankings $\mathrm{R}_{-i,agg}$ is evaluated on the left-out dataset $\mathscr{D}_i$ and FFA_{-i} and RFA_{-i} curves are constructed. The final estimated FFA and RFA curve, for the aggregated ranking is calculated as the average of each of these FFA_{-i} and RFA_{-i} curves. This gives the average quality of the aggregated feature ranking across all ET datasets.

The values of the FFA and RFA curves, considering target was numeric, are the correlation coefficient between the true and the predicted values. For generating the predicted values SVMs for regression with polynomial (quadratic) kernel were used and 10-fold cross validation was performed, on the dataset under consideration. The epsilon parameter of the SVMs was set to 1.0E-12, while the complexity was 0.1.

For constructing the gene networks, in accordance with the individual ET dataset analysis, we consider a subset of top-50 ranked genes from the aggregated gene lists. These top ranked genes are provided as input of the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) (Szklarczyk et al., 2011), version 9.0.

Results

We present the results of the aggregated ET analysis structured in the same way as the results of the individual ET datasets analysis, according to the knowledge discovery scenario in Figure 20. We begin by presenting the comparison of the FFA and RFA curves of the aggregated ranking with the individual rankings. We then present the gene networks reconstructed from the top 50 genes of the aggregated rankings.

The summary of the results from the comparison are given in Table 9 and the FFA/RFA curve plots are given in Appendix B.5 . The rows of the table refer to different datasets, while

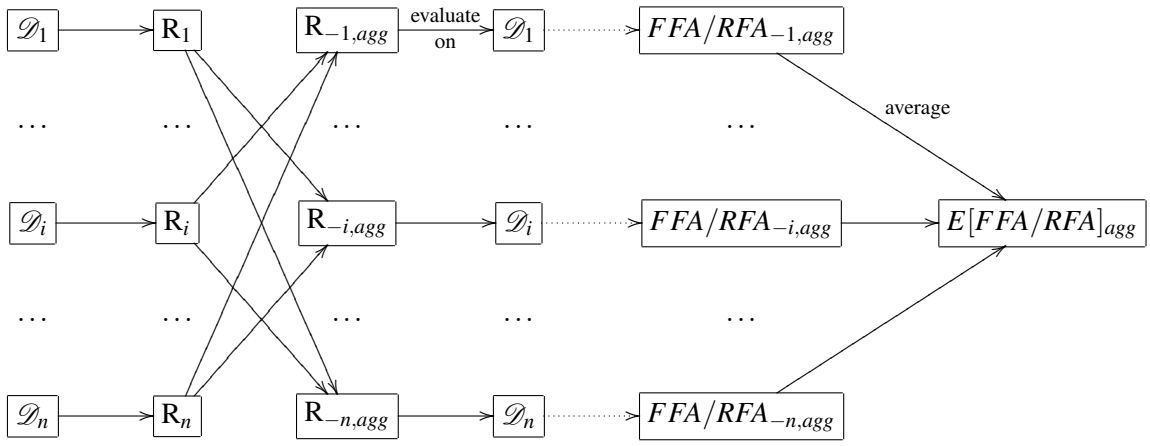


Figure 24: Graphical representation of the error estimation for the aggregated ranking induced from all of the datasets. The evaluation is in a "leave-one-dataset-out" manner. First, aggregated gene rankings $\mathrm{R}_{-i,agg}$ are induced, which do not include a single dataset from the ET collection. Then the aggregated ranking is evaluated on the excluded dataset $\mathscr{D}_i$. At the end all of the individual error evaluations are averaged.

the columns refer to a different combination of ranking method and aggregation function. The calculated ECA differences are between the FFA/RFA curves produced by a specific ranking method/aggregation function and the FFA/RFA curves produced by the individual feature rankings.

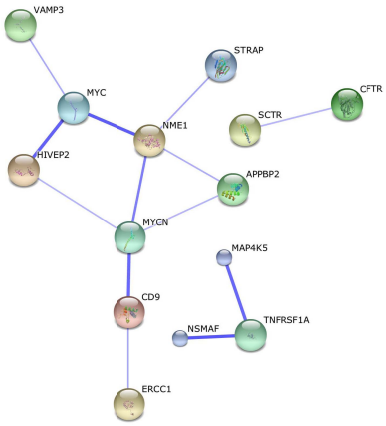
If we first examine the results of the aggregated rankings of random forests (first four columns of Table 9), we can notice that most of the time the ECA differences have the value of zero. This means that the aggregated ranking produced from feature rankings of random forest does not provide any improvement over the individual rankings, evaluated across the ET domain. However, the results of RReliefF are quite different. They show that for different aggregation functions, there is a difference between the quality of the aggregated feature ranking over the individual feature rankings. Most of the improvement of quality comes if the minimum or the mean aggregation function is used to aggregate the feature rankings. In fact, although with a small difference, the min aggregation function is constantly better than mean.

Based on this ECA differences analysis, we construct a gene network from the top-50 genes of the aggregated ranking produced by the RRelieF individual rankings with the minimum aggregation function. Additionally, we compare this network with the one constructed from random forests individual rankings also with the minimum aggregation function. The network for RReliefF is presented in Figure 25b, while the network for random forests is presented in Figure 25a. The full set of networks for the different aggregation functions can be seen in Appendix B.6.

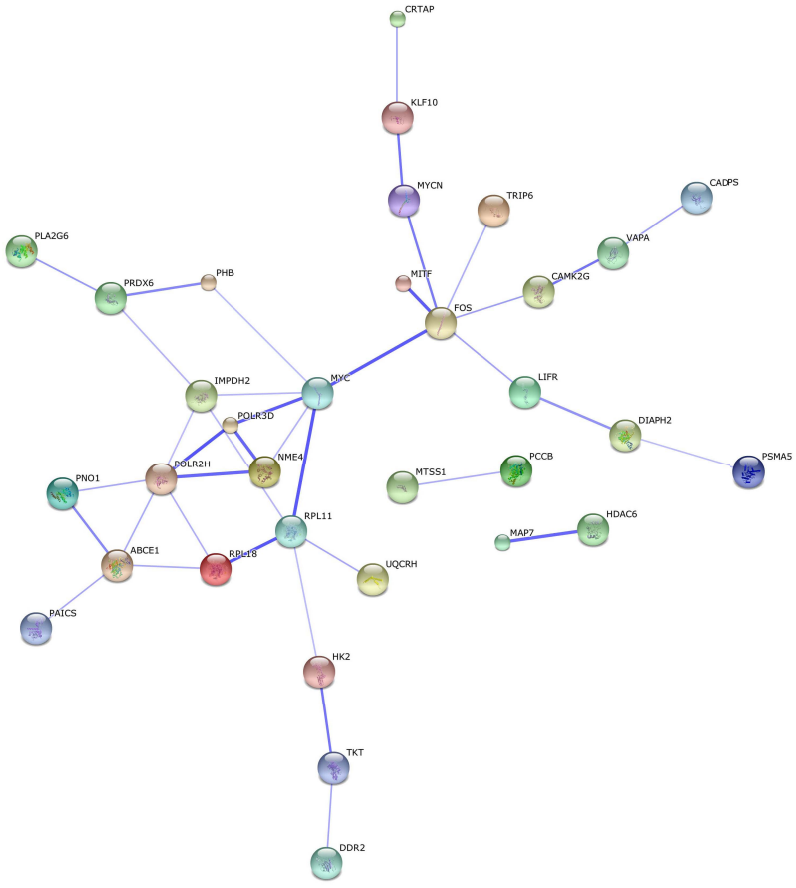
Upon inspection of the results, it can be noticed that the same discussion applies as in the case of the networks constructed from individual ET feature rankings, from Section 6.3.3. Namely, the gene network constructed from RReliefF and minimum aggregation function, contains much more nodes that are highly interconnected as compared to the gene network of random forests. This again can be interpreted that the aggregated gene lists provided by RReliefF are of better quality, as knowledge about their relation to the MYC genes is already implied in the biological domain, unlike the genes provided by the aggregation of random forests gene lists.

Table 9: ECA differences calculated between the FFA/RFA curves of various combinations of baseline feature ranking and aggregation methods. The differences are calculated w.r.t. the curves of an individual ranking.

Datasets	**Random Forests**				**RReliefF**			
	mean	**median**	**max**	**min**	**mean**	**median**	**max**	**min**
ews12102	0.01	0.01	0.01	0	0.08	0.05	0.03	0.1
mb10327	0	0	0	0	0.03	0	-0.02	0.04
mb12992	0	0	0	0	0.05	0.02	0	0.06
mbDKFZ	0	0	0	0	-0.02	-0.05	-0.07	0
nbCol251	0	0	0	0	0.04	0.01	-0.01	0.06
nbEssen	0	0	0	0	0.03	0	-0.02	0.04
rtCurie	0.01	0.01	0.01	0	0.04	0.01	-0.01	0.06
wt10320	0	0	0	0	0.04	0.01	-0.01	0.05
wt11024	0.01	0.01	0.01	0	0.04	0.01	-0.01	0.05
wtETABM53	0.01	0.01	0.01	0.01	0.08	0.05	0.03	0.09



(a) Gene network constructed from the top-50 genes provided by RFs



(b) Gene network constructed from the top-50 genes provided by ReliefF

Figure 25: Gene networks constructed from the top-50 genes of different feature ranking algorithms with the *min* aggregation functions. Nodes represent a gene/protein and vertices previously shown connections between genes. The genes that are not connected are omitted from the graph.

6.3.5 Conclusions

In this section we present a very specific, knowledge oriented analysis of a collection of embryonal tumours datasets. The biological question that drove the analysis was the discovery of key genes involved in embryonal tumour aggressiveness. Besides the biological knowledge discovery, the purpose of this analysis was to also demonstrate how our feature ranking evaluation method can be an integral part of a generic knowledge discovery scenario.

We divided the analysis of the ET datasets in two parts, the first one concerning individual ET datasets analysis and the second part an integrative ET analysis. Both types of analysis followed the generic knowledge discovery scenario from Figure 20. This scenario involved comparison of different feature rankings (gene lists), selection of the best one and finally, selection of top-k genes used for gene network re-construction.

The individual analysis of the ET datasets, compared two feature ranking methods RReliefF and random forests. The evaluation and comparison with FFA and RFA curves revealed that RReliefF provided, most of the time, a better gene list than random forests. This was also confirmed by the subsequent gene network construction, from the top-50 genes. Namely, the gene lists provided by RReliefF had a better overlap with the existing biological knowledge for genes involved in the mechanism of tumour aggressiveness.

The integrative analysis, examined the consensus (aggregated) gene lists of the various ET datasets. Various pairs of feature ranking methods and aggregation function were examined that were compared to the gene lists induced from the individual datasets. The results demonstrated that by aggregating the gene lists produced by RReliefF, with the min aggregation function provided an aggregated ranking which was of a good quality and valid across all ET datasets. This was demonstrated both through the use of FFA and RFA curves and the subsequent gene network re-construction.

6.4 Summary

In this chapter we presented experimental work concerning different applicative domains. The general purpose of the experiments was to show the usefulness of our feature ranking evaluation method by demonstrating its use on various problems. We investigated three diverse feature-ranking related problems, which were appropriately addressed with our feature ranking evaluation methodology.

The first problem, presented in Section 6.1, was a purely machine-learning problem concerning feature ranking ensemble. In this work we tried to extend the concept of ensemble learning to feature ranking problems. We considered a variety of possible parameters while constructing the feature ranking ensembles and evaluated the aggregated ranking for each. The results demonstrated a limited improvement in the feature ranking quality constructed with feature ranking ensembles. The main advantage of feature ranking ensembles was the stabilising effect they had on feature ranking methods. Consequently, only when using random forests as baseline classifiers there was a benefit from feature ranking ensembles

The second problem was a comparative analysis of feature ranking methods on diverse domains. The datasets included various domains, different number of instances and features and different unknown feature interaction structures. However, the high diversity of the datasets did not provide highly diverse results. The statistical tests showed that info gain and ReliefF provided feature rankings which had insignificant differences of quality across all of the datasets. However, they both seemed to outperform random forests, which sometimes provided peculiar results by inducing the inverse ranking from info gain and ReliefF.

The third problem investigated was a very specific biological problem from the area of embryonal tumours research. The data was a collection of microarray gene expression data of various embryonal tumour entities. The purpose of the analysis was to identify key genes involved in the process of tumour aggressiveness. In addition, it also aimed to demonstrate

how our feature ranking evaluation method, can be incorporated into a generic microarray analysis scenarios. From a biological perspective, the results demonstrated that a common gene set related to tumour aggressiveness can be identified, which is valid for all the different embryonal tumour entities. In terms of feature ranking methods, the results showed that RReliefF provided gene list that were better than those of random forest. This was concluded both by evaluation with FFA and RFA curves, but also in terms of consistency with previous biological knowledge.

7 Conclusions and Further Work

In this chapter, we summarise the work presented in this thesis and provide an account of each of the individual contributions. The main topic of our research was the machine learning task of feature ranking, in the context of supervised. Our work has been to first define feature ranking by using various sources in the literature. As feature relevance is the basis of feature ranking, we also provide the definition of a feature relevance index based on feature interactions.

The proposed method for evaluating feature rankings is placed in the context of filter methods for feature selection. More precisely, it performs a stepwise construction of predictive models for different sets of top-k and bottom-k features, as provided by a feature ranking. The aim of our work was to provide a method that answers the question: "Is feature ranking R_A better than feature ranking R_b?". The extensive experiments performed on synthetic data that involved applying uniform and non-uniform noise to the ground truth ranking have demonstrated that the evaluation method works as intended. The application-oriented experiments have demonstrated the usefulness of our method for evaluating feature rankings in real-world domains.

We discuss each of the contributions of our work in more detail in Section 7.1. Finally, we conclude with ideas and directions for further work in Section 7.2.

7.1 Contributions

We structure and present the individual contributions of this thesis into three general categories:

- contributions to defining and quantifying feature ranking,
- contributions to establishing a feature ranking evaluation methodology,
- contributions to practical data analysis problems.

Defining Feature Ranking: The first contributions, involving the definition and quantification of ground truth ranking, include the work presented in Chapters 2 and 3. By surveying the literature, we concluded that feature ranking has always been subsumed by feature subset selection, although they are essentially different tasks. This stems from their joint utility as machine learning tasks that precede the predictive model induction phase. Namely, their joint purpose is to select features that can be used to induce a better predictive model.

By combining the various definitions provided in the literature, we managed to provide a definition of feature ranking that includes all the essential characteristics of that process. We defined the feature ranking as a process that solves the all-relevant feature selection process and additionally provides a correct ordering of the relevant features. The output of the feature ranking process is a linear ordering of features.

For quantifying feature relevance we used a feature relevance index based on feature interactions and grounded in information theory. The advantage of this index is that it does

not treat the relevance of a feature independently from the other features in the dataset. Namely, sometimes the relevance of a feature w.r.t. to a target concept can be only obvious in conjunction with other features. This is taken into account when calculating the relevance of a feature and is rewarded with a higher relevance value.

Evaluating Feature Rankings:

Our second group of contributions is related to defining a method for evaluating feature rankings. The theoretical groundwork has been described in Chapter 4, while the supporting experiments are presented in Chapter 5. The main contribution here is the definition of a formal algorithmic procedure that can be used for evaluating the distribution of relevant features within a feature ranking. The evaluation method extends the basic intuition found in the literature of evaluating feature ranking algorithms as filter methods. It not only considers the top-k features as important for evaluating the feature ranking, but also the bottom-k features, which are necessary for evaluating the overall distribution of relevant features. In addition, by defining the expected error curves, we provide a baseline to which every feature ranking method can be compared to.

We performed experiments on synthetic data, which support the claim that this evaluation method can distinguish between feature rankings of different quality. As the ground truth ranking is known for the synthetic datasets, we perform experiments by adding different levels of noise to the ground truth ranking. We consider adding either uniform or non-uniform noise (as provided by different feature ranking methods). In both cases, the evaluation method for feature rankings was able to detect and properly quantify the difference in quality of the rankings with different levels of noise.

Feature Ranking Empirical Studies:

Our final contributions in this thesis are related to the experimental analysis and results, presented partly in Chapter 5 and mostly in Chapter 6. These are mostly empirical studies, focused on practical questions originating from different domains. The conclusion from these studies can be considered as a direct contribution to the different domains. They can be divided into:

- a comparative study of different feature ranking algorithms,
- evaluating the usefulness of feature ranking ensembles,
- discovering key genes involved in embryonal tumour aggressiveness.

The first empirical study comparing different feature ranking algorithms, is performed both on synthetic datasets and on real-world data. On the synthetic data, we perform analysis of the individual feature ranking algorithms with our feature ranking evaluation method. Additionally, we consider the stability of the algorithms. This analysis reveals that ReliefF is the best performing ranking method among the ones that were analysed, both in terms of correct feature ranking and stability. The comparative analysis of the feature ranking algorithms on real-world data from various domains revealed that, on average, there is no significant difference between the performance of the ReliefF and the information gain methods for feature ranking.

The second empirical study concerns a special type of feature ranking methods, namely feature ranking ensembles. They are inspired by ensemble learning and are basically used to produce a consensus feature ranking from multiple baseline rankings. Different parameters can be varied in the process of constructing the feature ranking ensembles and we explore several of them. From the experiments, it can be concluded that the major advantage of using ensembles for producing a feature ranking is in their stabilising effect on the base rankings. As expected, this effect is only visible when combining unstable baseline rankings, such as those produced by Random forests.

The last applied experimental work presented in this thesis addresses the area of cancer research, more specifically embryonal tumours. The purpose of this analysis was to find key

genes involved in embryonal tumour aggressiveness. We mapped this problem to a feature ranking problem, where the relevance of the genes is assessed w.r.t. the expression of c-MYC and N-MYC genes.

We analysed individual tumour datasets, but also provided an integrated analysis of all of them. The analysis consisted of producing feature rankings with several methods and comparing them by using our feature ranking evaluation method. From the analysis of the individual tumour datasets, we managed to determine the best feature ranking method for identifying the genes related to tumour aggressiveness. The integrated analysis showed that there is a common set of genes for all tumour entities that can be related to tumour aggressiveness. Our findings were also backed up by the subsequent re-construction of gene networks based on previous knowledge about gene interactions.

7.2 Further Work

After summarising all of the contributions of the work presented in this thesis, we now discuss some possibilities for further work. The present work can be extended in three general directions:

- to evaluate the feature rankings by considering different utility for the feature ranking, e.g. weighted classifiers such as weighted kNN,
- to incorporate different measures for evaluation which directly incorporate feature ranking stability,
- to extend the methodology towards evaluating feature rankings in the context of predicting different types of structured outputs.

The first direction of development that we would like to consider is the use of different utility for the feature ranking. Our work, and most of the evaluation of feature ranking algorithms is placed in the context of the use of filter methods for feature selection and the subsequent construction of predictive models. We would like to extend this by plugging in feature weights directly into the model induction phase.

The second direction is to incorporate different measures for evaluating the quality of the feature ranking. Namely, with our evaluation method, we investigate feature rankings indirectly via constructing predictive models. As previously mentioned, there is a lot of work which evaluates directly the robustness of feature ranking methods by considering their stability. Given that the desired properties of a feature ranking algorithm would be to correctly rank relevant feature and to be stable, these two criteria can be combined into a single measure for evaluating the feature ranking, in a manner similar to the one proposed by Saeys et al. (2008).

The third direction is inspired by the recent increase of interest in analysing and predicting structured data (Bakır et al., 2007). Although the evaluation method that we propose is general, in our experimental work, we have so far performed experiments which mainly involved (binary) classification problems and regression problems. Therefore, we would like to expand and adapt our evaluation method to include different types of structured data. To this end, we need to use a feature ranking method for structured targets and a couple it with a predictive model also for structured outputs (Kocev et al., 2008).

Acknowledgements

In the process of making this thesis many people and institutions were involved. They all contributed to this work in their own way and made the whole experience worthy. I would like to express my gratitude to all of them together and acknowledge each of them separately.

First, I would like to express my gratitude to my thesis supervisor Prof. Dr. Sašo Džeroski. He supported my research work by both providing funding and offering scientific insights and ideas. I owe to him the chance to work on various research projects, visit different research institutions and attend international conferences. This provided me with the necessary experience needed for scientific work.

I would like to especially thank the committee members for reviewing the work presented in this thesis. Namely, Asst. Prof. Bernard Ženko, Assoc. Prof. Dr. Marko Robnik-Šikonja and Dr. Benedikt Brors offered constructive advice on how to improve the quality of the work. For that I am grateful.

Multiple institutions kindly provided funding for my research work. They are the Ad Futura Scientific and Educational Foundation, the Department of Knowledge Technologies (Jožef Stefan Institute, Ljubljana) and the Department of Medical Genetics (University Medical Center, Ljubljana). In addition to funding, the colleagues from the Department of Knowledge Technologies provided a pleasant and friendly working environment as well as interesting scientific debates for various research problems. From the Department of Medical Genetics, I would like to acknowledge Prof. Dr. Borut Peterlin and Asst. Prof. Luca Lovrečić for their cooperation.

Also, it was a real pleasure to work with all of the people involved in the "E.E.T.-Pipeline" project, coordinated by Prof. Dr. Angelika Eggert. It was inspiring, at times challenging but most importantly- it was really fun. It provided me with knowledge related to cancer research and to practical analysis of biological data. In addition, I would like to especially thank Dr. Kathy Astrahantseff for providing insights about biological research, but most importantly for her friendliness and long discussions in the beer gardens of Essen.

There are also many people that supported me personally while making this thesis. My companions from day one were Dragi Kocev and Panče Panov. We have been together through some good times and through some bad times. Together with Aleksandra Raškovska, I consider them as my family here in Ljubljana and I have to acknowledge each of them individually.

Dragi Kocev, has been my best friend at all times. Thank you for all the laughs we had together, for all the drama we endured and for all the good times we had. Putting up with me all these years shows how kind-hearted you are.

Panče Panov is the living proof that big people have big hearts. I have to thank you for always being there for me, for your world-famous cooking and for showing me how not to worry too much in life.

Aleksandra Raškovska is the person that always tries to do the right thing. This is obvious in her sense for style, in her meticulousness at work and especially in her quiet way of being good to people. One can be only glad to have such a friend.

Also, I have been surrounded by a lot of other wonderful people in Ljubljana and Skopje,

which have helped my personal growth. I would like to give them all an honourable mention. In no particular order, I acknowledge each one of them individually.

Vesna Andova, for her easy going attitude and her willingness to do new things in life, which are an inspiration to me. Marija Drenkovska, my always-on-the-run friend, whose energy and will to challenge herself brings out the best in me. Elena Ikonomovska, for showing me that one can indeed have it all in life, if one is willing to work for it. Živa Antauer, for being the best flatmate and a dear friend. Mirjana Batinić and Nevena Aleksovski, the artsy friends that are always interesting to be with, during walks in the park and talks in the bars. Nikola Simidjievski, for being a guy with an attitude that is evenly matched by his inherent goodness and great talent for photography. Jovan Tanevski for being the smart guy, with encyclopaedic knowledge about everything. And finally, Dušana Majera, for (over)analysing and laughing at life with me.

From the Skopje crew, I would also acknowledge many people, for their long distance support while making this thesis. Those are the twin sisters Ana and Aleksandra Hristovi, Mila Stankovi ́k, Nataša Kaceska, Dana Gapkovska, Ksenija Putilin, Ivana Gavriloska, Jana Karčeska, Lilian Kandikjan, Miloš Gjuroski, Ana Veta, Tomislav Zografski, Gorazd Titizov, Filip Neškoski and Biljana Trajkovska. Because of all of them, I still feel deeply connected to my home town. They have all made my trips back to Skopje pleasant, warm and most of all fun.

Also, a special thanks goes to my aunt Snežana Slavkova, who offered invaluable support to me and my family, during the whole time.

Finally, my sincerest gratitude goes to my loving family that supported me selflessly during the whole time of making this thesis. They have always been the wind at my back and without them none of this would be possible.

I want to thank my father Duško for many things, but I will name just a few. I am grateful for his quiet way of worrying about me, but I also cherished his loud advices in times of need. His constant belief in me and his support lead me strive for more in life.

My wonderful mother Nataša, also leaves me with countless things to be grateful about. She has been caring and loving, but also surprisingly strong when I least expect it. I thank her for being a beacon of light, in times good or bad.

Having an older brother is always something to be grateful about. My brother Marjan, so different but yet so similar to me, has done his role as an older brother well. Telling me to be more ambitious when I am not and grounding me when I am irrational. Thank you for supporting me.

8 References

Amit, Y.; Geman, D. Shape quantization and recognition with randomized trees. *Neural Computing* **9**, 1545–1588 (1997).

Bakır, G. H.; Hofmann, T.; Schölkopf, B.; Smola, A. J.; Taskar, B.; Vishwanathan, S. V. N. (eds.) *Predicting structured data* (The MIT Press, 2007).

Bell, D. A.; Wang, H. A formalism for relevance and its application in feature subset selection. *Machine Learning* **41**, 175–195 (2000).

Blum, A. L.; Langley, P. Selection of relevant features and examples in machine learning. *Artificial Intelligence* **97**, 245–271 (1997).

Boser, B. E.; Guyon, I. M.; Vapnik, V. N. A training algorithm for optimal margin classifiers. In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory* 144–152 (ACM, 1992).

Bratko, I. *Prolog Programming for Artificial Intelligence*, 3rd ed. (Addison Wesley, 2000).

Breiman, L. Bagging predictors. *Machine Learning* **24**, 123–140 (1996).

Breiman, L. Random forests. *Machine Learning* **45**, 5–32 (2001).

Breiman, L.; Friedman, J.; Olshen, R.; Stone, C. J. *Classification and Regression Trees* (Chapman & Hall/CRC, 1984).

Caruana, R.; Freitag, D. *How Useful Is Relevance?* Tech. rep., In: *Relevance, Papers from the 1994 AAAI Fall Symposium* (1994).

Corbin, M.; de Reynies, A.; Rickman, D. S.; Berrebi, D.; Boccon-Gibod, L.; Cohen-Gogo, S.; Fabre, M.; Jaubert, F.; Faussillon, M.; Yilmaz, F.; Sarnacki, S.; Landman-Parker, J.; Patte, C.; Schleiermacher, G.; Antignac, C.; Jeanpierre, C. Wnt/SS-catenin pathway activation in wilms tumors: A unifying mechanism with multiple entries? *Genes, Chromosomes and Cancer* **48**, 816–827 (2009).

Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods* (Cambridge University Press, 2010).

Demšar, J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research* **7**, 1–30 (2006).

Dietterich, T. G. Ensemble methods in machine learning. In: *Proceedings of the First International Workshop on Multiple Classifier Systems* 1–15 (Springer, 2000a).

Dietterich, T. G. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning* **40**, 139–157 (2000b).

Duch, W. Filter methods. In: *Feature extraction, foundations and applications* **207**, 89–118 (Springer, 2006).

Dunne, K.; Cunningham, P.; Azuaje, F. Solutions to Instability Problems with Sequential Wrapper-Based Approaches To Feature Selection. Tech. rep., *Journal of Machine Learning Research* (2002).

Džeroski, S.; Demšar, D.; Grbović, J. Predicting chemical parameters of river water quality from bioindicator data. *Applied Intelligence* **13**, 7–17 (2000).

Džeroski, S.; Potamias, G.; Moustakis, V.; Charissis, G. Automated revision of expert rules for treating acute abdominal pain in children. In: Keravnou, E. T.; Garbay, C.; Baud, R. H.; Wyatt, J. C. (eds.) *AIME*, *Lecture Notes in Computer Science* **1211**, 98–109 (Springer, 1997).

E. E. T. P. FP6 - European Embryonal Tumor Pipeline Project (2007–2009). Grant number LifeSciHealth-2005-037260.

Fattet, S.; Haberler, C.; Legoix, P.; Varlet, P.; Lellouch-Tubiana, A.; Lair, S.; Manie, E.; Raquin, M.-A.; Bours, D.; Carpentier, S.; Barillot, E.; Grill, J.; Doz, F.; Puget, S.; Janoueix-Lerosey, I.; Delattre, O. Beta-catenin status in paediatric medulloblastomas: correlation of immunohistochemical expression with mutational status, genetic profiles, and clinical characteristics. *The Journal of Pathology* **218**, 86–94 (2009).

Friedman, M. A comparison of alternative tests of significance for the problem of m rankings. *Annals of Mathematical Statistics* **11**, 86–92 (1940).

Furlanello, C.; Serafini, M.; Merler, S.; Jurman, G. Entropy-based gene ranking without selection bias for the predictive classification of microarray data. *BMC Bioinformatics* **4**, 54 (2003).

Gosset, W. S. S. The probable error of a mean. *Biometrika* **6**, 1–25 (1908).

Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *Journal of Machine Learning Research* **3**, 1157–1182 (2003).

Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. *Machine Learning* **46**, 389–422 (2002).

Guzmán-Martínez, R.; Alaiz-Rodríguez, R. Feature selection stability assessment based on the jensen-shannon divergence. In: *Proceedings of the 2011 European conference on Machine learning and knowledge discovery in databases - Volume Part I*. 597–612 (Springer-Verlag, Berlin, Heidelberg, 2011).

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. The weka data mining software: an update. *SIGKDD Explor. Newsl.* **11**, 10–18 (2009).

Hastie, T.; Tibshirani, R.; Friedman, J. H. *The Elements of Statistical Learning* (Springer, 2003).

He, Z.; Yu, W. Review article: Stable feature selection for biomarker discovery. *Comput. Biol. Chem.* **34**, 215–225 (2010).

Ho, T. K. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**, 832–844 (1998).

Huang, C.-C.; Gadd, S.; Breslow, N.; Cutcliffe, C.; Sredni, S. T.; Helenowski, I. B.; Dome, J. S.; Grundy, P. E.; Green, D. M.; Fritsch, M. K.; Perlman, E. J. Predicting relapse in favorable histology wilms tumor using gene expression analysis: A report from the renal tumor committee of the children's oncology group. *Clinical Cancer Research* **15**, 1770–1778 (2009).

Iman, R. L.; Davenport, J. M. Approximations of the critical region of the friedman statistic. *Communications in Statistics - Theory and Methods* **9**, 571–595 (1980).

Jakulin, A.; Bratko, I. Testing the significance of attribute interactions. In: Brodley, C. E. (ed.) *ICML*, *ACM International Conference Proceeding Series* **69** (ACM, 2004).

Jeffreys, H. An invariant for the prior probability in estimation problems. *Proceedings of the Royal Society A* **186**, 454–461 (1946).

John, G. H.; Kohavi, R.; Pfleger, K. Irrelevant features and the subset selection problem. In: *Proceedings of the Eleventh International Conference of Machine Learning* 121–129 (Morgan Kaufmann, 1994).

Jong, K.; Mary, J.; Cornuéjols, A.; Marchiori, E.; Sebag, M. Ensemble feature ranking. In: *PKDD* 267–278 (2004).

Jurman, G.; Merler, S.; Barla, A.; Paoli, S.; Galea, A.; Furlanello, C. Algebraic stability indicators for ranked lists in molecular profiling. *Bioinformatics* **24**, 258–264 (2008).

Kalousis, A.; Prados, J.; Hilario, M. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and Information Systems* **12**, 95–116 (2007).

Kenney, J. F.; Keeping, E. S. *Mathematics of Statistics, Pt. 2* (Princeton, NJ, 1951).

Kira, K.; Rendell, L. A. A practical approach to feature selection. In: *ML92: Proceedings of the ninth international workshop on Machine learning*. 249–256 (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1992).

Kocev, D. *Ensembles for predicting structured outputs*. Ph. D. thesis, IPS Jožef Stefan, Ljubljana, Slovenia (2011).

Kocev, D.; Slavkov, I.; Džeroski, S. More is better: ranking with multiple targets for biomarker discovery. In: *Proceeding of the Second International Workshop on Machine Learning in Systems Biology* 133 (2008).

Kohavi, R.; John, G. H. Wrappers for feature subset selection. *Artif. Intell.* **97**, 273–324 (1997).

Kolmogorov, A. N. Three approaches to the quantitative definition of information. *Problems of information transmission* **1**, 381–399 (1965).

Kononenko, I. Estimating attributes: Analysis and extensions of relief. In: *European Conference on Machine Learning*. 171–182 (1994).

Kononenko, I.; Kukar, M. *Machine Learning and Data Mining* (Horwood Publishing, 2007).

Kool, M.; Koster, J.; Bunt, J.; Hasselt, N. E.; Lakeman, A.; van Sluis, P.; Troost, D.; Meeteren, N. S.-v.; Caron, H. N.; Cloos, J.; Mršia, A.; Ylstra, B.; Grajkowska, W.; Hartmann, W.; Pietsch, T.; Ellison, D.; Clifford, S. C.; Versteeg, R. Integrated genomics identifies five medulloblastoma subtypes with distinct genetic profiles, pathway signatures and clinicopathological features. *PLoS ONE* **3**, e3088 (2008).

Kort, E. J.; Farber, L.; Tretiakova, M.; Petillo, D.; Furge, K. A.; Yang, X. J.; Cornelius, A.; Teh, B. T. The e2f3-oncomir-1 axis is activated in wilms' tumor. *Cancer Research* **68**, 4034–4038 (2008).

Kullback, S.; Leibler, R. A. On information and sufficiency. *Annals of Mathematical Statistics* **22**, 79–86 (1951).

Kuncheva, L. I. A stability index for feature selection. In: Devedzic, V. (ed.) *Artificial Intelligence and Applications*. 421–427 (IASTED/ACTA Press, 2007).

Lance, G. N.; Williams, W. T. Computer programs for hierarchical polythetic classification ('similarity analyses'). *The Computer Journal* **9**, 60–64 (1966).

Lance, G. N.; Williams, W. T. Mixed-data classificatory programs i-agglomerative systems. *Australian Computer Journal* **1** (1967).

Langley, P. *Elements of machine learning* (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1996).

Lin, J. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory* **37**, 145–151 (1991).

Liu, H.; Motoda, H. *Computational Methods of Feature Selection*. (Chapman & Hall, 2008).

Loscalzo, S.; Yu, L.; Ding, C. H. Q. Consensus group stable feature selection. In: IV, J. F. E.; Fogelman-Soulié, F.; Flach, P. A.; Zaki, M. J. (eds.) *KDD*. 567–576 (ACM, 2009).

Matsuda, H. Physical nature of higher-order mutual information: intrinsic correlations and frustration. *Physical Review E* **62**, :3096–3102 (2000).

Matusita, K. On the theory of statistical decision functions. *Annals of the Institute of Statistical Mathematics (Tokyo)* **186**, 17–35 (1951).

McCarthy, J.; Minsky, M.; Rochester, N.; Shannon, C. A proposal for the dartmouth summer research project on artificial intelligence. URL `http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html`. (accessed: June 2012).

McGill, W. Multivariate information transmission. *Psychometrika* **19**, 97–116 (1954).

Michie, D. Personal models of rationality. *Journal Statistical Planning and Inference* **21**, 381–399 (1990).

Mitchell, T. *Machine learning* (McGraw Hill, 1997).

Molina, L. C.; Belanche, L.; Nebot, A. Feature selection algorithms: A survey and experimental evaluation. In: *Proceedings of the 2002 IEEE International Conference on Data Mining*. 306– (IEEE Computer Society, Washington, DC, USA, 2002).

Nemenyi, P. B. *Distribution-free multiple comparisons*. Ph. D. thesis, Princeton University, Princeton, NY, USA (1963).

Newman, C. B. D.; Merz, C. UCI repository of machine learning databases (1998). URL `http://www.ics.uci.edu/$\sim$mlearn/MLRepository.html`. (accessed: June 2012).

Nilsson, R.; Peña, J. M.; Björkegren, J.; Tegnér, J. Consistent feature selection for pattern recognition in polynomial time. *Journal of Machine Learning Research* **8**, 589–612 (2007).

Oberthuer, A.; Berthold, F.; Warnat, P.; Hero, B.; Kahlert, Y.; Spitz, R.; Ernestus, K.; König, R.; Haas, S.; Eils, R.; Schwab, M.; Brors, B.; Westermann, F.; Fischer, M. Customized oligonucleotide microarray gene expression-based classification of neuroblastoma patients outperforms current clinical risk stratification. *Journal of Clinical Oncology* **24**, 5070–5078 (2006).

Paoli, S.; Jurman, G.; Albanese, D.; Merler, S.; Furlanello, C. Semisupervised profiling of gene expressions and clinical data. In: *Proceedings of the Sixth international conference on Fuzzy Logic and Applications*. 284–289 (2005).

Pearson, K. Mathematical contributions to the theory of evolution (series of papers). *Philosophical Transactions of the Royal Society of London A* (1896–1912).

Press, W.; Teukolsky, S.; Vetterling, W.; Flannery, B. *Numerical Recipes in C* (Cambridge University Press, Cambridge, UK, 1992).

Rissanen, J. Modeling by shortest data description. *Automatica* **14**, 465–471 (1978).

Robnik-Šikonja, M.; Kononenko, I. An adaptation of relief for attribute estimation in regression. In: Fisher, D. H. (ed.) *ICML*. 296–304 (Morgan Kaufmann, 1997).

Robnik-Šikonja, M.; Kononenko, I. Theoretical and empirical analysis of relieff and rrelieff. *Mach. Learn.* **53**, 23–69 (2003).

Saeys, Y.; Abeel, T.; de Peer, Y. V. Robust feature selection using ensemble feature selection techniques. In: Daelemans, W.; Goethals, B.; Morik, K. (eds.) *ECML/PKDD (2)*, *Lecture Notes in Computer Science* **5212**, 313–325 (Springer, 2008).

Saeys, Y.; Inza, I.; Larrañaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **23**, 2507–2517 (2007).

Schramm, A.; Schulte, J. H.; Klein-Hitpass, L.; Havers, W.; Sieverts, H.; Berwanger, B.; Christiansen, H.; Warnat, P.; Brors, B.; Eils, J.; Eils, R.; Eggert, A. Prediction of clinical outcome and biological characterization of neuroblastoma by expression profiling. *Oncogene* **24**, 7902–7912 (2005).

Scotlandi, K.; Remondini, D.; Castellani, G.; Manara, M. C.; Nardi, F.; Cantiani, L.; Francesconi, M.; Mercuri, M.; Caccuri, A. M.; Serra, M.; Knuutila, S.; Picci, P. Overcoming resistance to conventional drugs in ewing sarcoma and identification of molecular predictors of outcome. *Journal of Clinical Oncology* **27**, 2209–2216 (2009).

Sebag, M.; Azé, J.; Lucas, N. Impact studies and sensitivity analysis in medical data mining with roc-based genetic learning. In: *Proceedings of the Third IEEE International Conference on Data Mining*. 637– (IEEE Computer Society, Washington, DC, USA, 2003).

Seni, G.; Elder, J. F. *Ensemble methods in data mining: Improving accuracy through combining predictions* (Morgan & Claypool Publishers, 2010).

Shannon, C. E. A mathematical theory of communication. *Bell System Technical Journal* **27**, 379–423 (1948).

Slonim, D. K. From patterns to pathways: gene expression data analysis comes of age. *Nature genetics* **32 (Suppl.)**,502–508 (2002).

Somol, P.; Novovicová, J. Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**, 1921–1939 (2010).

Stanfill, C.; Waltz, D. Toward memory-based reasoning. *Communications of the ACM* **29**, 1213–1228 (1986).

Szklarczyk, D.; Franceschini, A.; Kuhn, M.; Simonovic, M.; Roth, A.; Minguez, P.; Doerks, T.; Stark, M.; Muller, J.; Bork, P.; Jensen, L. J.; von Mering, C. The string database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research* **39**, D561–D568 (2011).

Turney, P. D. Technical note: Bias and the quantification of stability. *Machine Learning* **20**, 23–33 (1995).

Vajda, I. *Theory of statistical inference and information* (Kluwer Academic Press, 1979).

Vapnik, V. N. *Statistical learning theory* (Wiley, 1998).

Verikas, A.; Gelzinis, A.; Bacauskiene, M. Mining data with random forests: A survey and results of new tests. *Pattern Recognition* **44**, 330–349 (2011).

Štrumbelj, E.; Kononenko, I.; Robnik-Šikonja, M. Explaining instance classifications with interactions of subsets of feature values. *Data Knowledge Engineering* **68**, 886–904 (2009).

Westermann, F.; Muth, D.; Benner, A.; Bauer, T.; Henrich, K.-O.; Oberthuer, A.; Brors, B.; Beissbarth, T.; Vandesompele, J.; Pattyn, F.; Hero, B.; Konig, R.; Fischer, M.; Schwab, M. Distinct transcriptional mycn/c-myc activities are associated with spontaneous regression or malignant progression in neuroblastomas. *Genome Biology* **9**, R150 (2008).

Zucknick, M.; Richardson, S.; Stronach, E. A. Comparing the characteristics of gene expression profiles derived by univariate and multivariate classification methods. *Statistical applications in genetics and molecular biology* **7** (2008).

Index of Figures

Index of Tables

List of Algorithms

A The Influence of Using Different Learning Methods to Construct FFA and RFA Curves

FFA and RFA curves are produced by using the error estimates for the predictive models built on feature sets of different sizes. According to Algorithm 5, the error curve estimates depend on the feature ranking method used, as it directly influences the choice of features used for constructing the predictive models. However, the error estimates also depend on the learning method used to construct the predictive models.

Therefore, in this appendix we investigate how different feature ranking methods compare when using different learning methods to construct the FFA and RFA curves. The purpose of this comparison is to empirically decide which learning methods are suitable for use within our feature ranking evaluation method. The basic requirement for a learning method to be used in this context, is that it should produce predictive models that utilise all of the information that the features have about the target concept. These predictive models would in turn be used to construct FFA and RFA curves that can distinguish between feature rankings of different quality.

Below, we first present the details of the experimental setup for this comparative analysis of FFA and RFA curves. Five different learning methods are investigated in the context of four feature ranking methods. We then present the resulting FFA and RFA curves and discuss them in detail.

Experimental Setup

When comparing the FFA and RFA curves of different ranking methods, constructed with different learning methods, we used the synthetic datasets described in Section 5.1. These datasets are appropriate for this analysis as they have a known feature interaction structure and therefore a known ground truth ranking. This is helpful for the later interpretation of the produced results.

We consider four feature ranking methods:

- **Information gain**, calculating the information gain of each feature F_i as: $IG(F_t, F_i) = H(F_t) - H(F_t|F_i)$. This does not require any specific parameter setting.
- **SVM-RFE**, the redundant feature elimination (RFE) procedure that employs an SVM to evaluate the feature weights at each iteration. A linear SVM was employed, as proposed by Guyon et al. (2002). The epsilon parameter of the SVM was set to 1.0E-12 , while the complexity parameter was set to 0.1.
- **ReliefF** algorithm, as proposed by Robnik-Šikonja and Kononenko (2003). The number of neighbours was set to 10 and all of the instances were used for estimating the relevance values.
- **Random forests**, which can be used for estimating feature relevance as described by Breiman (2001). A forest of 100 trees was used, constructed by randomly choosing a $\log_2$ of the number of features.

We compare these ranking methods by using five learning methods to produce classifiers and generate error estimates for the FFA and RFA curves. More specifically, we consider the following learning methods:

- **Naïve Bayes**
- **Decision Trees**
- **Random Forests:** a forest of 100 trees was used, constructed by randomly choosing $\log_2$ of the number of features.
- **SVMs:** polynomial (quadratic) kernel was employed, with the epsilon parameter set to 1.0E-12 and the complexity parameter set to 0.1.
- **k-NN:** with a value of $k = 10$

We present the obtained FFA and RFA curve comparisons of the four feature ranking methods obtained by each of the five learning methods in the following section.

Results

The summary of the results can be seen in Figure 26 trough 31. Each of the figures shows five comparisons of the four ranking methods, where each figure contains FFA (RFA) curves constructed by one of the five learning methods. The comparison of each type of curve (FFA or RFA) for the three synthetic dataset yields a total of six figures.

Before discussing the results, it is worth noting that methods such as information gain and SVM-RFE with a linear SVM, can only properly estimate the relevance of individually correlated features. The other two methods, ReliefF and random forests can detect the interactions. This claim is in line with the analysis from Section 5.3.2, more specifically Figures 11a, 11b, 12a and 12b.

We would expect this to be reflected in the FFA/RFA curves in such a way that the FFA curves of ReliefF and random forests, achieve an overall higher accuracy or at least earlier in the curve. For the RFA curves we expect the opposite, namely that information gain and SVM-RFE should have higher accuracy as compared to the ReliefF and random forests, at least earlier in the curve.

We now consider the results, for the FFA and RFA curves given in Figure 26 and Figure 27 correspondingly. All of the classifiers considered, yield almost identical accuracy estimates. These estimates are for the "single" dataset that includes only features individually associated with the target class. Considering the previous discussion from Section 5.3.2, all of the feature ranking methods are able to properly estimate the relevance of the individually correlated features. Therefore, equally accurate classifiers can be constructed independently of the classifier method.

Next, we consider the FFA and RFA curves for the "pair" dataset given in Figure 28 and 29. As mentioned earlier, ranking methods such as information gain and SVM-RFE with a linear SVM should not be able to detect the relevance of the XOR features present in the data. Therefore the produced FFA and RFA curves for information gain and SVM-RFE, should differ from those of random forests and ReliefF.

For the Naïve Bayes classifier, both the FFA and RFA curves in Figures 28a and 29a show virtually no difference between the four ranking methods. The reason behind this is that the Naïve Bayes classifier simply can not use the information from the interactions of higher order. This means that Naïve Bayes is not appropriate for use in the considered context.

All the other learning methods seem to be able to detect the difference between the feature rankings, up to a certain degree. Namely the error curves constructed by decision

trees (Figures 28b and 29b) show differences between the ranking methods. However, the error curves are highly variable, especially the estimates for the RFA curve.

The other three learning methods (random forests, SVMs and k-NNs) produce comparable FFA and RFA curves. The are several differences, however. Random forests and SVMs seem to be able to produce in general more accurate models as compared to k-NNs. This means that they can both distinguish more clearly between the feature rankings of different quality, which is especially evident when constructing the RFA curves. The magnitude of the error estimates of the models produced by random forests and SVMs is quite similar, but the FFA and RFA curves of random forests are more variable than those of SVMs.

Similar observations can be made for the "combined" dataset. The error curves constructed by Naïve Bayes in Figures 30a and 31a are not sensitive to the different feature rankings, as there is no difference between the FFA and RFA curves. Also, the error curves constructed by decision trees in Figure 30b do not show any difference between the different feature rankings. There is a visible difference in Figure 31b, but the estimates are highly variable.

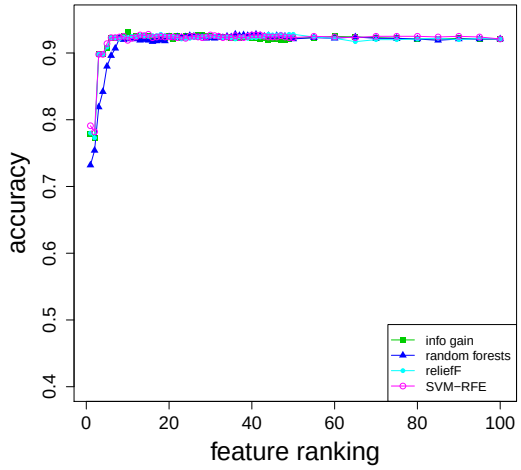
For the remaining three learning methods (random forests, SVMs and k-NNs), similar observations as for the "pair" dataset can be made . Namely, they all distinguish between the different feature ranking methods, except when random forest are used for constructing the FFA curve (Figure 30c). Otherwise, it is again visible that random forests (RFA curve) and SVMs provide error curves with the greatest separation for the different ranking methods. The RFA curves of random forests are more variable then the RFA curves produced by SVMs.

Conclusions

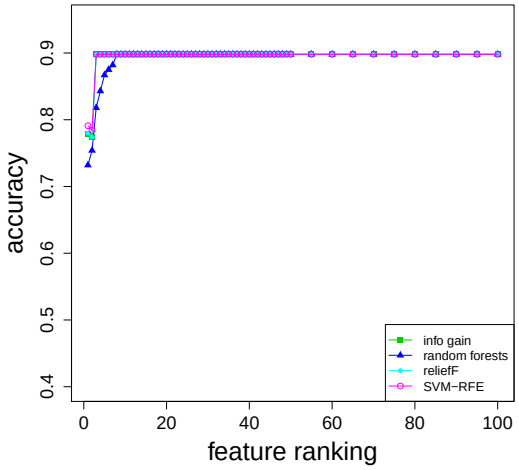
From the previous discussion of the obtained results, we can draw several conclusions about the learning methods that can be used to generate the FFA and RFA curves. First, Naïve Bayes and decision trees should not be used as the first is not sensitive to different feature rankings and the second produces curves that are very variable.

The other three methods, in general, can be used to construct error curves for evaluating feature rankings. They satisfy the basic requirement that they produce error curves that are able to distinguish between feature rankings of different quality. However, random forests and SVMs produce predictive models that are better able to utilise the information that the features have about the target concept, as compared to k-NN. In addition, SVMs produce FFA and RFA curves that are less variable than those produced by random forests. Therefore, in this thesis, we employ SVMs as a learning method when constructing FFA and RFA curves.

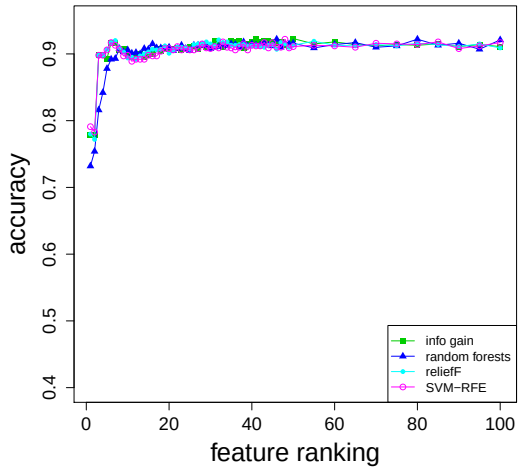
Finally, it should be noted that the above experiments were performed on specific synthetic datasets. It may not be so straightforward to generalise these conclusions. However, we believe that they provide an initial empirical guideline on what learning methods should be used when evaluating feature rankings by constructing predictive models.



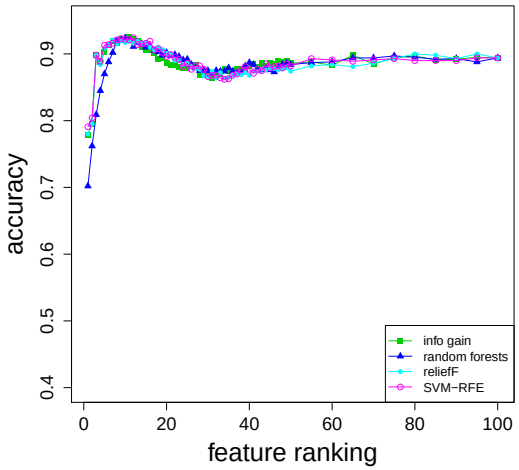
(a) FFA curves obtained with **Naïve Bayes**



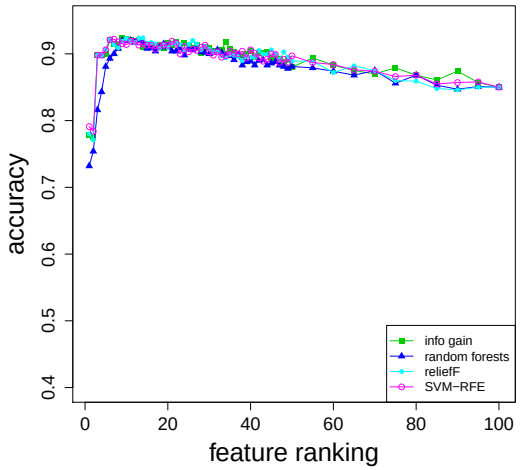
(b) FFA curves obtained with **Decision Trees**



(c) FFA curves obtained with **Random Forests**

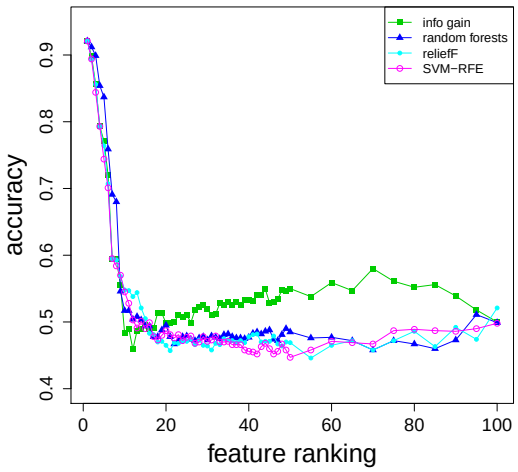


(d) FFA curves obtained with **SVMs**

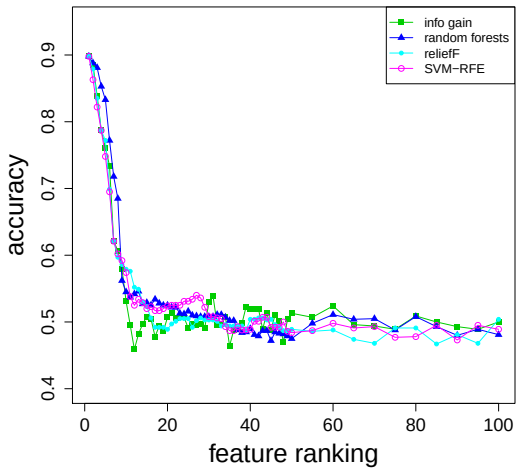


(e) FFA curves obtained with **k-NN**

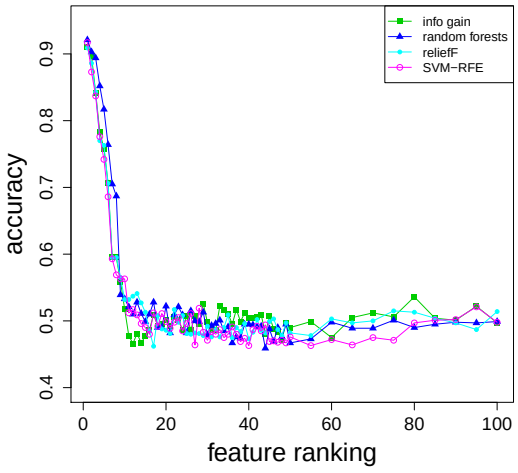
Figure 26: Comparison of **FFA** curves for the four different ranking methods for the **"single"** dataset. Each figure contains FFA curves for the same four feature ranking methods obtained by a different learning method.



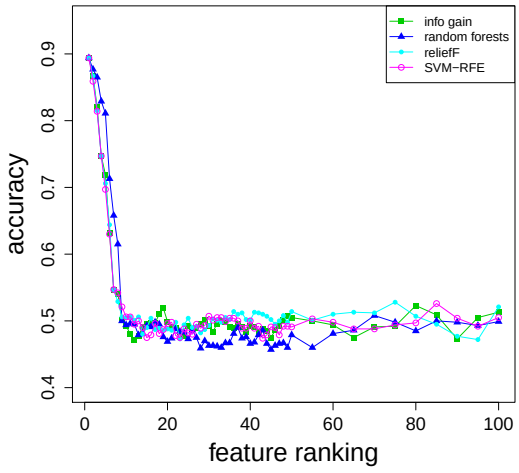
(a) RFA curves obtained with **Naïve Bayes**



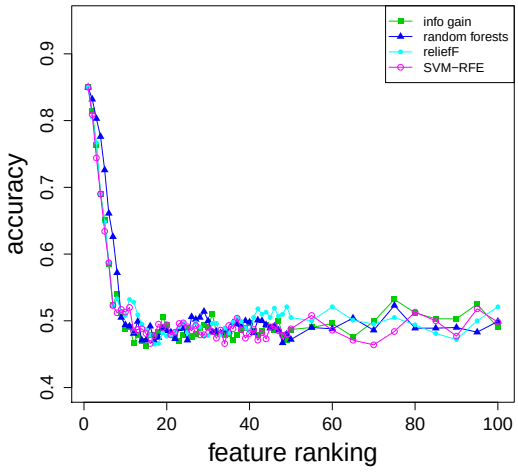
(b) RFA curves obtained with **Decision Trees**



(c) RFA curves obtained with **Random Forests**

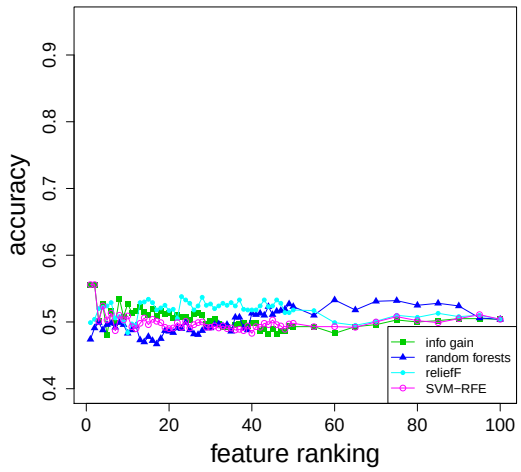


(d) RFA curves obtained with **SVMs**

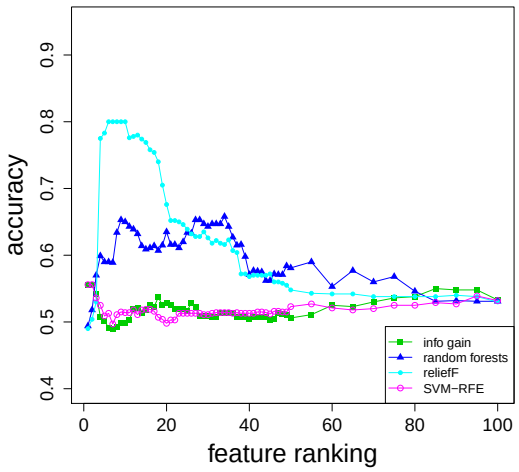


(e) RFA curves obtained with **k-NN**

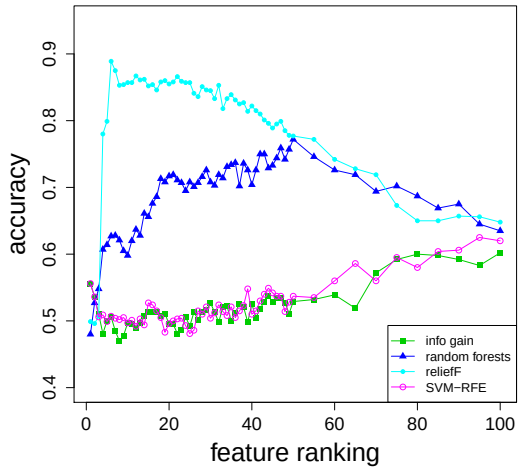
Figure 27: Comparison of **RFA** curves for the four different ranking methods for the **"single"** dataset. Each figure contains RFA curves for the same four feature ranking methods obtained by a different learning method.



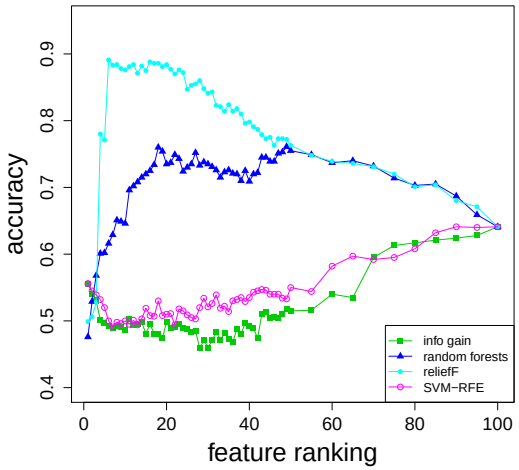
(a) FFA curves obtained with **Naïve Bayes**



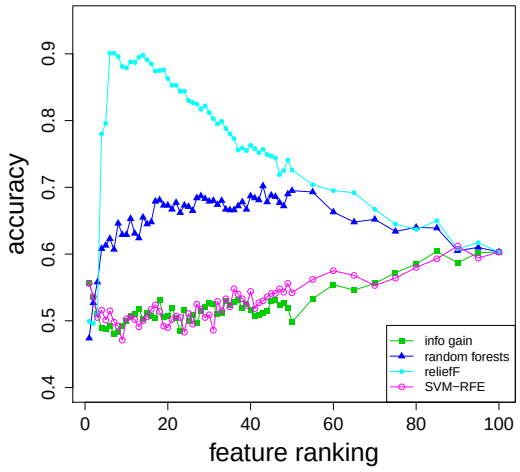
(b) FFA curves obtained with **Decision Trees**



(c) FFA curves obtained with **Random Forests**

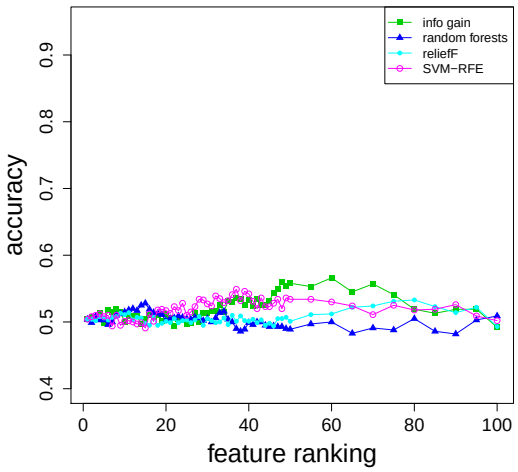


(d) FFA curves obtained with **SVMs**

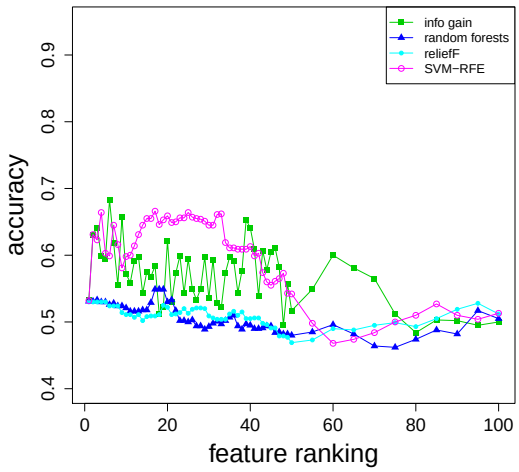


(e) FFA curves obtained with **k-NN**

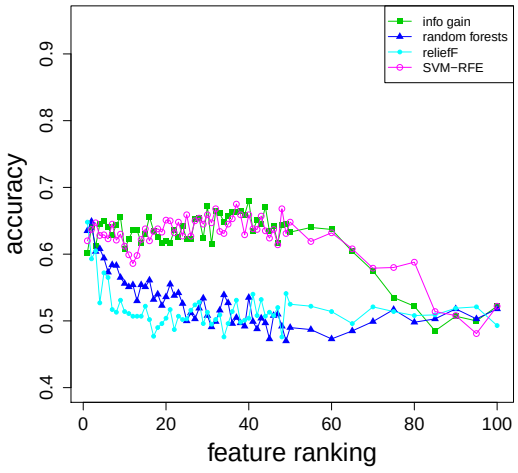
Figure 28: Comparison of **FFA** curves for the four different ranking methods for the **"pair"** dataset. Each figure contains FFA curves for the same four feature ranking methods obtained by a different learning method.



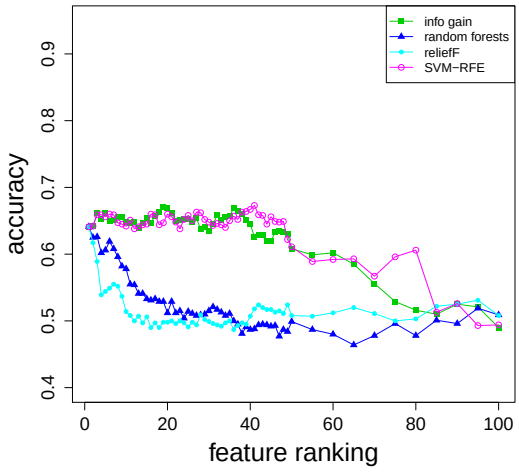
(a) RFA curves obtained with **Naïve Bayes**



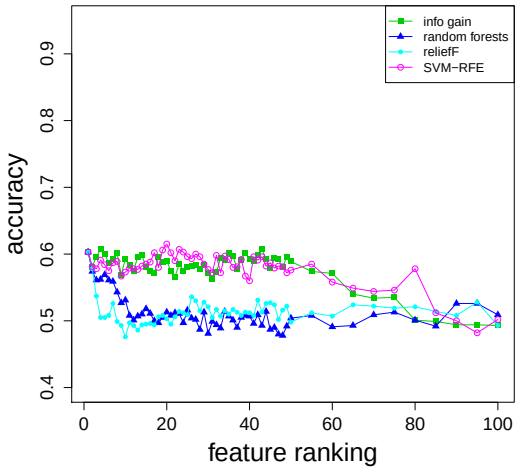
(b) RFA curves obtained with **Decision Trees**



(c) RFA curves obtained with **Random Forests**

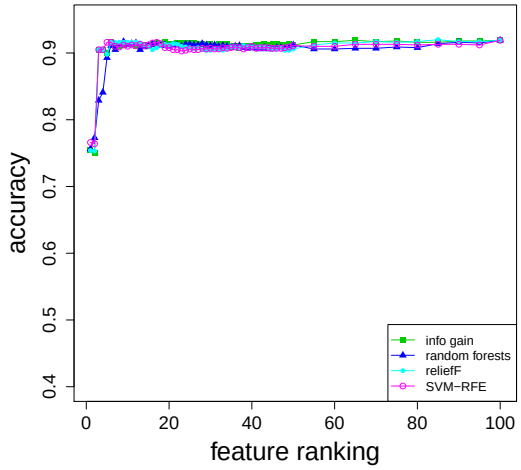


(d) RFA curves obtained with **SVMs**

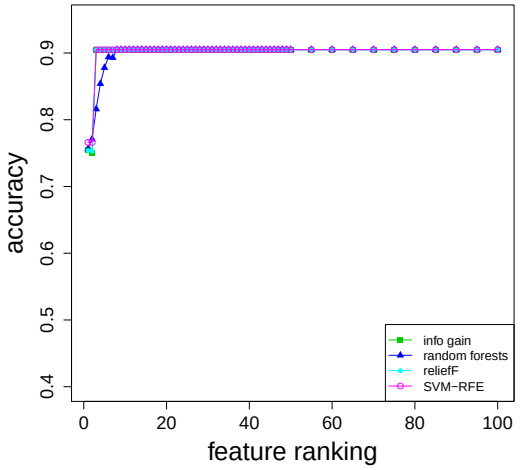


(e) RFA curves obtained with **k-NN**

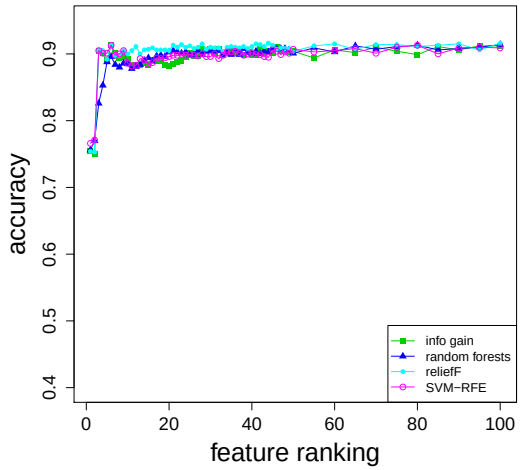
Figure 29: Comparison of **RFA** curves for the four different ranking methods for the **"pair"** dataset. Each figure contains RFA curves for the same four feature ranking methods obtained by a different learning method.



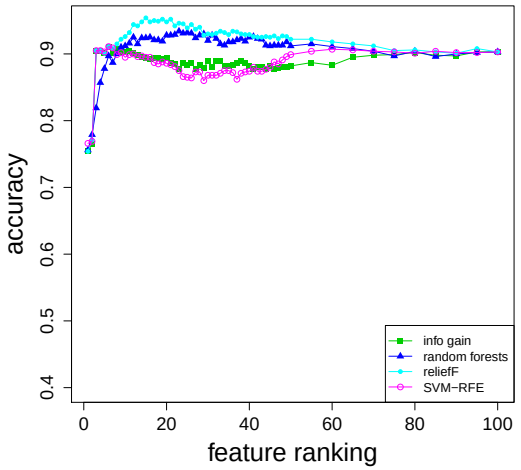
(a) FFA curves obtained with **Naïve Bayes**



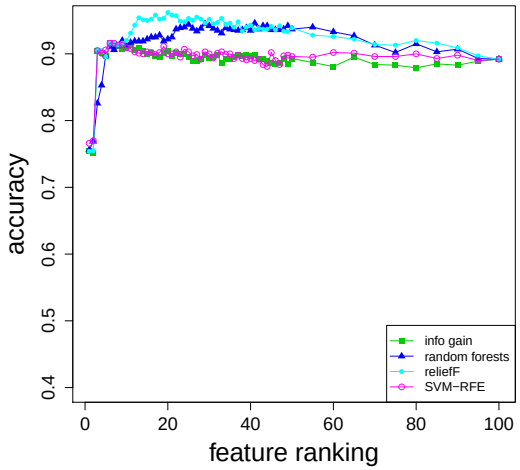
(b) FFA curves obtained with **Decision Trees**



(c) FFA curves obtained with **Random Forests**

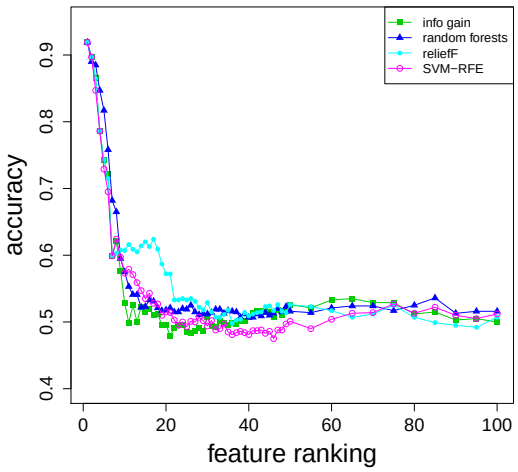


(d) FFA curves obtained with **SVMs**

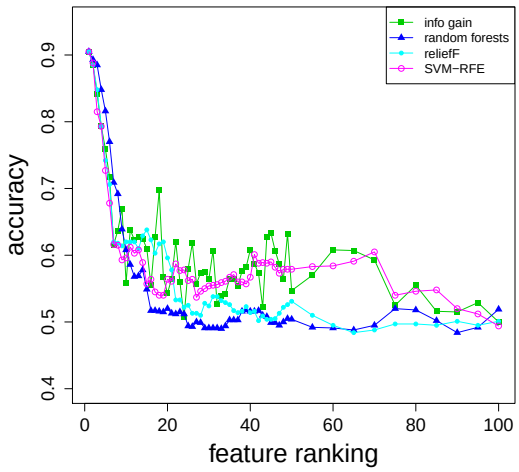


(e) FFA curves obtained with **k-NN**

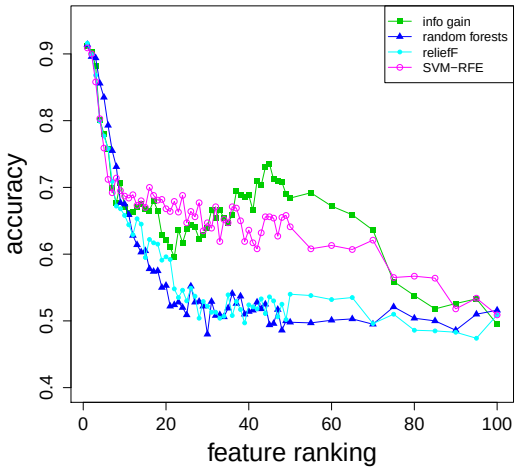
Figure 30: Comparison of **FFA** curves for the four different ranking methods for the **"combined"** dataset. Each figure contains FFA curves for the same four feature ranking methods obtained by a different learning method.



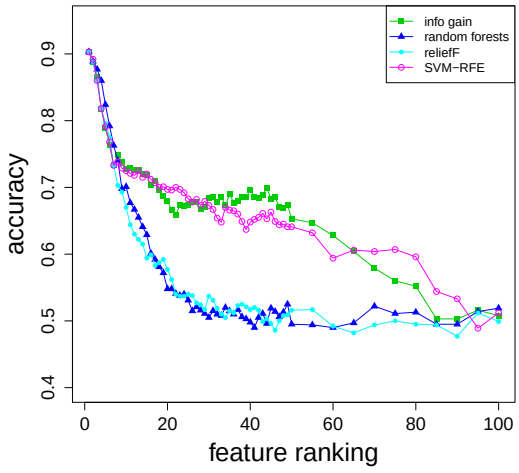
(a) RFA curves obtained with **Naïve Bayes**



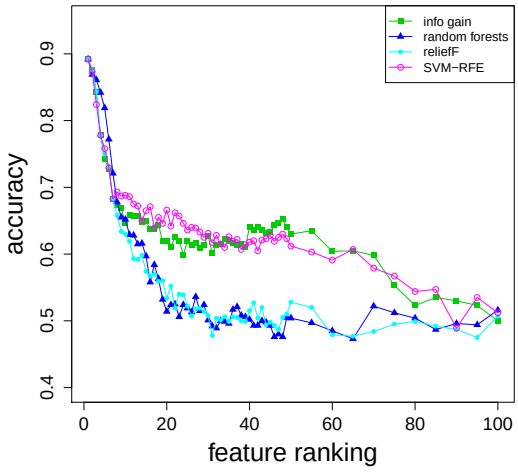
(b) RFA curves obtained with **Decision Trees**



(c) RFA curves obtained with **Random Forests**



(d) RFA curves obtained with **SVMs**



(e) RFA curves obtained with **k-NN**

Figure 31: Comparison of **RFA** curves for the four different ranking methods for the **"combined"** dataset. Each figure contains RFA curves for the same four feature ranking methods obtained by a different learning method.

B Complete Experimental Results

Here we present the complete set of results of the experiments described in Chapter 6. They include three different sets of results. First, in Section B.1 we present all of the obtained FFA and RFA curves for the feature ranking ensembles experiments. Next, in Section B.2 we present the FFA and RFA curves of the comparative analysis of feature ranking algorithms on various benchmark datasets. The final set of results are related to the case study of aggressiveness of embryonal tumours. In Section B.3 we present the obtained FFA and RFA curves of the individual analysis of the ET data, while in Section B.4 we present the gene networks constructed by applying the STRING database to the topmost ranked genes. The results from the integrated analysis of the ET datasets (with aggregated feature rankings) are presented in Section B.5 (FFA and RFA curves) and in Section B.6 (gene networks).

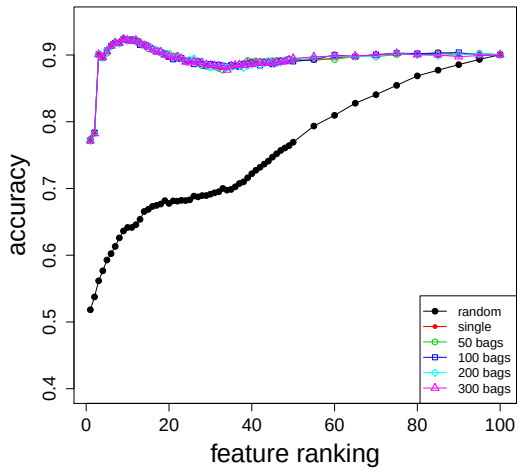
B.1 Feature Ranking Ensembles

Feature Ranking Ensembles (FRE) are aggregated feature rankings $\mathbf{R}_{agg}$, produced by combining base rankings from different data subsamples. The intuition for constructing ensembles of feature rankings is similar to the one for constructing ensembles of predictive models. The aim of our experiments is to answer the practical question of whether, and under what conditions there is an advantage in using feature ranking ensembles.

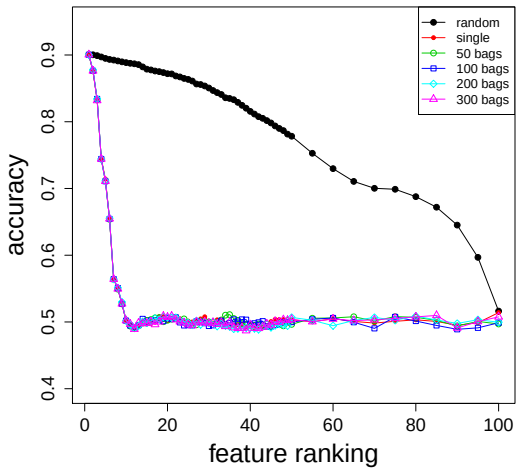
We considered the following specific experimental conditions:

- Synthetic data: "single", "pair" and "combined",
- Different number k of bootstrap replicates: 50, 100, 200 and 300,
- Different base rankers: Info Gain, Random Forests, ReliefF and SVM-RFE,
- Different aggregation functions: mean, median, maximal and minimal.

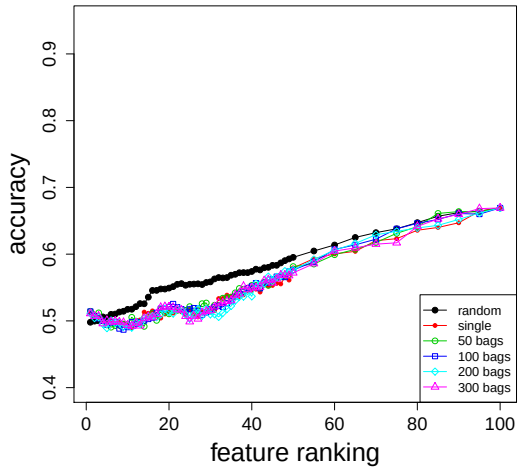
We present the FFA and RFA curves obtained under different combination of the above experimental conditions. Each figure shows the FFA/RFA curves for the three synthetic datasets, for a particular combination of a method for generating base rankings and aggregation function. This yields a total of 16 figures.



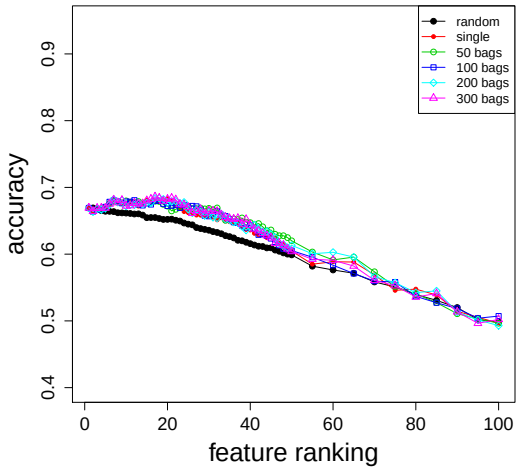
(a) FFA curves of the "single" data.



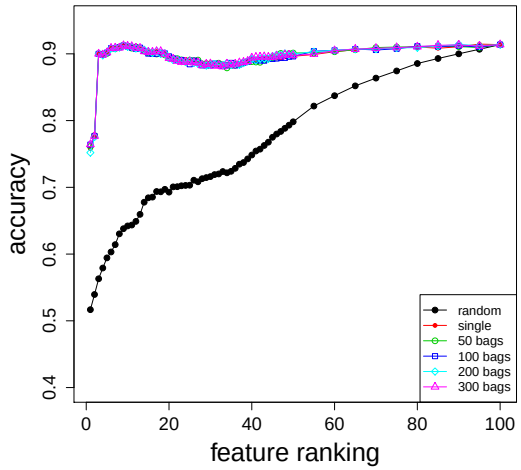
(b) RFA curves of the "single" data.



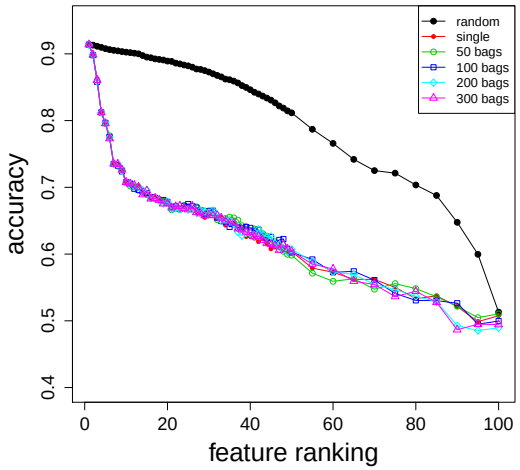
(c) FFA curves of the "pair" data.



(d) RFA curves of the "pair" data.

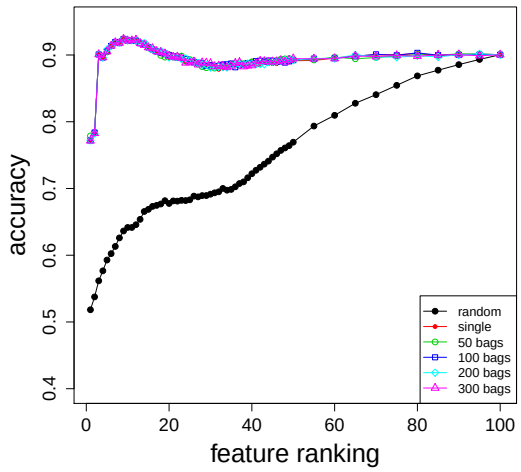


(e) FFA curves of the "combined" data.

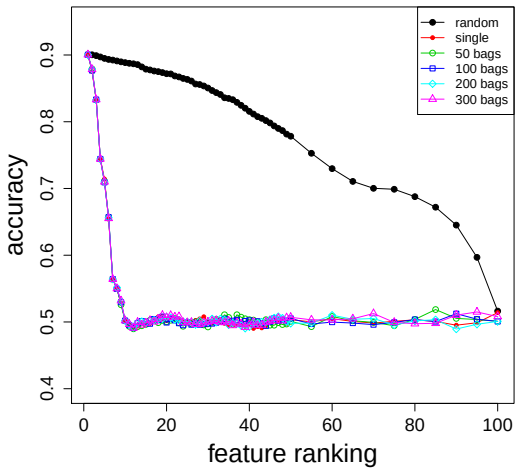


(f) RFA curves of the "combined" data.

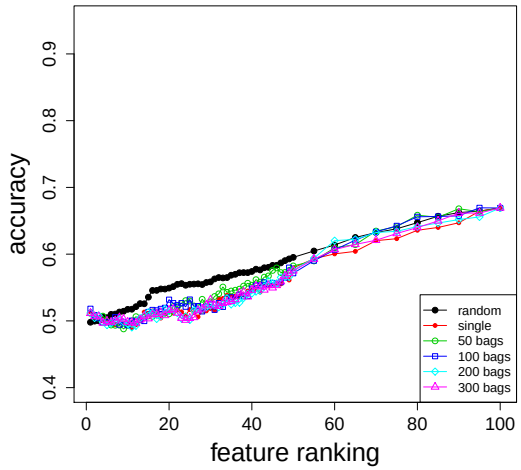
Figure 32: Comparison of FFA (left) and RFA (right) curves of **single** and **ensemble** rankings produced by considering a **different number** k of base rankings. The base rankings are produced by **info gain** and aggregated with the **mean** function.



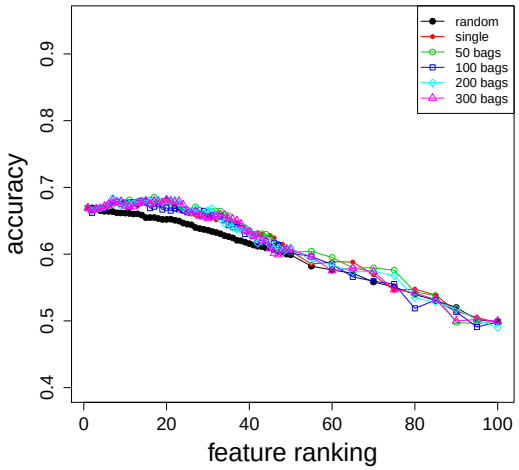
(a) FFA curves of the "single" data.



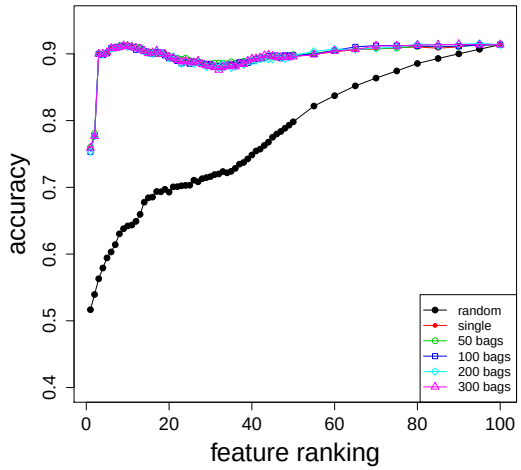
(b) RFA curves of the "single" data.



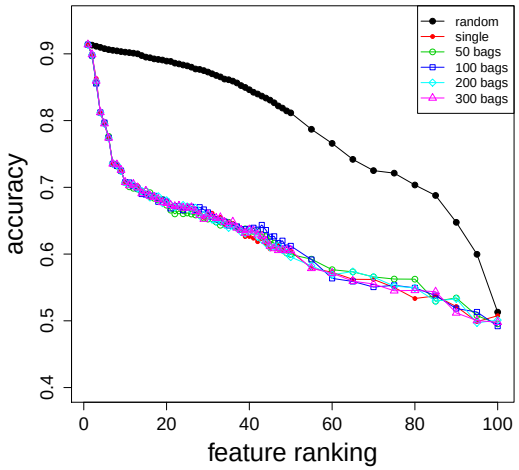
(c) FFA curves of the "pair" data.



(d) RFA curves of the "pair" data.

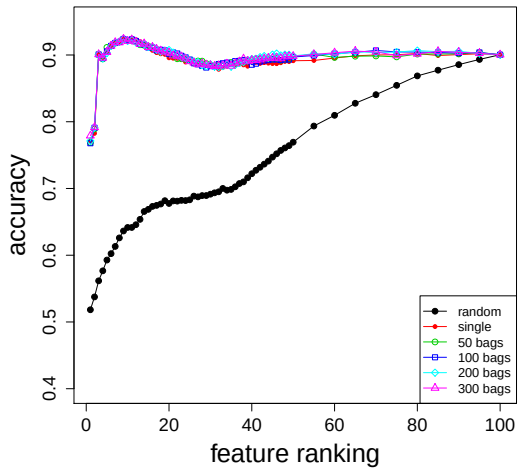


(e) FFA curves of the "combined" data.

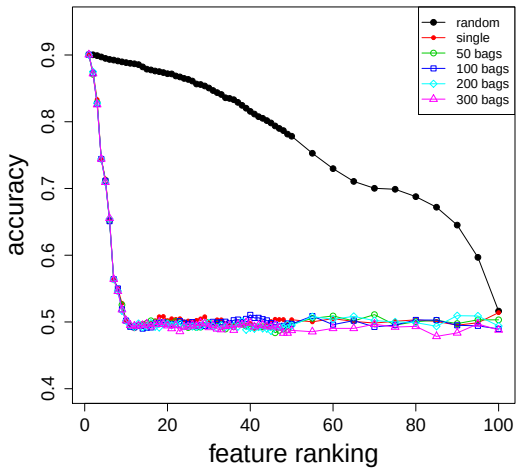


(f) RFA curves of the "combined" data.

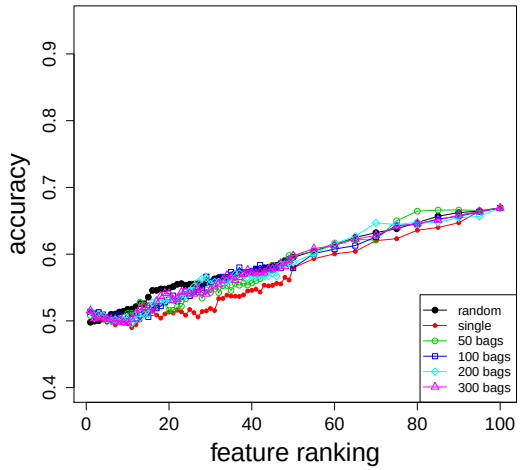
Figure 33: Comparison of FFA (left) and RFA (right) curves of **single** and **ensemble** rankings produced by considering a **different number** k of base rankings. The base rankings are produced by **info gain** and aggregated with the **median** function.



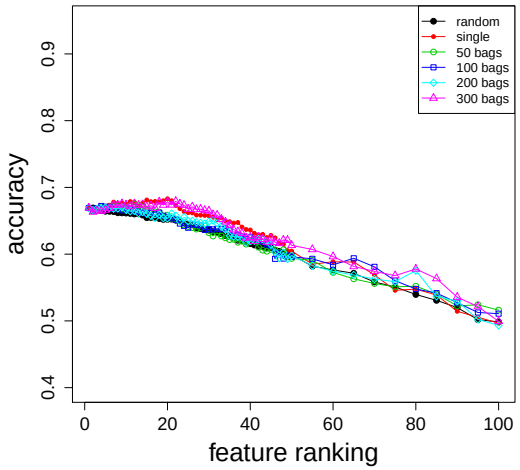
(a) FFA curves of the "single" data.



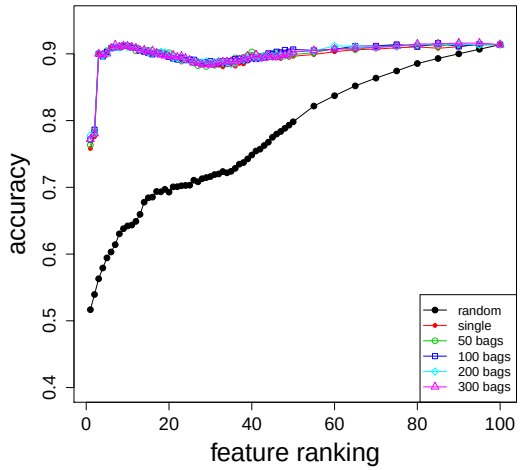
(b) RFA curves of the "single" data.



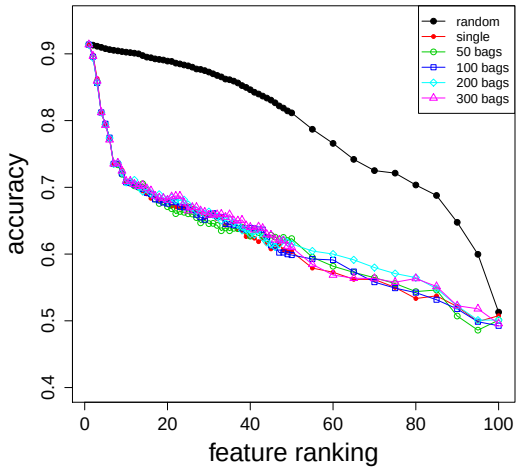
(c) FFA curves of the "pair" data.



(d) RFA curves of the "pair" data.

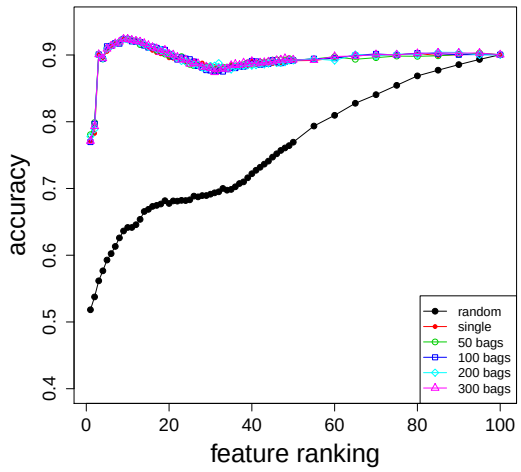


(e) FFA curves of the "combined" data.

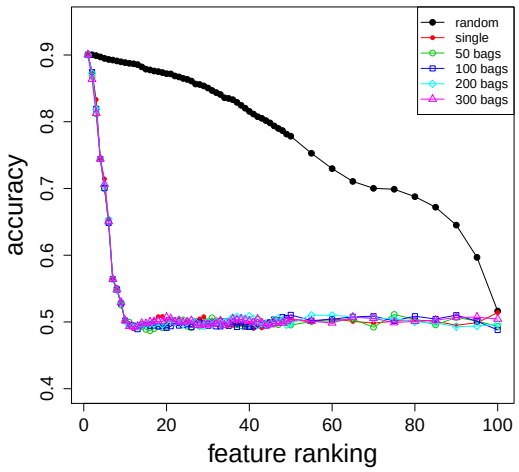


(f) RFA curves of the "combined" data.

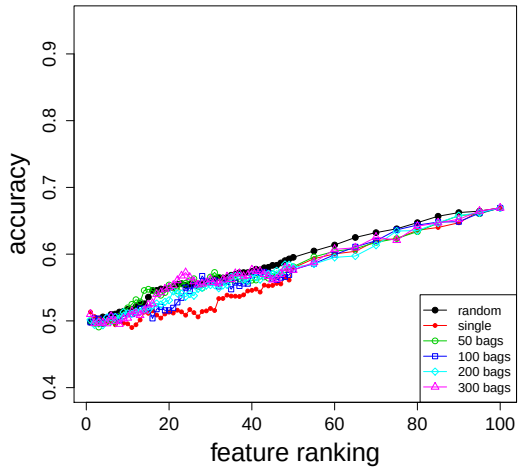
Figure 34: Comparison of FFA (left) and RFA (right) curves of **single** and **ensemble** rankings produced by considering a **different number** k of base rankings. The base rankings are produced by **info gain** and aggregated with the **min** function.



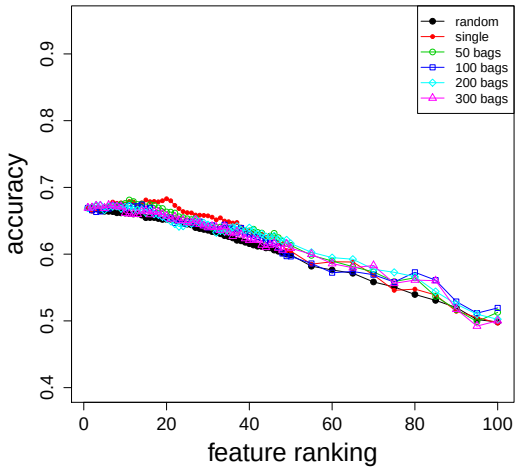
(a) FFA curves of the "single" data.



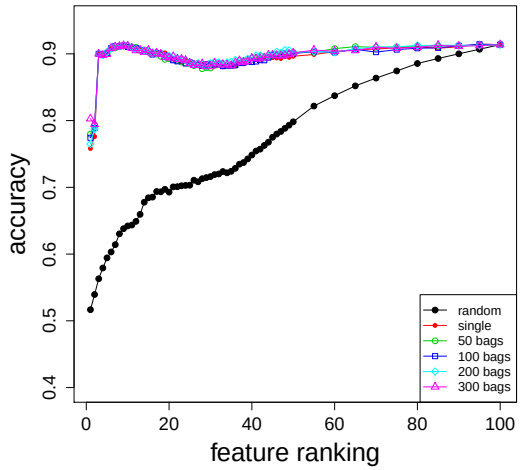
(b) RFA curves of the "single" data.



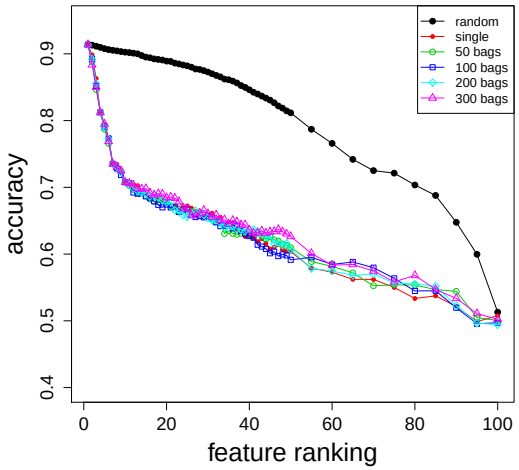
(c) FFA curves of the "pair" data.



(d) RFA curves of the "pair" data.

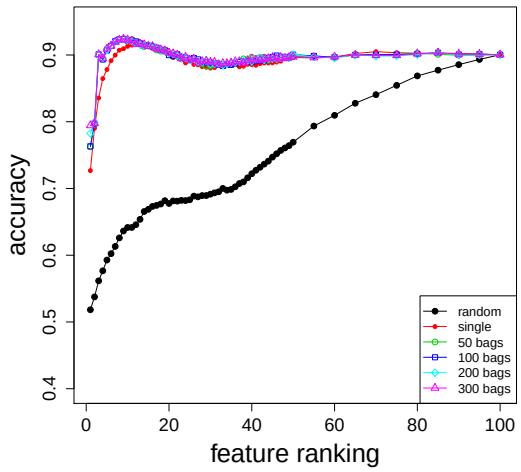


(e) FFA curves of the "combined" data.

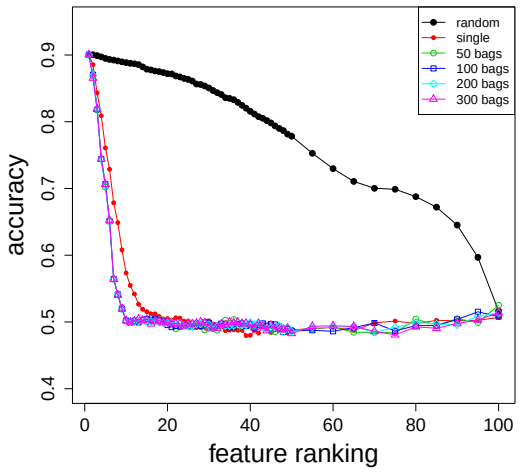


(f) RFA curves of the "combined" data.

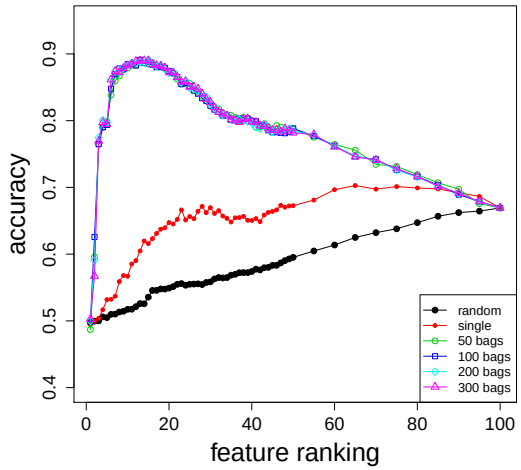
Figure 35: Comparison of FFA (left) and RFA (right) curves of **single** and **ensemble** rankings produced by considering a **different number** k of base rankings. The base rankings are produced by **info gain** and aggregated with the **max** function.



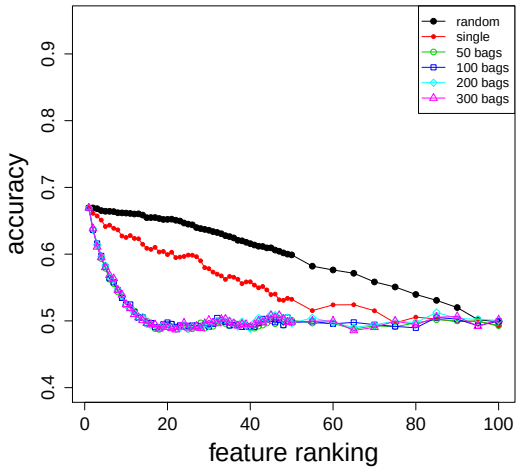
(a) FFA curves of the "single" data.



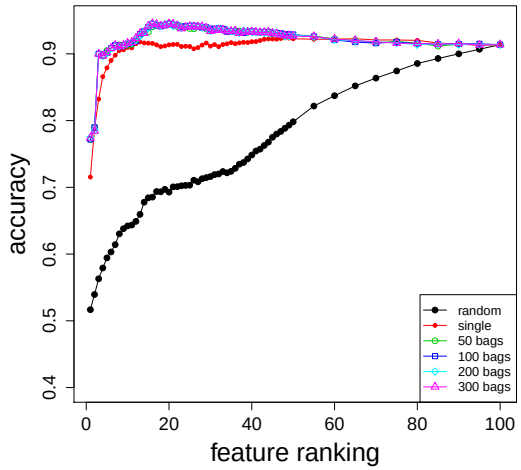
(b) RFA curves of the "single" data.



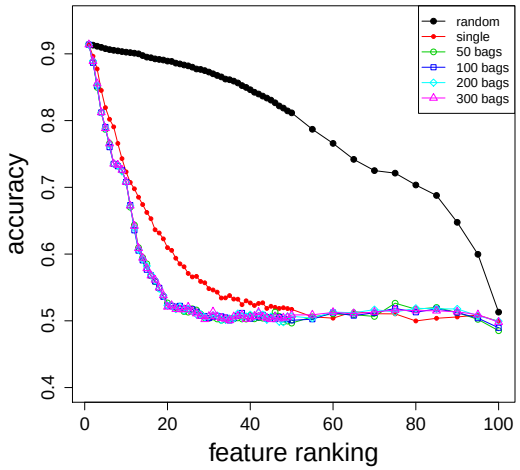
(c) FFA curves of the "pair" data.



(d) RFA curves of the "pair" data.

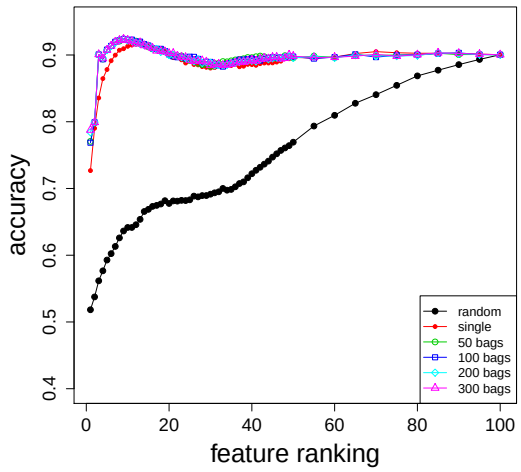


(e) FFA curves of the "combined" data.

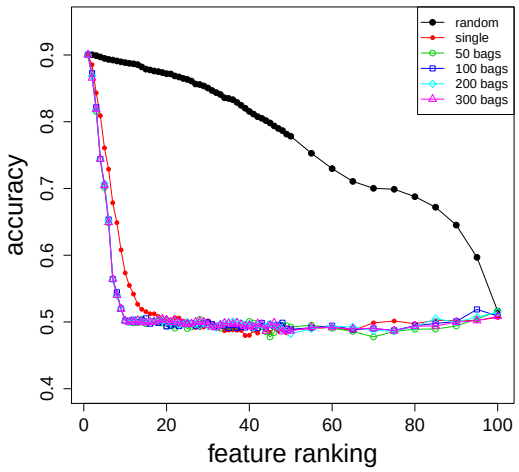


(f) RFA curves of the "combined" data.

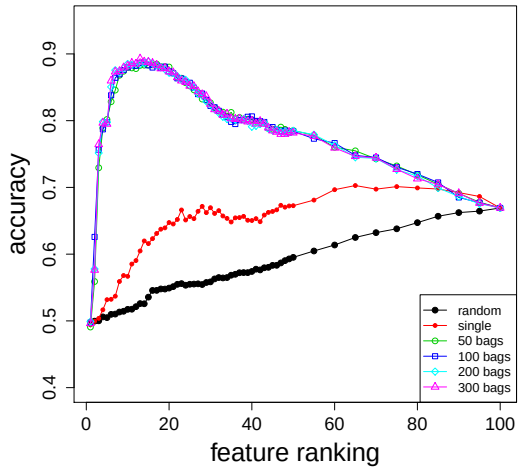
Figure 36: Comparison of FFA (left) and RFA (right) curves of **single** and **ensemble** rankings produced by considering a **different number** k of base rankings. The base rankings are produced by **random forests** and aggregated with the **mean** function.



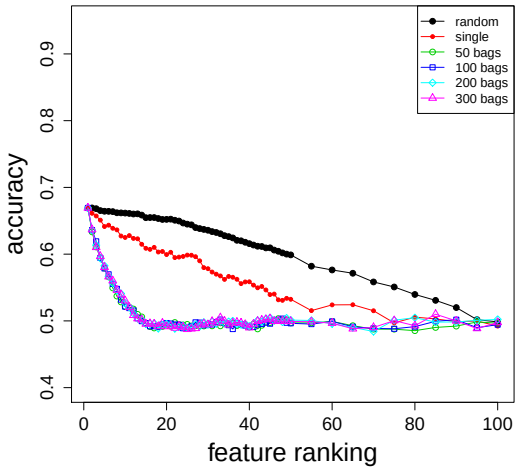
(a) FFA curves of the "single" data.



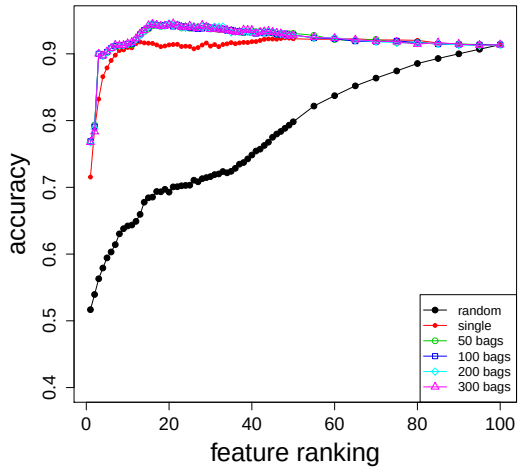
(b) RFA curves of the "single" data.



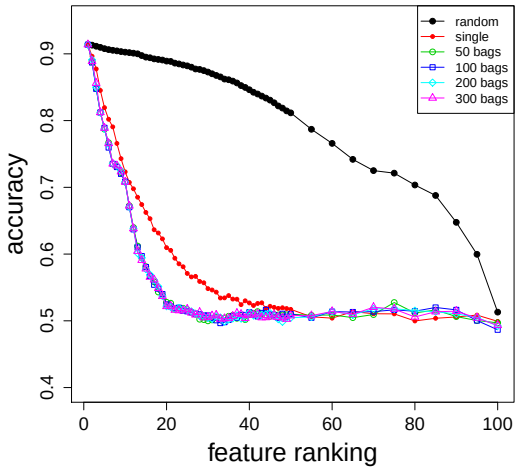
(c) FFA curves of the "pair" data.



(d) RFA curves of the "pair" data.

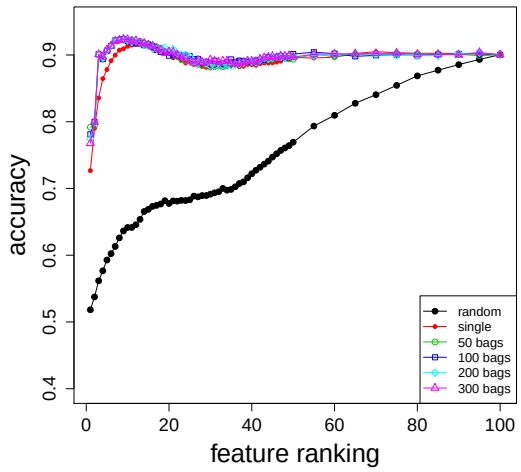


(e) FFA curves of the "combined" data.

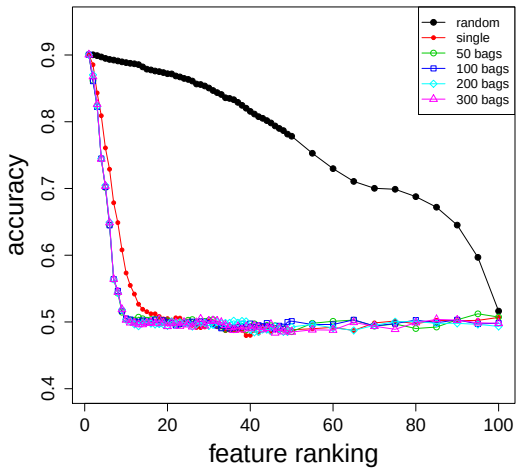


(f) RFA curves of the "combined" data.

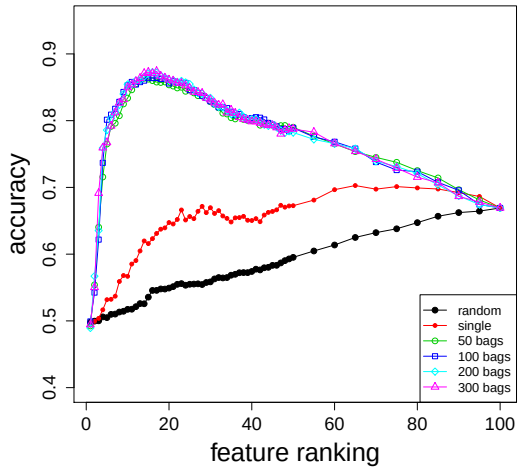
Figure 37: Comparison of FFA (left) and RFA (right) curves of **single** and **ensemble** rankings produced by considering a **different number** k of base rankings. The base rankings are produced by **random forests** and aggregated with the **median** function.



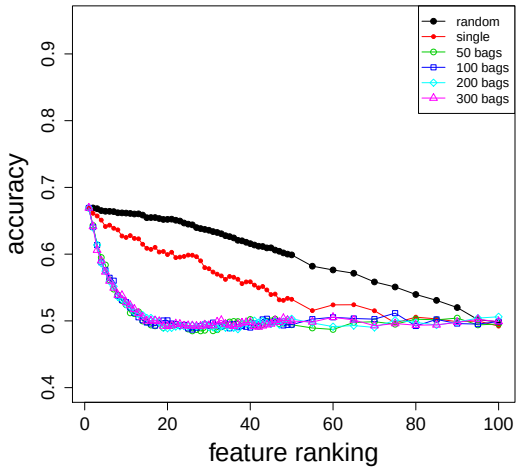
(a) FFA curves of the "single" data.



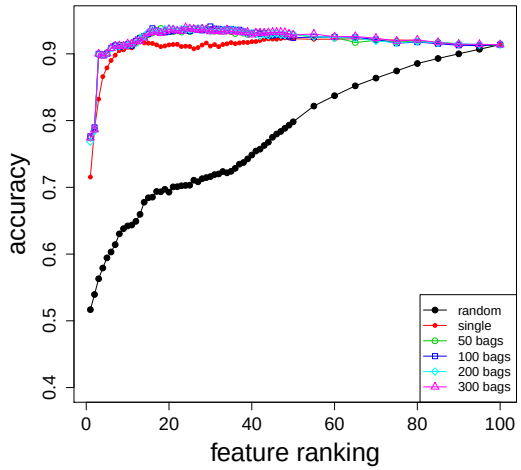
(b) RFA curves of the "single" data.



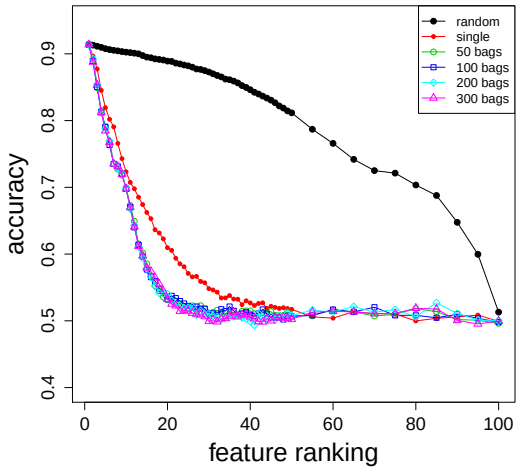
(c) FFA curves of the "pair" data.



(d) RFA curves of the "pair" data.

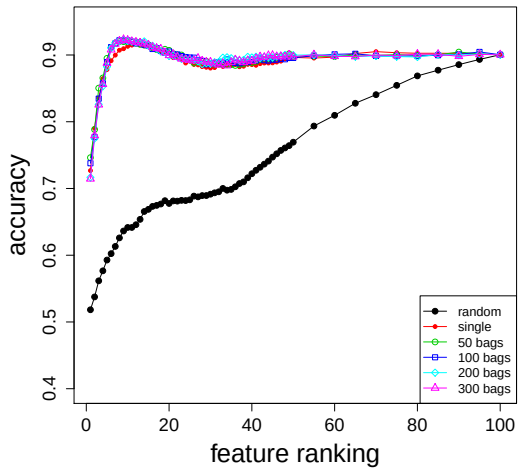


(e) FFA curves of the "combined" data.

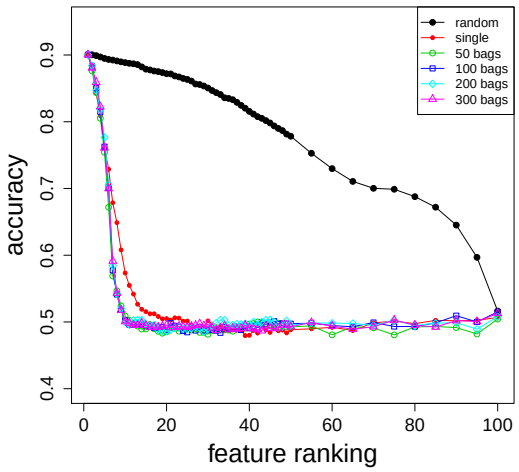


(f) RFA curves of the "combined" data.

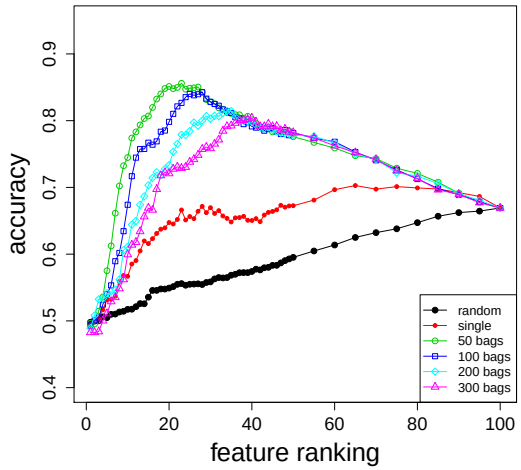
Figure 38: Comparison of FFA (left) and RFA (right) curves of **single** and **ensemble** rankings produced by considering a **different number** k of base rankings. The base rankings are produced by **random forests** and aggregated with the **min** function.



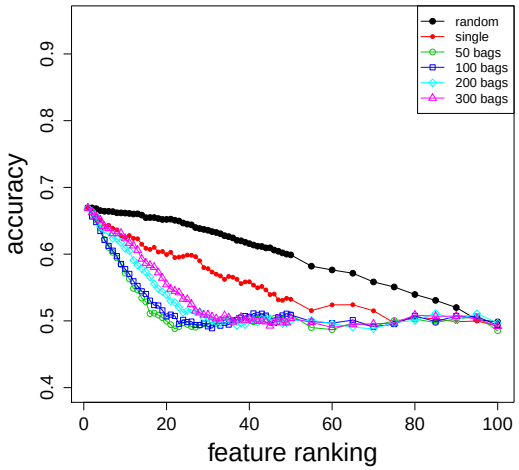
(a) FFA curves of the "single" data.



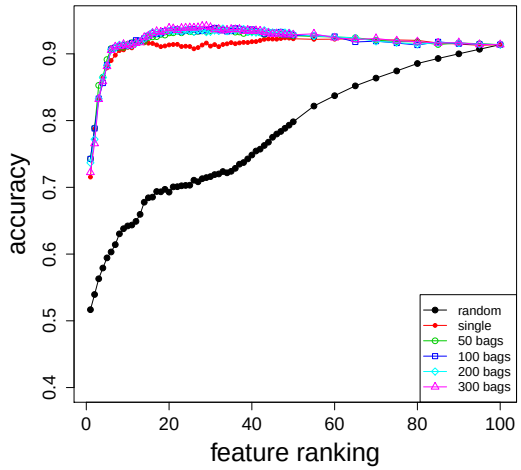
(b) RFA curves of the "single" data.



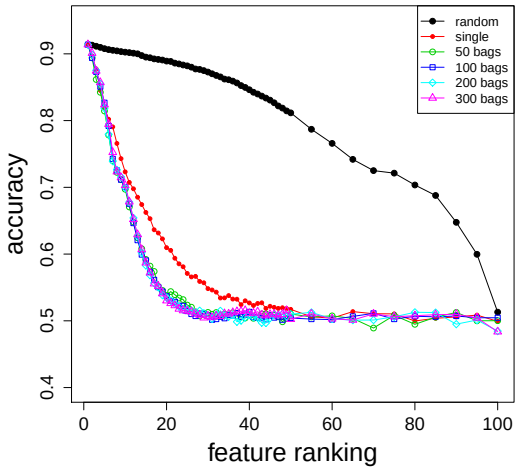
(c) FFA curves of the "pair" data.



(d) RFA curves of the "pair" data.

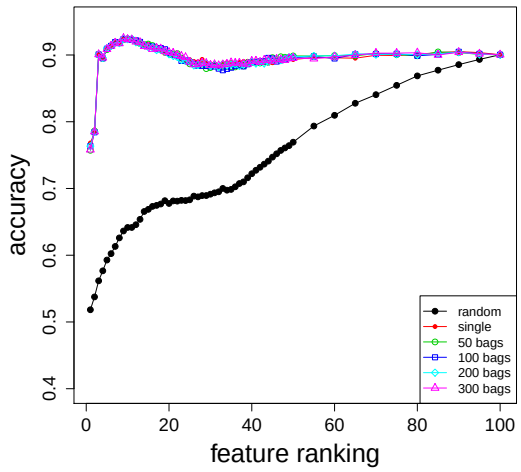


(e) FFA curves of the "combined" data.

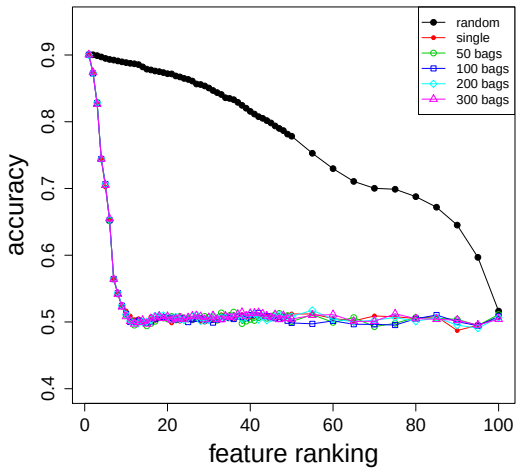


(f) RFA curves of the "combined" data.

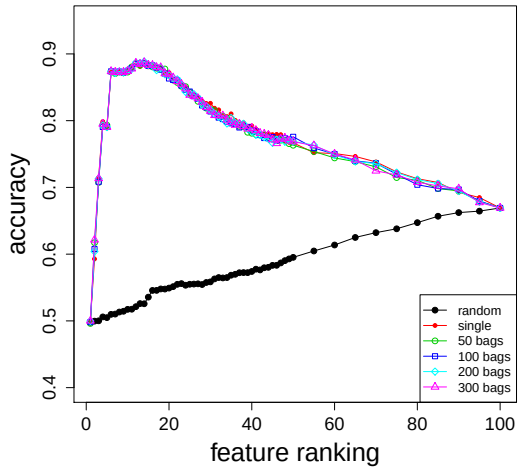
Figure 39: Comparison of FFA (left) and RFA (right) curves of **single** and **ensemble** rankings produced by considering a **different number** k of base rankings. The base rankings are produced by **random forests** and aggregated with the **max** function.



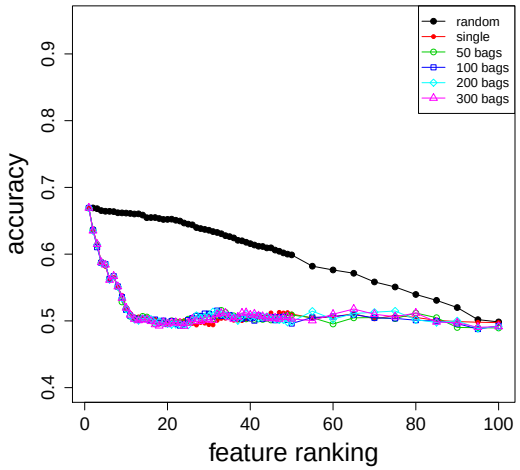
(a) FFA curves of the "single" data.



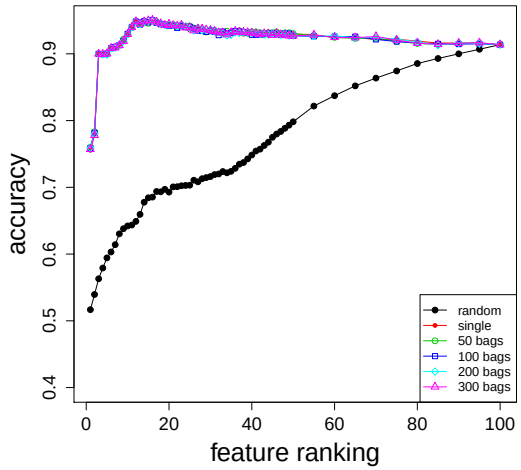
(b) RFA curves of the "single" data.



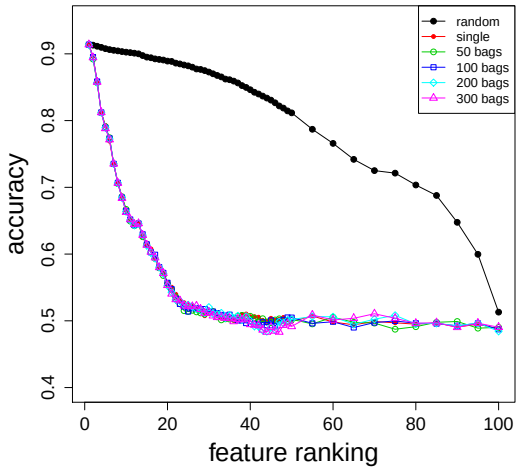
(c) FFA curves of the "pair" data.



(d) RFA curves of the "pair" data.

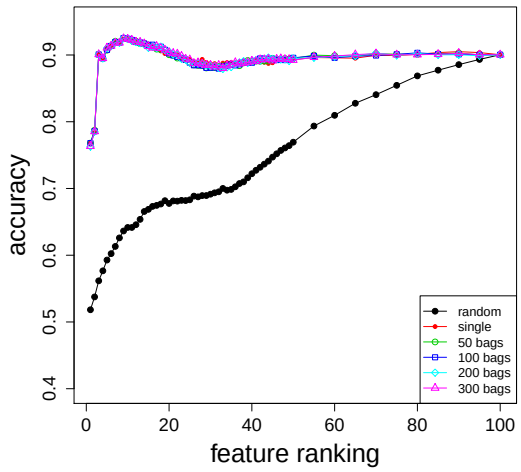


(e) FFA curves of the "combined" data.

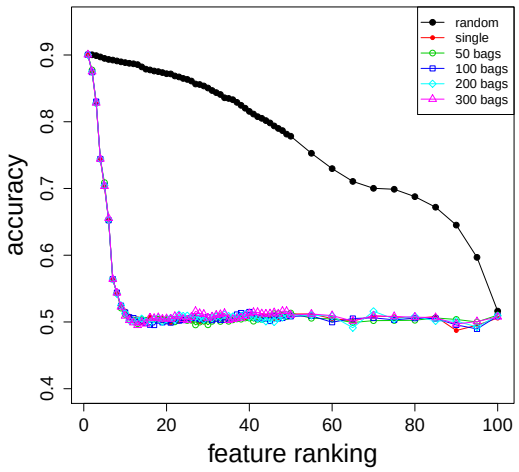


(f) RFA curves of the "combined" data.

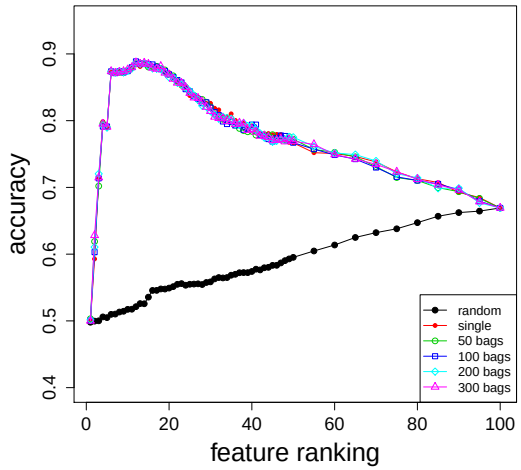
Figure 40: Comparison of FFA (left) and RFA (right) curves of **single** and **ensemble** rankings produced by considering a **different number** k of base rankings. The base rankings are produced by **reliefF** and aggregated with the **mean** function.



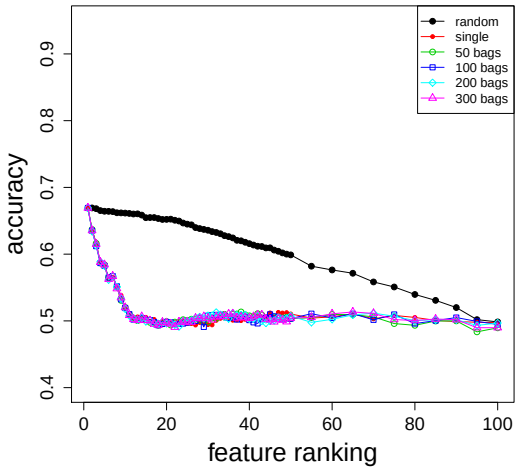
(a) FFA curves of the "single" data.



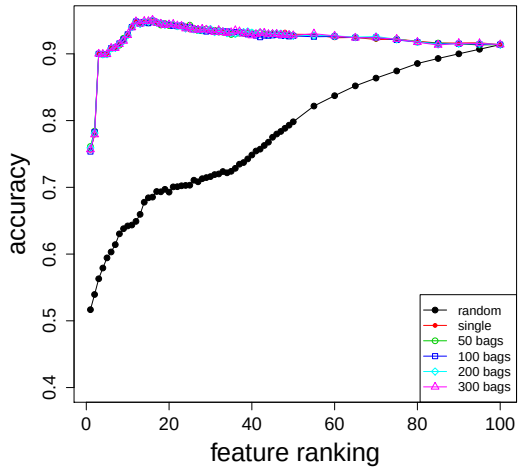
(b) RFA curves of the "single" data.



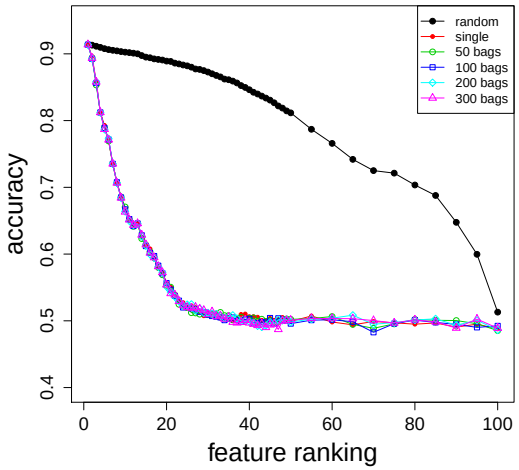
(c) FFA curves of the "pair" data.



(d) RFA curves of the "pair" data.

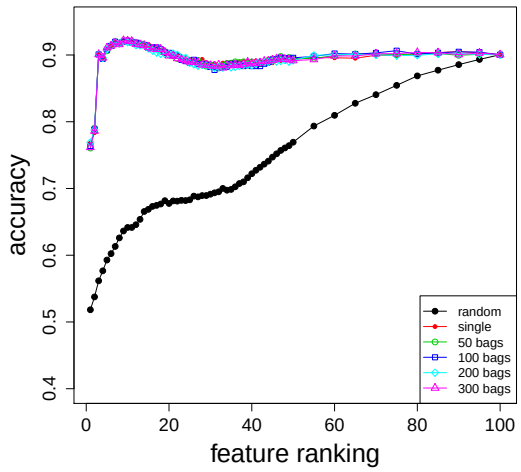


(e) FFA curves of the "combined" data.

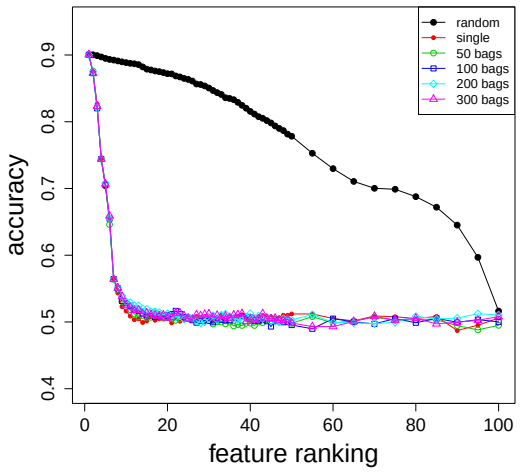


(f) RFA curves of the "combined" data.

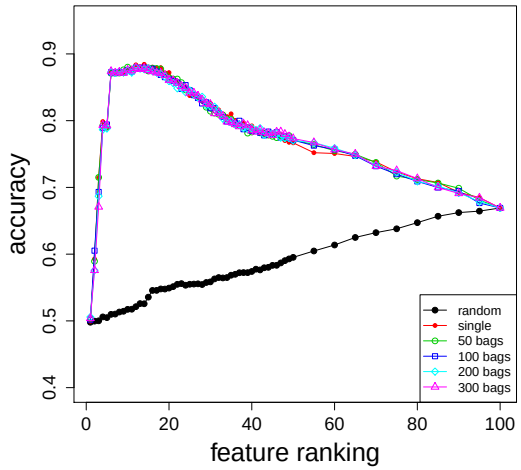
Figure 41: Comparison of FFA (left) and RFA (right) curves of **single** and **ensemble** rankings produced by considering a **different number** k of base rankings. The base rankings are produced by **reliefF** and aggregated with the **median** function.



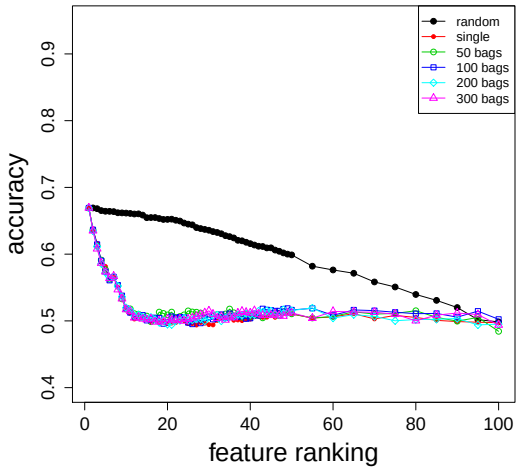
(a) FFA curves of the "single" data.



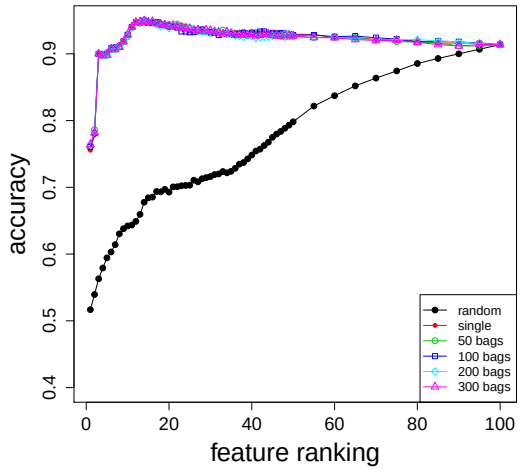
(b) RFA curves of the "single" data.



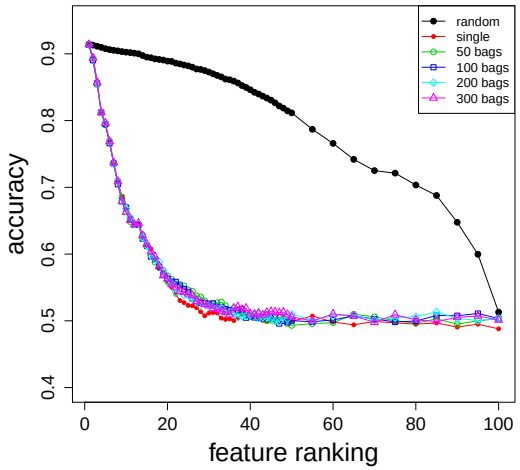
(c) FFA curves of the "pair" data.



(d) RFA curves of the "pair" data.

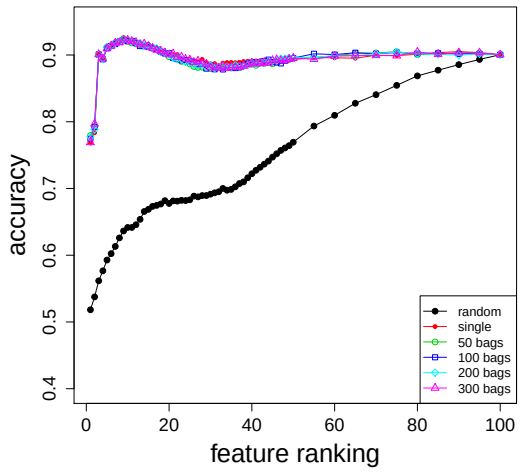


(e) FFA curves of the "combined" data.

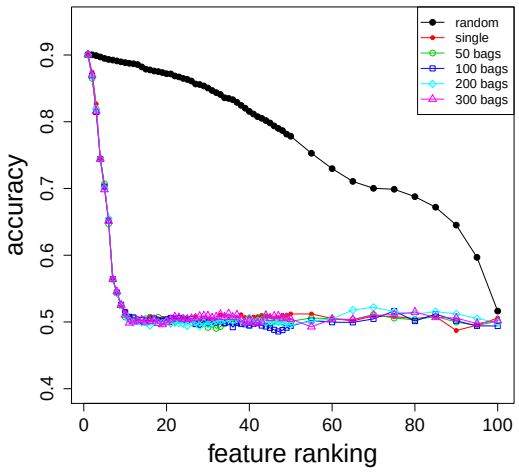


(f) RFA curves of the "combined" data.

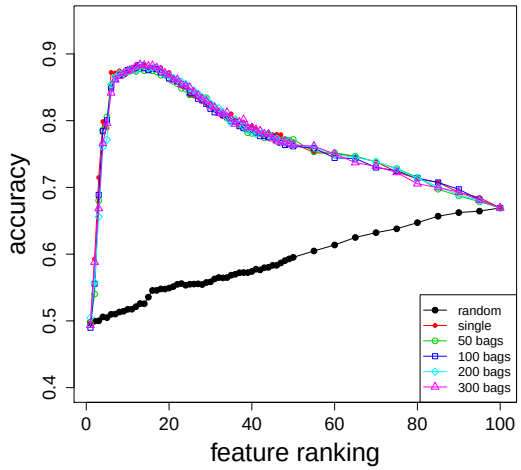
Figure 42: Comparison of FFA (left) and RFA (right) curves of **single** and **ensemble** rankings produced by considering a **different number** k of base rankings. The base rankings are produced by **reliefF** and aggregated with the **min** function.



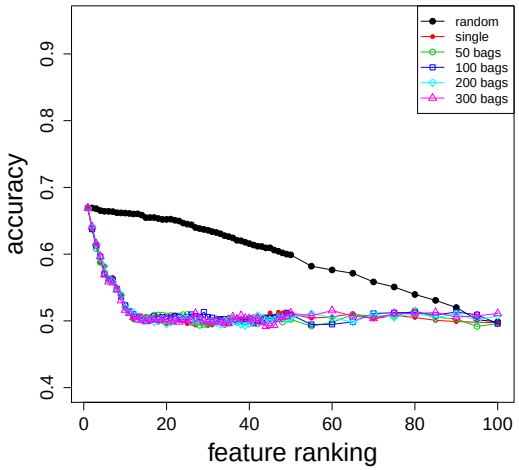
(a) FFA curves of the "single" data.



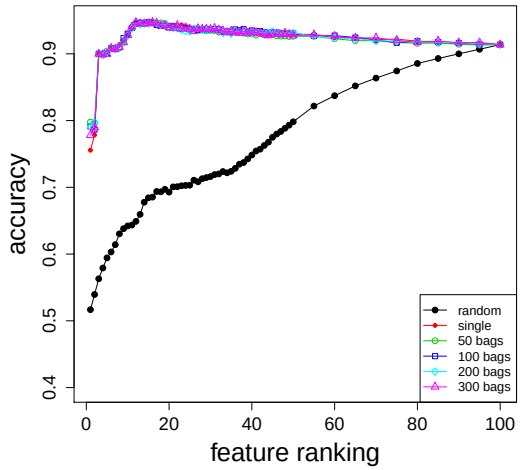
(b) RFA curves of the "single" data.



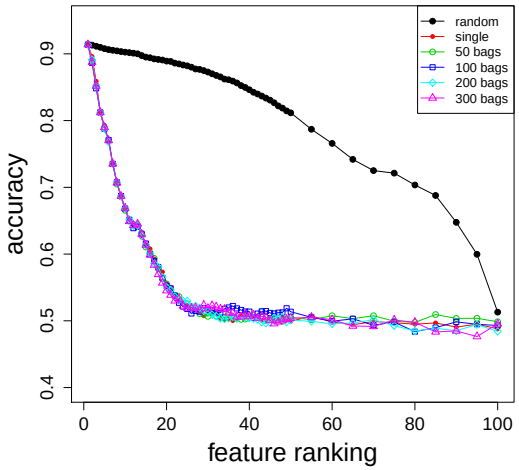
(c) FFA curves of the "pair" data.



(d) RFA curves of the "pair" data.

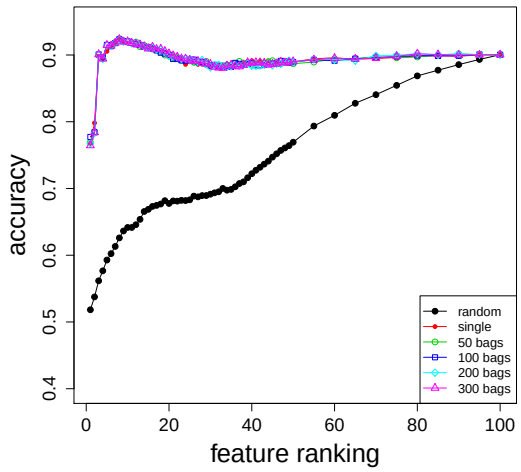


(e) FFA curves of the "combined" data.

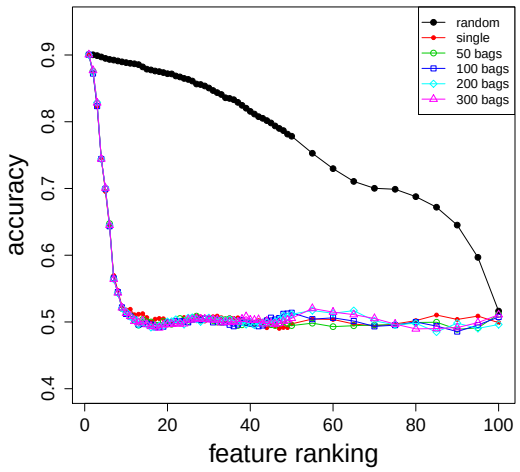


(f) RFA curves of the "combined" data.

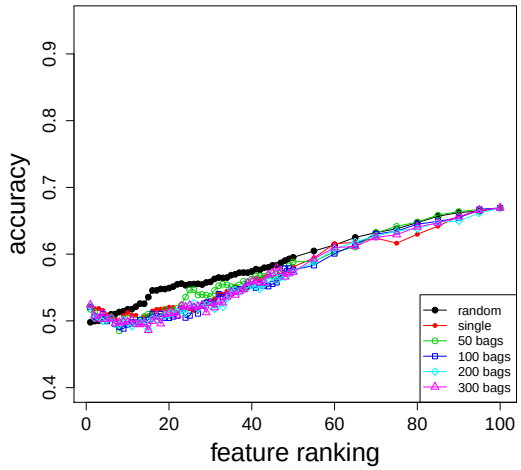
Figure 43: Comparison of FFA (left) and RFA (right) curves of **single** and **ensemble** rankings produced by considering a **different number** k of base rankings. The base rankings are produced by **reliefF** and aggregated with the **max** function.



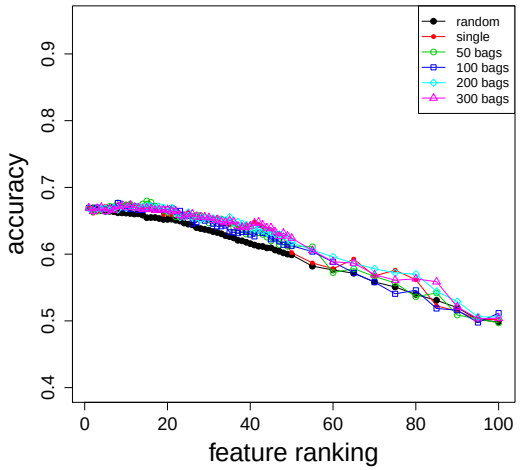
(a) FFA curves of the "single" data.



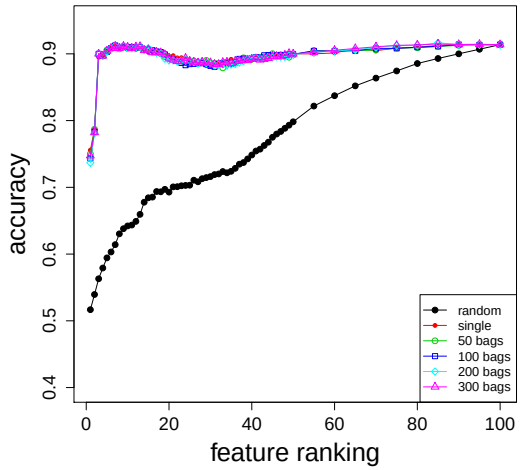
(b) RFA curves of the "single" data.



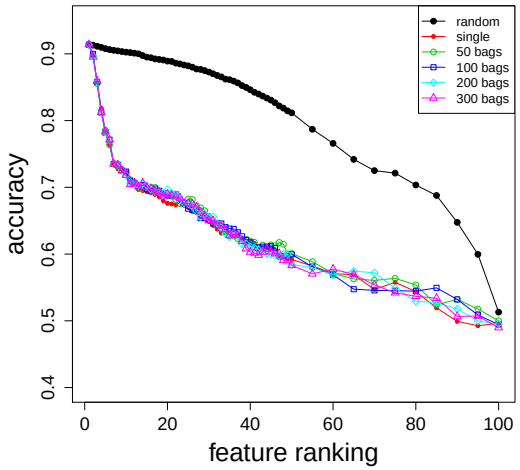
(c) FFA curves of the "pair" data.



(d) RFA curves of the "pair" data.

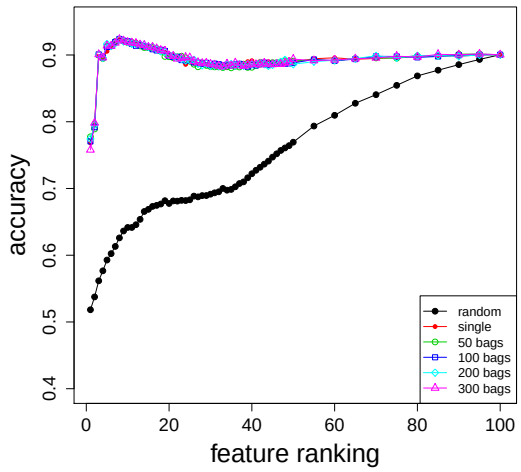


(e) FFA curves of the "combined" data.

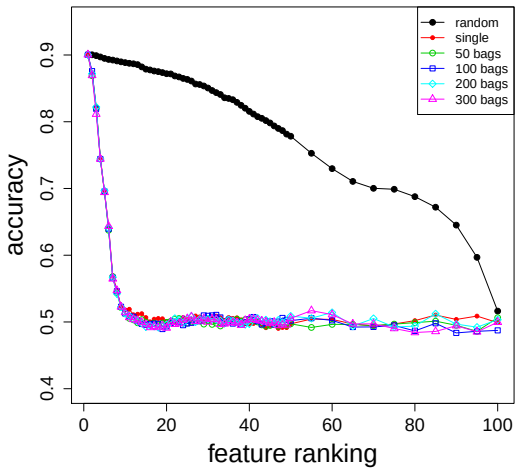


(f) RFA curves of the "combined" data.

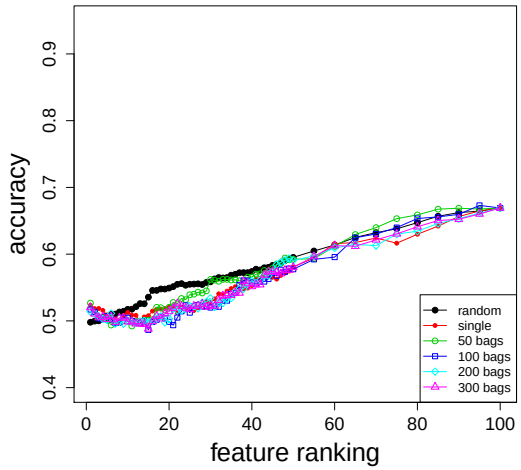
Figure 44: Comparison of FFA (left) and RFA (right) curves of **single** and **ensemble** rankings produced by considering a **different number** k of base rankings. The base rankings are produced by **SVM-RFE** and aggregated with the **mean** function.



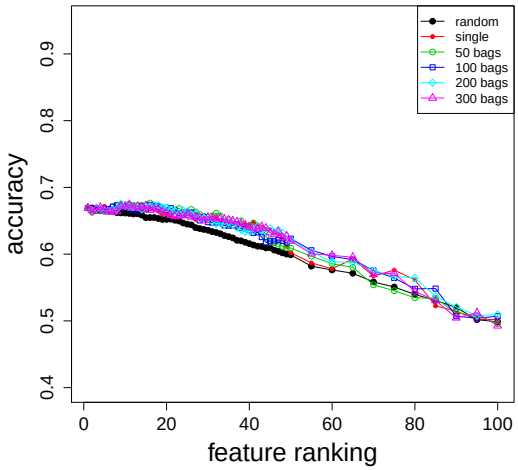
(a) FFA curves of the "single" data.



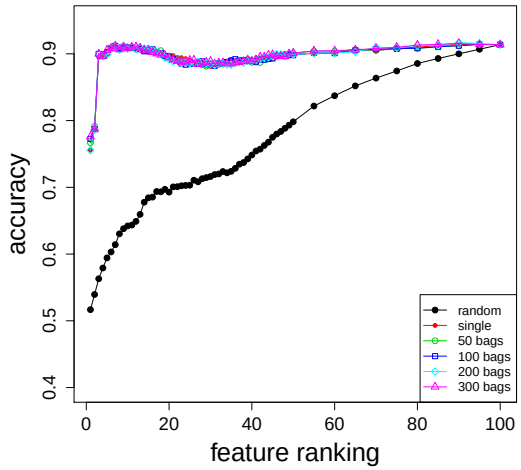
(b) RFA curves of the "single" data.



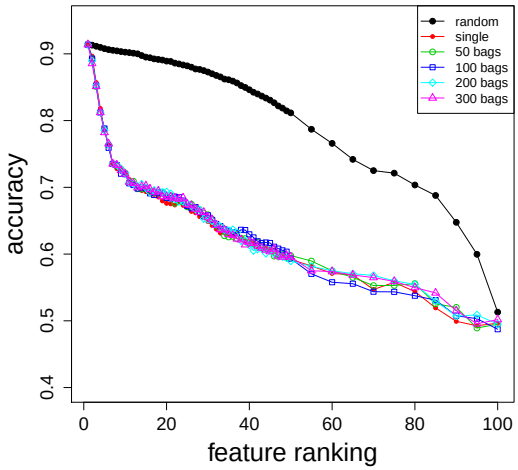
(c) FFA curves of the "pair" data.



(d) RFA curves of the "pair" data.

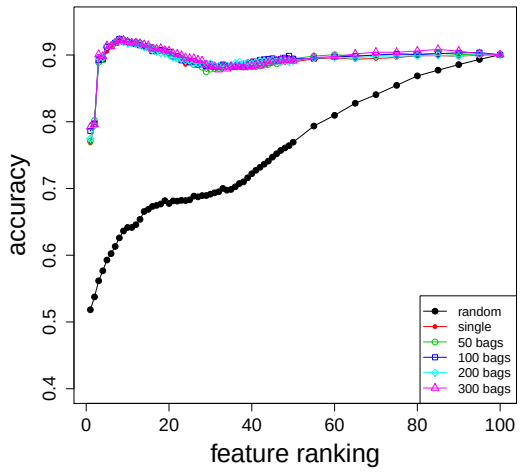


(e) FFA curves of the "combined" data.

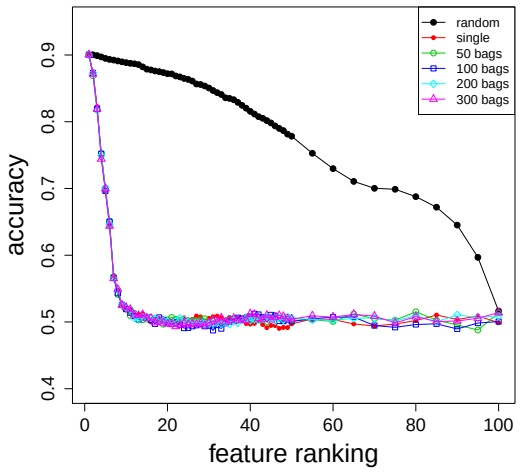


(f) RFA curves of the "combined" data.

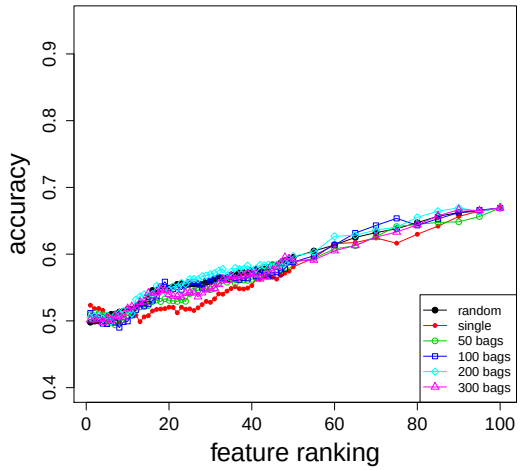
Figure 45: Comparison of FFA (left) and RFA (right) curves of **single** and **ensemble** rankings produced by considering a **different number** k of base rankings. The base rankings are produced by **SVM-RFE** and aggregated with the **median** function.



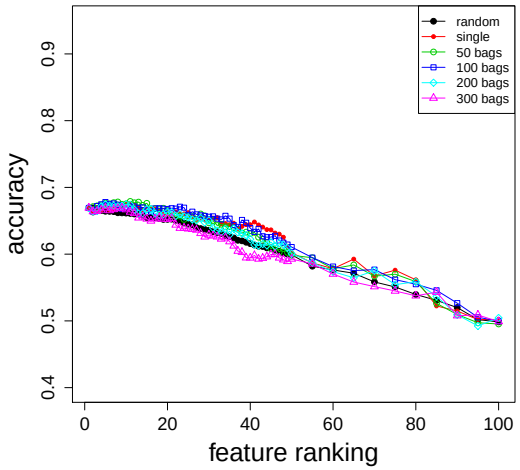
(a) FFA curves of the "single" data.



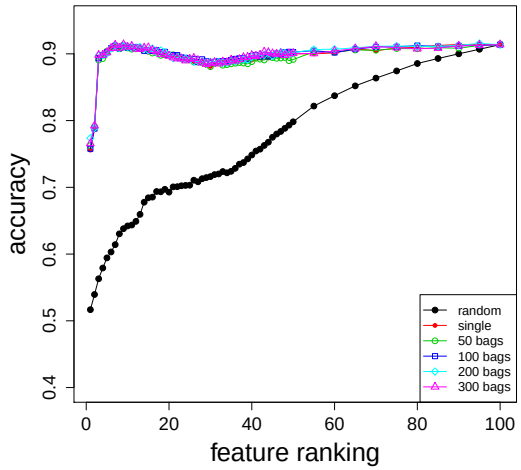
(b) RFA curves of the "single" data.



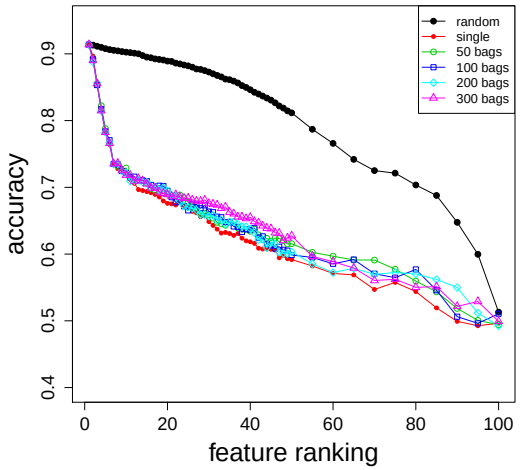
(c) FFA curves of the "pair" data.



(d) RFA curves of the "pair" data.

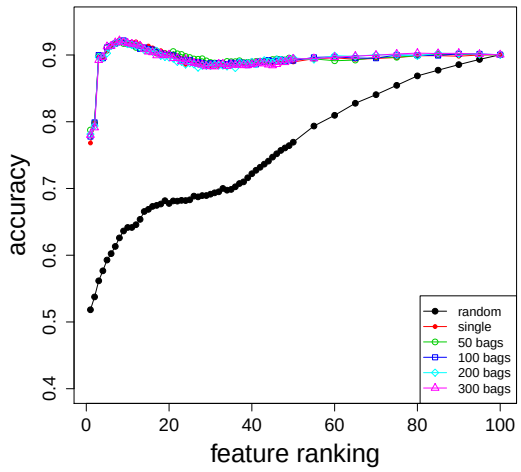


(e) FFA curves of the "combined" data.

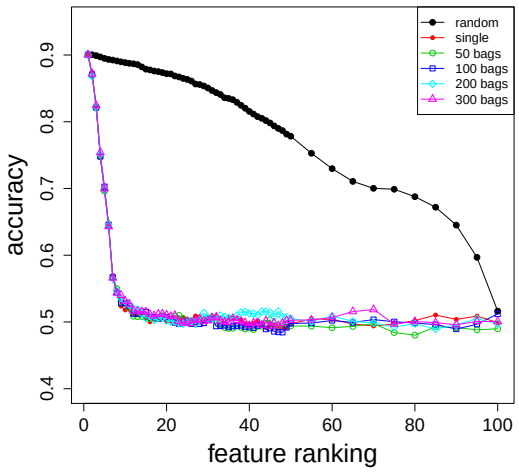


(f) RFA curves of the "combined" data.

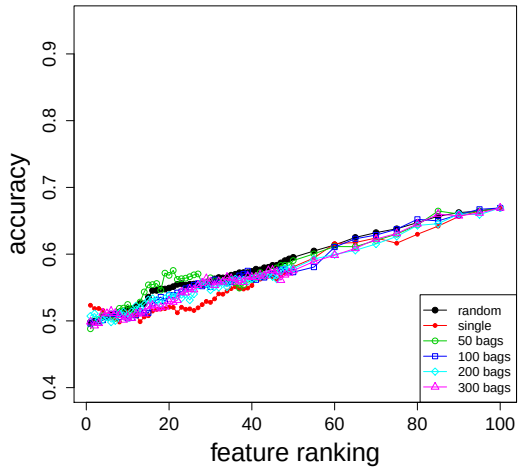
Figure 46: Comparison of FFA (left) and RFA (right) curves of **single** and **ensemble** rankings produced by considering a **different number** k of base rankings. The base rankings are produced by **SVM-RFE** and aggregated with the **min** function.



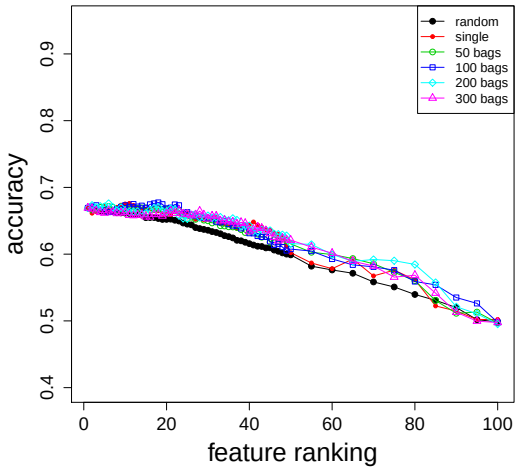
(a) FFA curves of the "single" data.



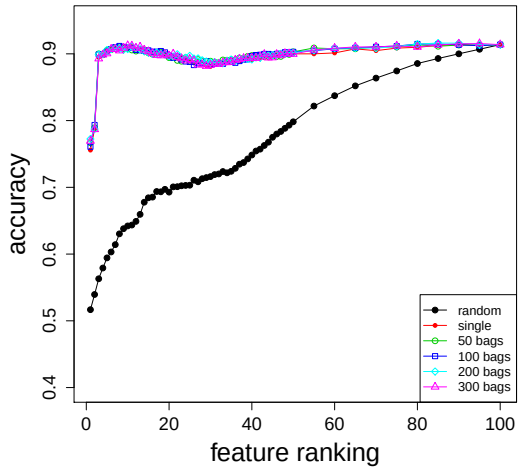
(b) RFA curves of the "single" data.



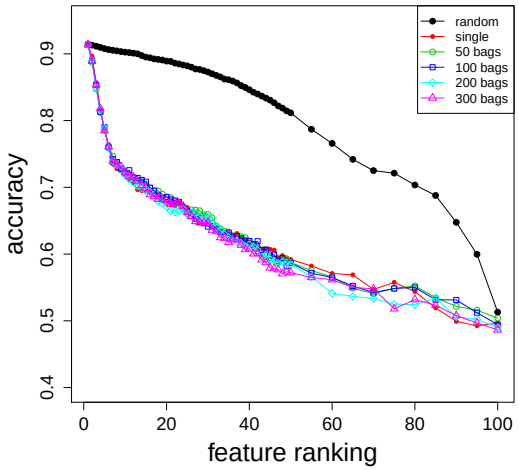
(c) FFA curves of the "pair" data.



(d) RFA curves of the "pair" data.



(e) FFA curves of the "combined" data.

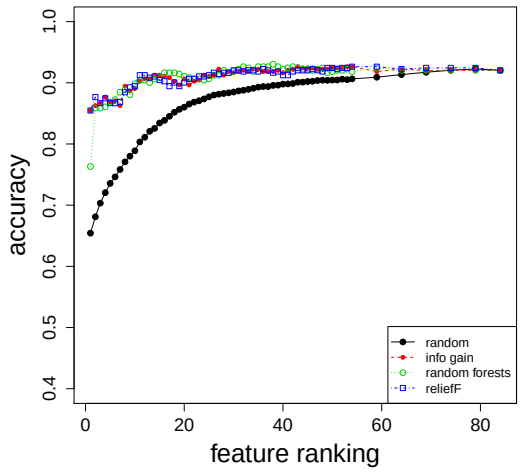


(f) RFA curves of the "combined" data.

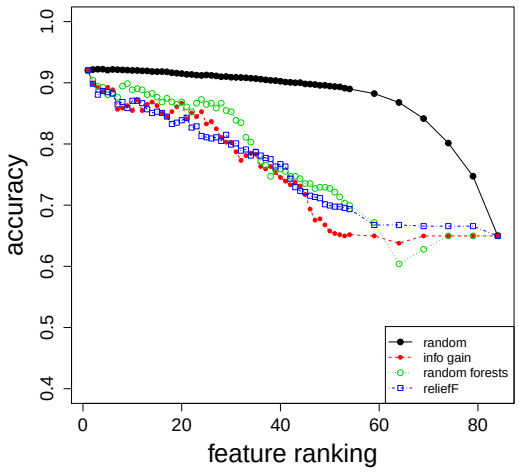
Figure 47: Comparison of FFA (left) and RFA (right) curves of **single** and **ensemble** rankings produced by considering a **different number** k of base rankings. The base rankings are produced by **SVM-RFE** and aggregated with the **max** function.

B.2 Experiments in Different Domains

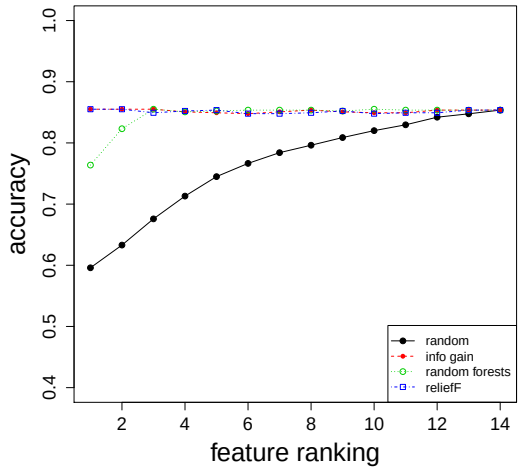
We analyse 23 classification datasets originating from various real-life domains. The purpose of the experiments is to examine the quality of the feature rankings produced by several feature ranking methods and to draw general conclusions about their performance across domains. We compare the FFA and RFA curves of four feature ranking algorithms (Info Gain, Random Forests, ReliefF and SVM-RFE) and as baseline we use the expected error curve of a random ranking.



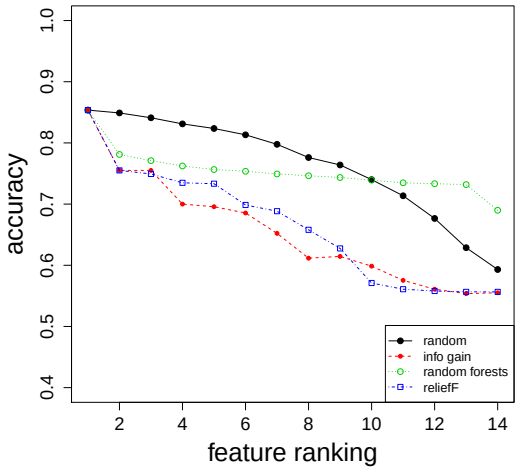
(a) FFA curves of the "aapc" data.



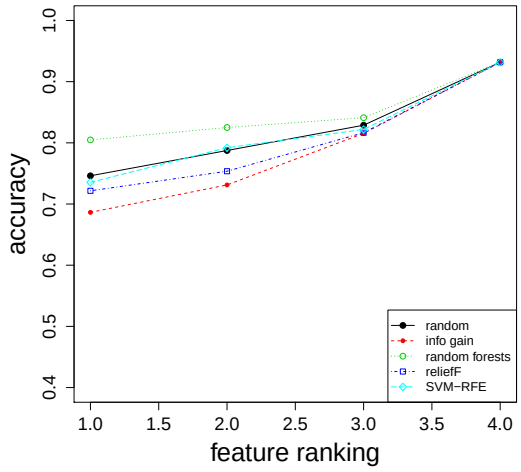
(b) RFA curves of the "aapc" data.



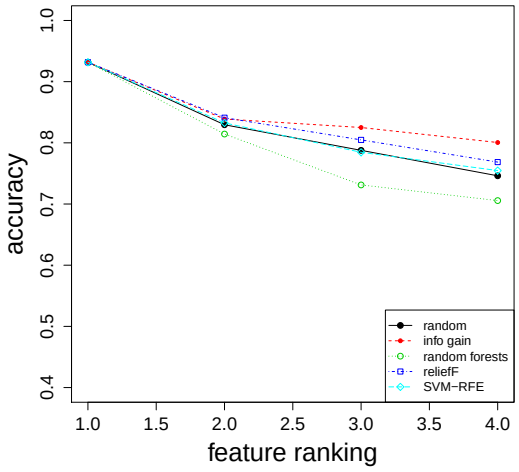
(c) FFA curves of the "australian" data.



(d) RFA curves of the "australian" data.

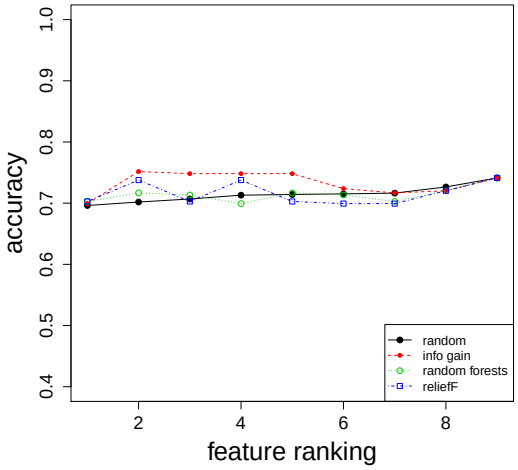


(e) FFA curves of the "balance" data.

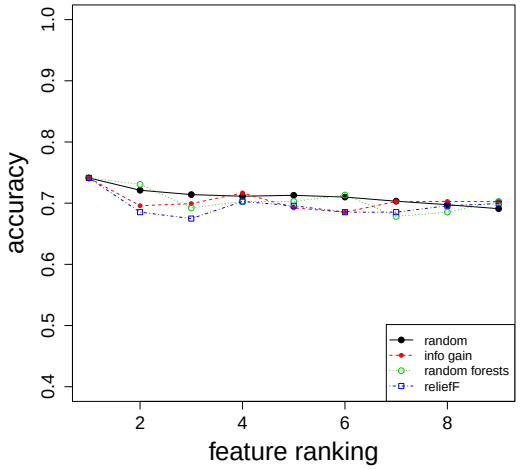


(f) RFA curves of the "balance" data.

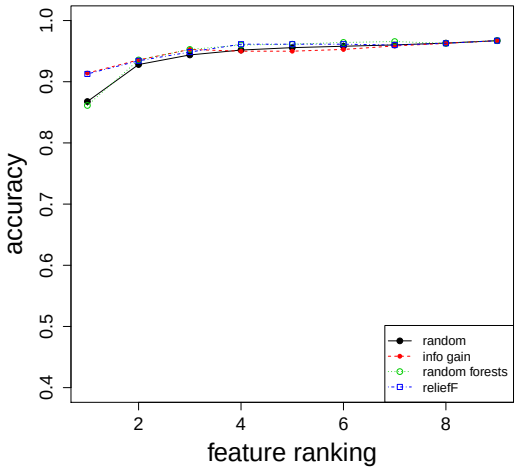
Figure 48: Comparison of the FFA (left) and RFA (right) of different feature ranking algorithms and a random ranking. The feature rankings are obtained for various benchmark datasets from different domains.



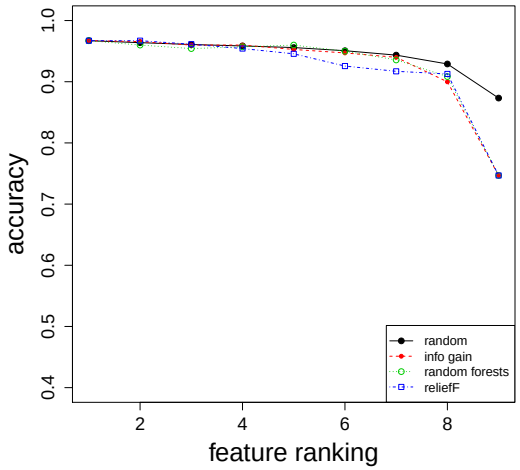
(a) FFA curves of the "breast-cancer" data.



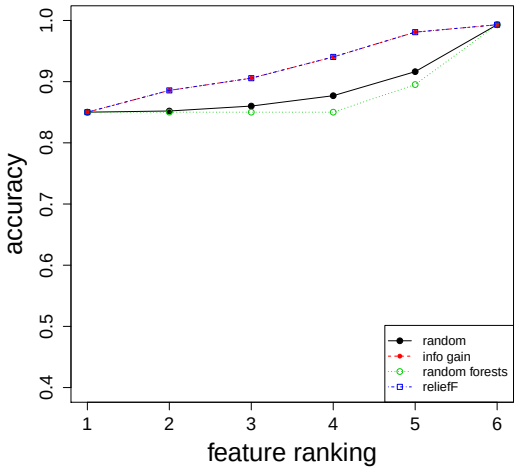
(b) RFA curves of the "breast-cancer" data.



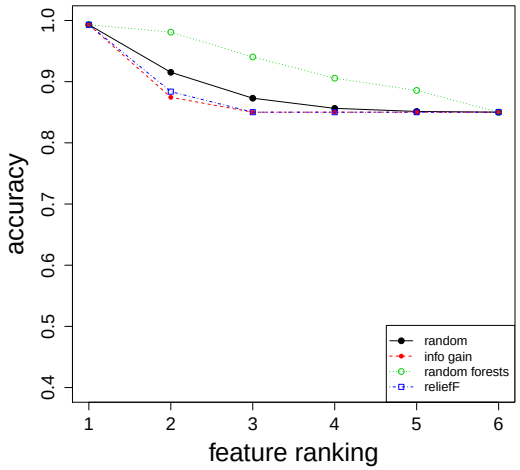
(c) FFA curves of the "breast-w" data.



(d) RFA curves of the "breast-w" data.

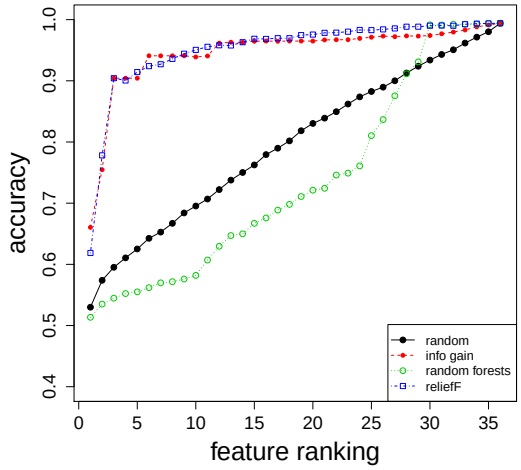


(e) FFA curves of the "car" data.

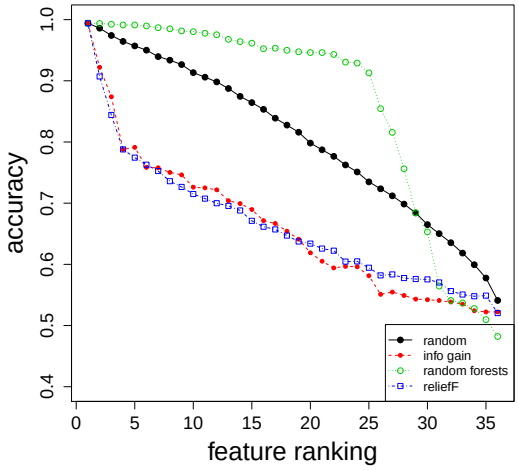


(f) RFA curves of the "car" data.

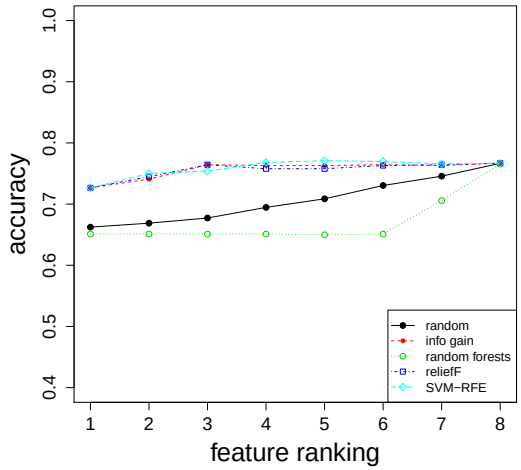
Figure 49: Comparison of the FFA (left) and RFA (right) of different feature ranking algorithms and a random ranking. The feature rankings are obtained for various benchmark datasets from different domains.



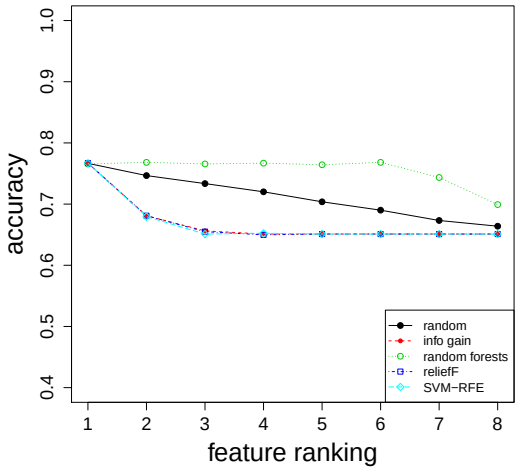
(a) FFA curves of the "chess" data.



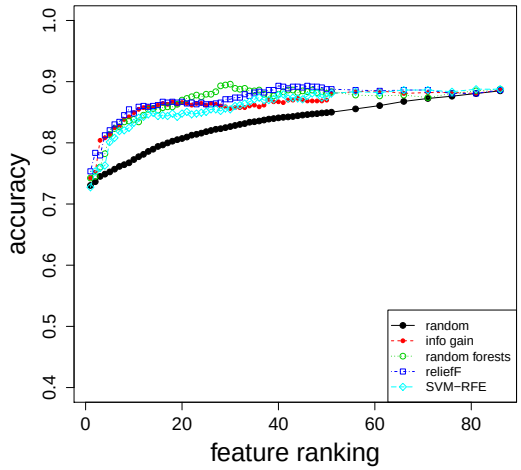
(b) RFA curves of the "chess" data.



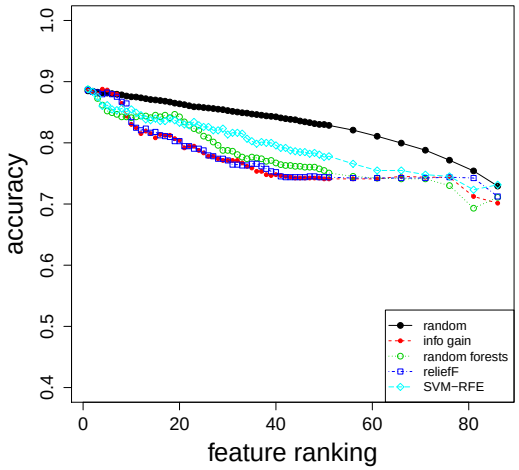
(c) FFA curves of the "diabetes" data.



(d) RFA curves of the "diabetes" data.

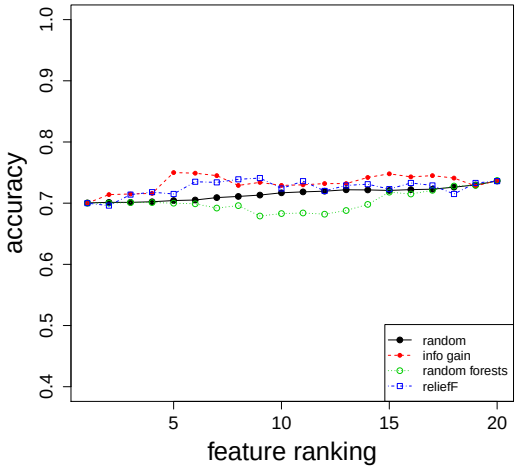


(e) FFA curves of the "diversity" data.

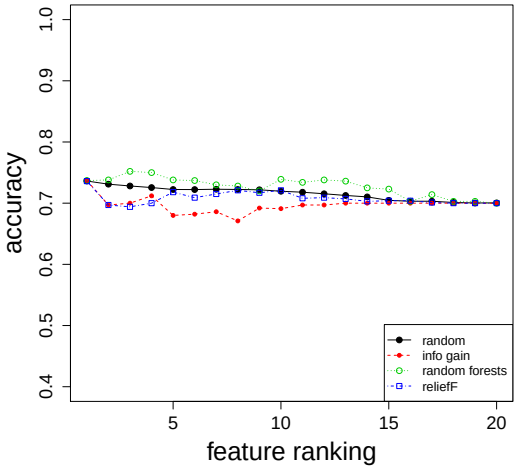


(f) RFA curves of the "diversity" data.

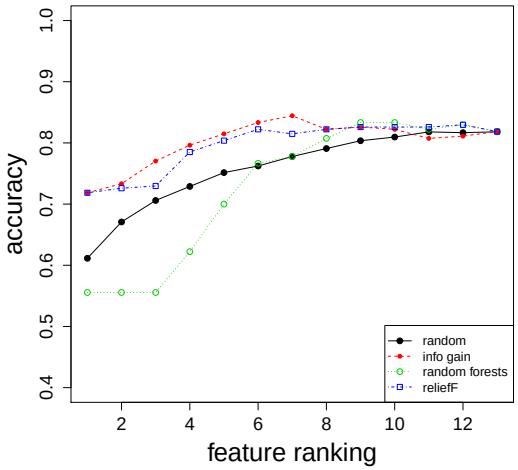
Figure 50: Comparison of the FFA (left) and RFA (right) of different feature ranking algorithms and a random ranking. The feature rankings are obtained for various benchmark datasets from different domains.



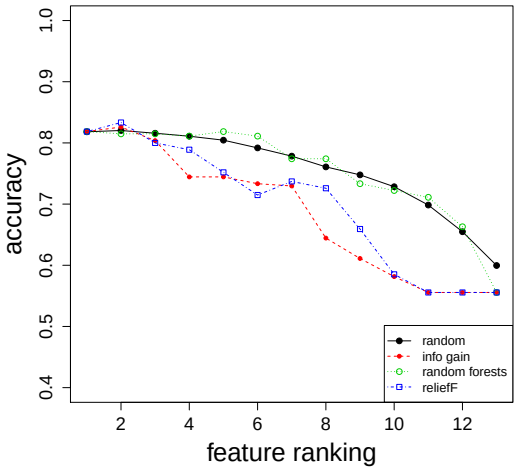
(a) FFA curves of the "german" data.



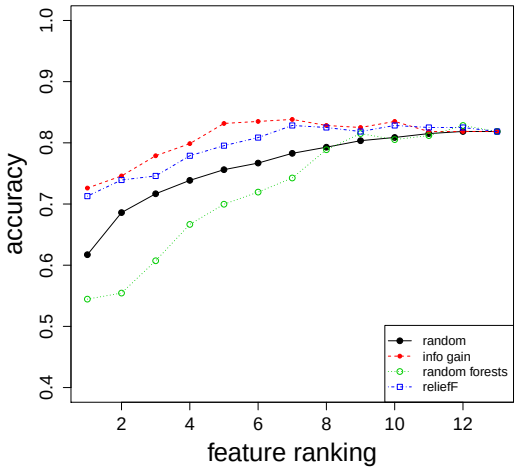
(b) RFA curves of the "german" data.



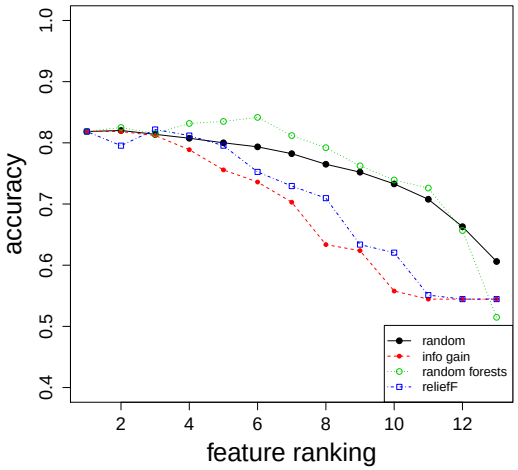
(c) FFA curves of the "heart" data.



(d) RFA curves of the "heart" data.

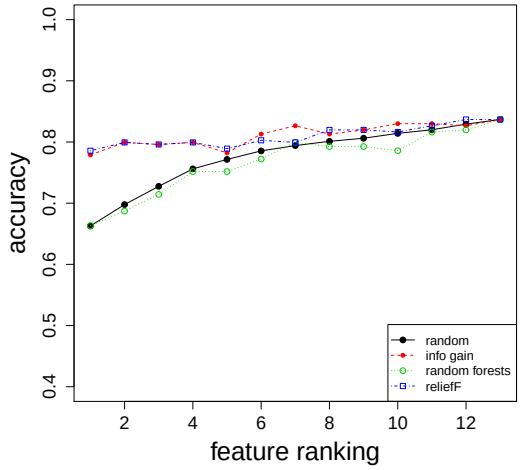


(e) FFA curves of the "heart-c" data.

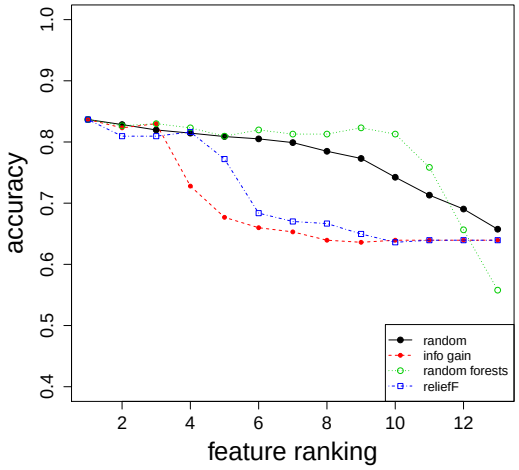


(f) RFA curves of the "heart-c" data.

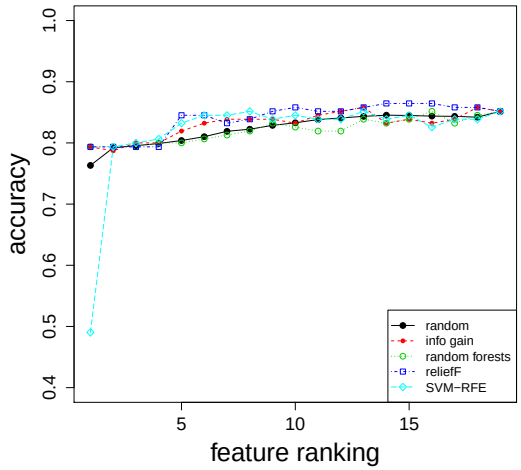
Figure 51: Comparison of the FFA (left) and RFA (right) of different feature ranking algorithms and a random ranking. The feature rankings are obtained for various benchmark datasets from different domains.



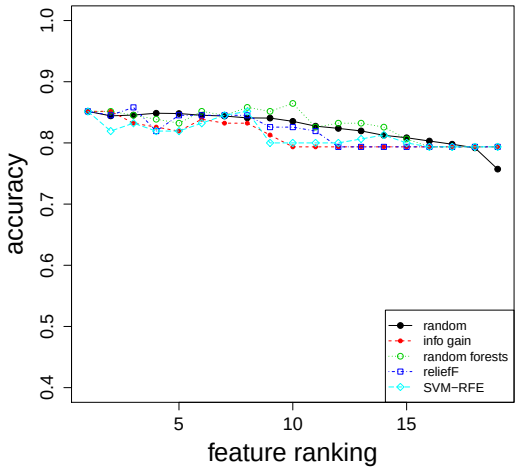
(a) FFA curves of the "heart-h" data.



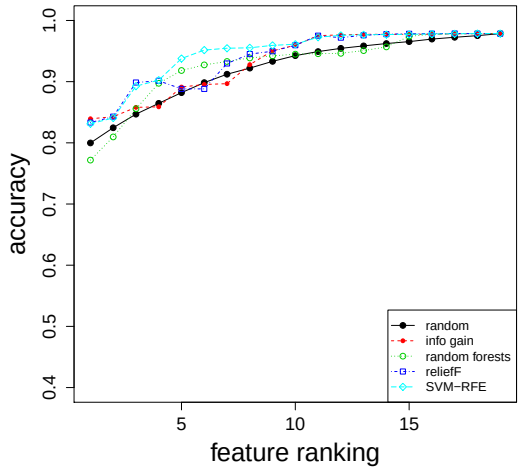
(b) RFA curves of the "heart-h" data.



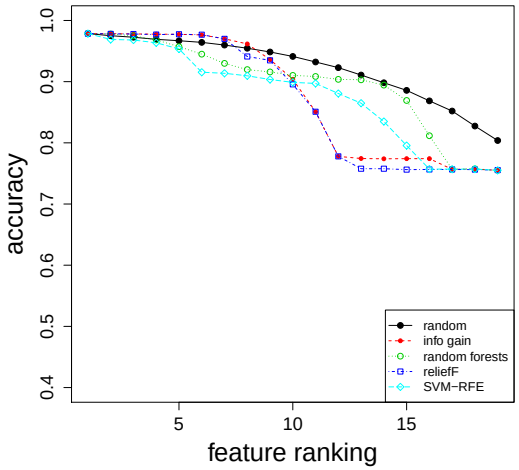
(c) FFA curves of the "hepatitis" data.



(d) RFA curves of the "hepatitis" data.

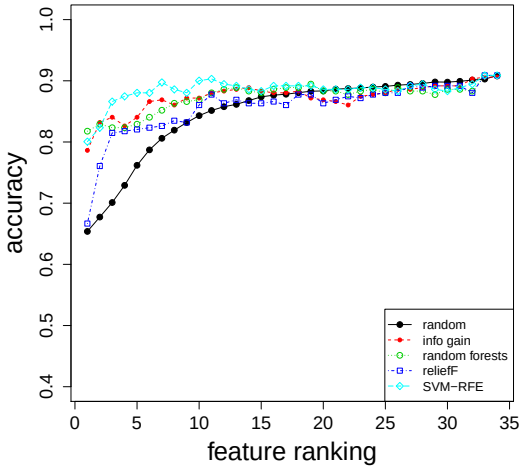


(e) FFA curves of the "image" data.

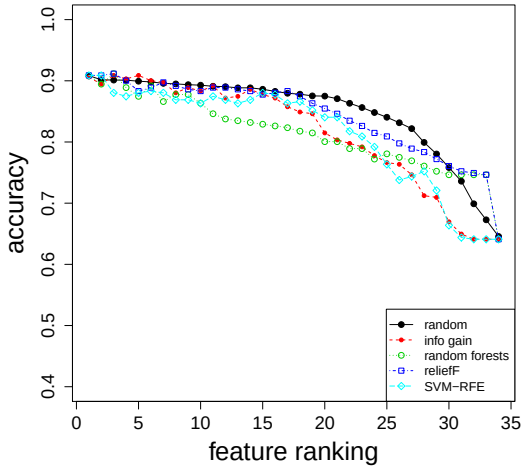


(f) RFA curves of the "image" data.

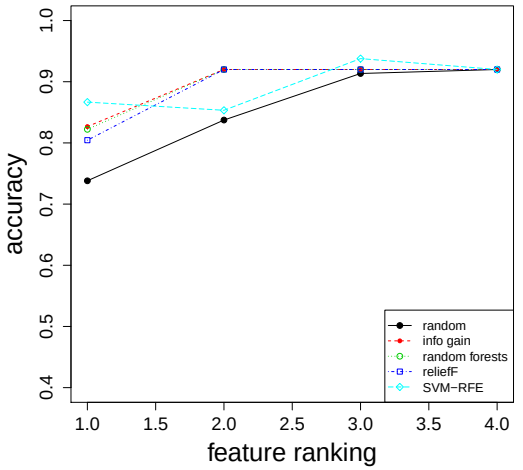
Figure 52: Comparison of the FFA (left) and RFA (right) of different feature ranking algorithms and a random ranking. The feature rankings are obtained for various benchmark datasets from different domains.



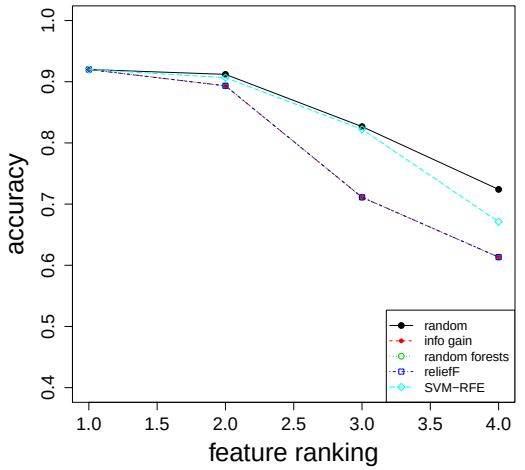
(a) FFA curves of the "ionosphere" data.



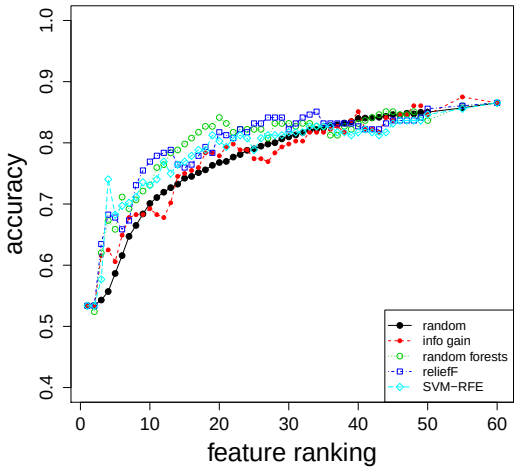
(b) RFA curves of the "ionosphere" data.



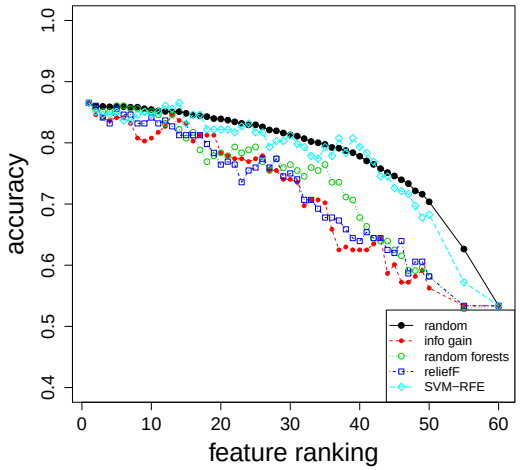
(c) FFA curves of the "iris" data.



(d) RFA curves of the "iris" data.

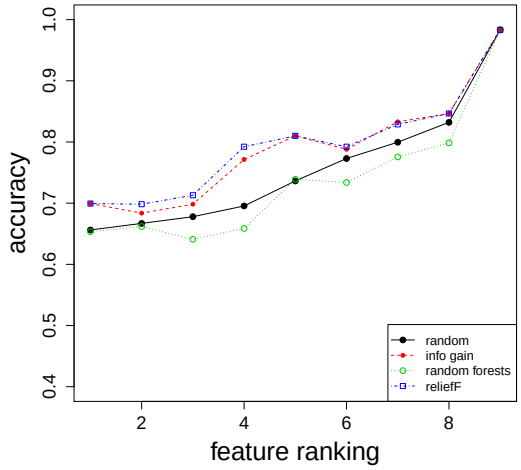


(e) FFA curves of the "sonar" data.

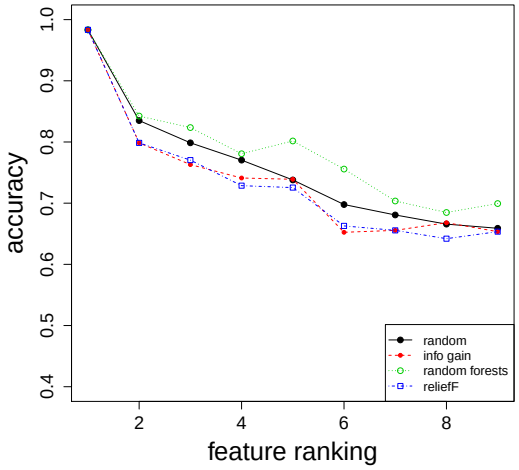


(f) RFA curves of the "sonar" data.

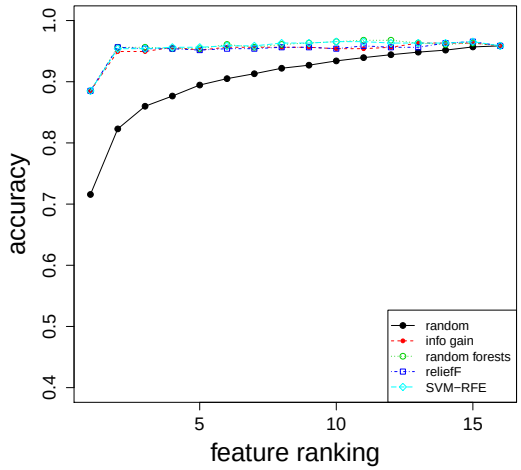
Figure 53: Comparison of the FFA (left) and RFA (right) of different feature ranking algorithms and a random ranking. The feature rankings are obtained for various benchmark datasets from different domains.



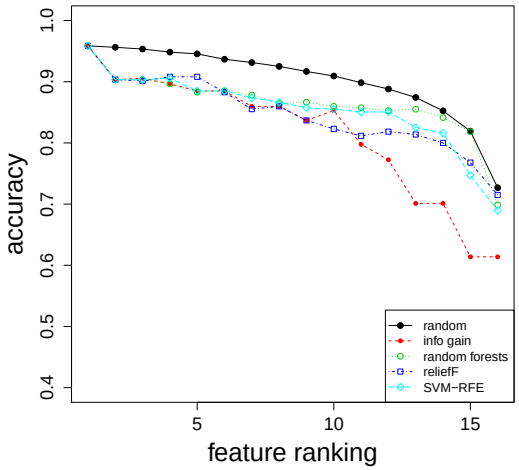
(a) FFA curves of the "tic-tac-toe" data.



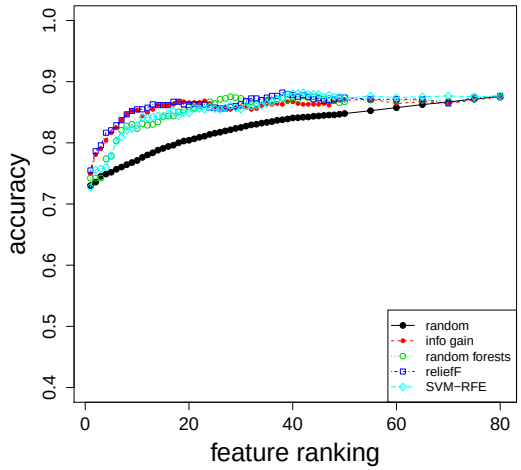
(b) RFA curves of the "tic-tac-toe" data.



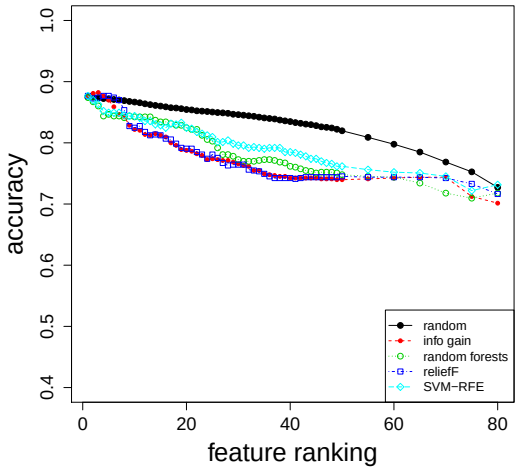
(c) FFA curves of the "vote" data.



(d) RFA curves of the "vote" data.

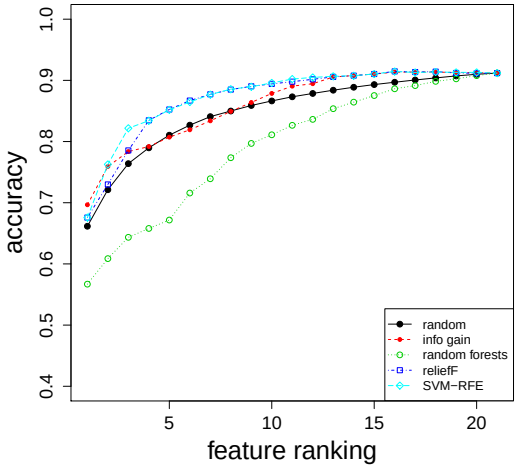


(e) FFA curves of the "water" data.

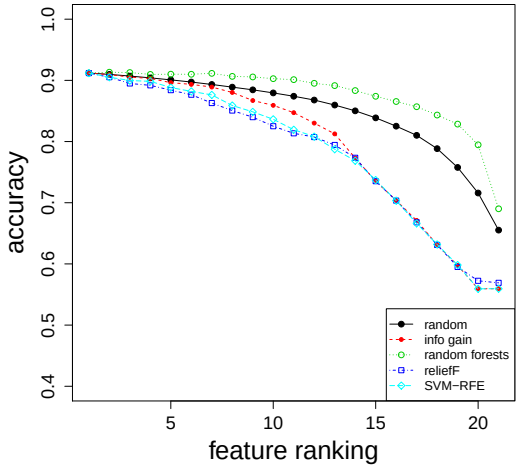


(f) RFA curves of the "water" data.

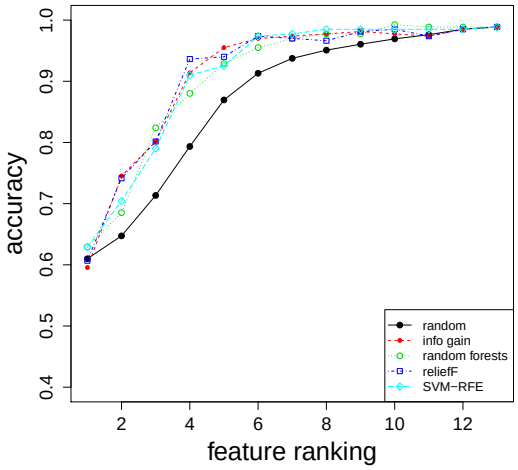
Figure 54: Comparison of the FFA (left) and RFA (right) of different feature ranking algorithms and a random ranking. The feature rankings are obtained for various benchmark datasets from different domains.



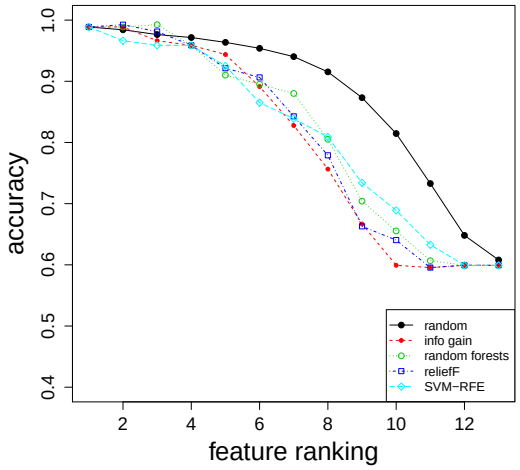
(a) FFA curves of the "waveform" data.



(b) RFA curves of the "waveform" data.



(c) FFA curves of the "wine" data.

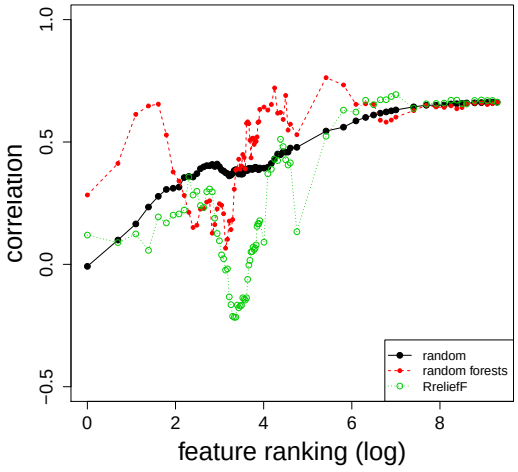


(d) RFA curves of the "wine" data.

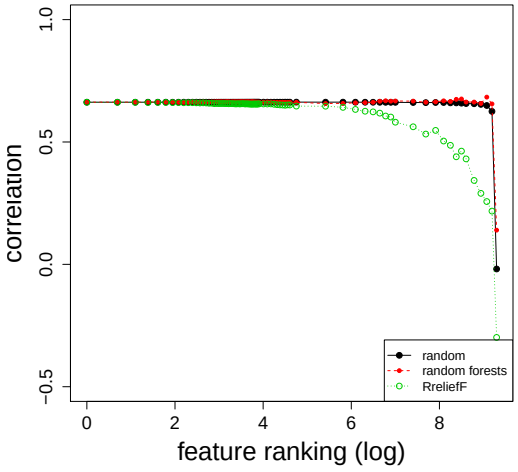
Figure 55: Comparison of the FFA (left) and RFA (right) of different feature ranking algorithms and a random ranking. The feature rankings are obtained for various benchmark datasets from different domains.

B.3 FFA and RFA curves of Individual ET Datasets

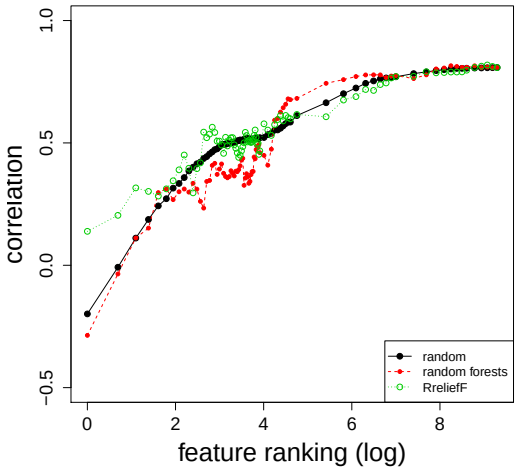
We present the FFA and RFA curves resulting from the individual analysis of a collection of 10 ET gene expression datasets. The purpose of the analysis is to discover key genes involved in the mechanism of tumour aggressiveness. Each figure contains a comparison of the FFA and RFA curves of the feature rankings produced by Random Forests and RReliefF. As a baseline for each dataset, the expected curve of random rankings is also included.



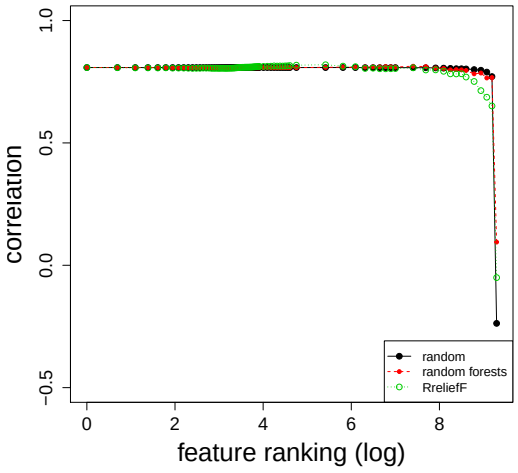
(a) FFA curves of the "ews12102" data.



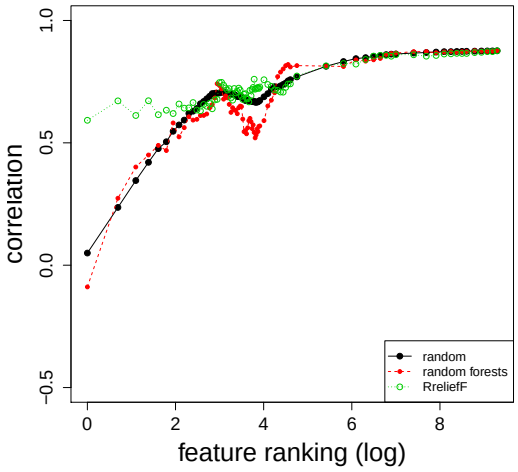
(b) RFA curves of the "ews12102" data.



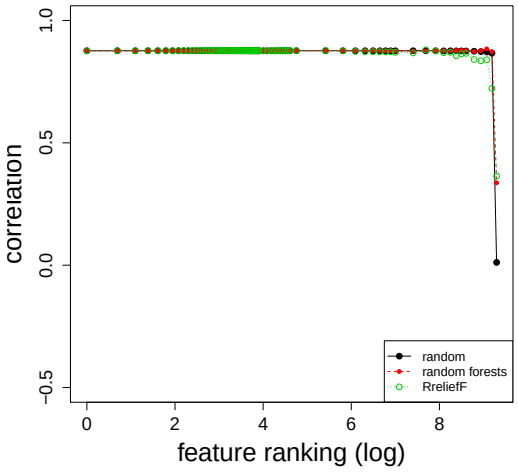
(c) FFA curves of the "mb10327" data.



(d) RFA curves of the "mb10327" data.

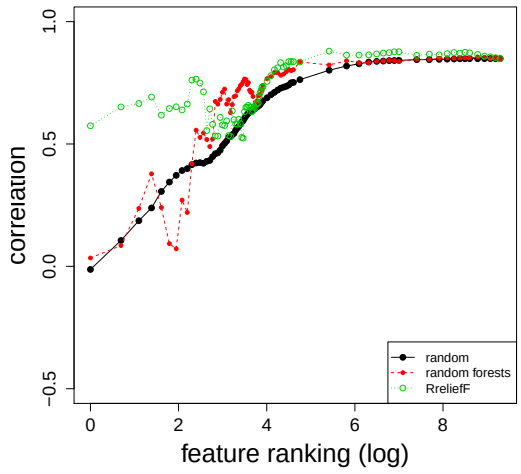


(e) FFA curves of the "mb12992" data.

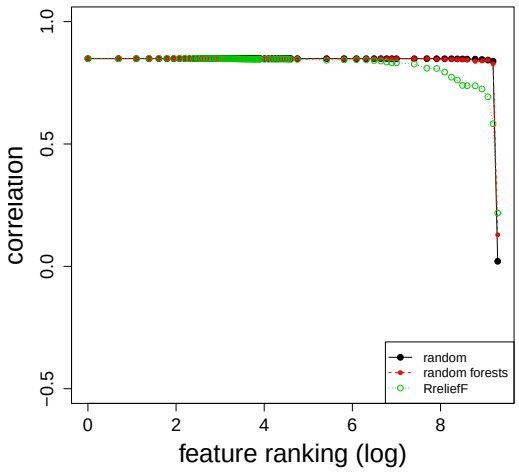


(f) RFA curves of the "mb12992" data.

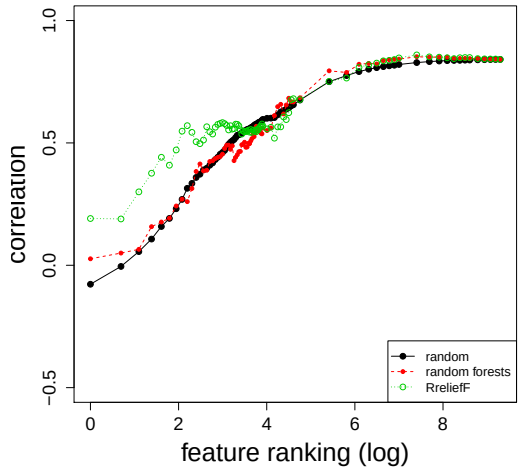
Figure 56: Comparison of the FFA (left) and RFA (right) of RReliefF, random forests and a random ranking. The feature rankings are obtained for different **embryonal tumor** (ET) datasets.



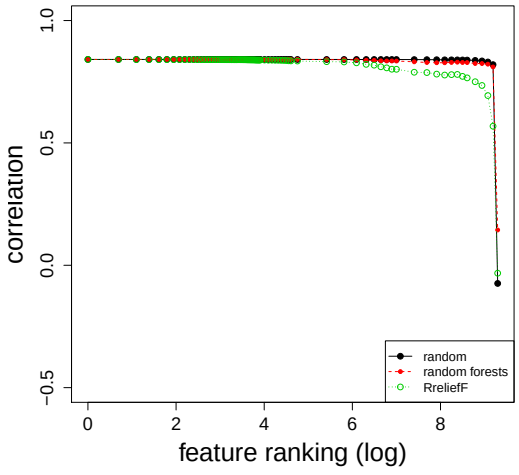
(a) FFA curves of the "mbDKFZ" data.



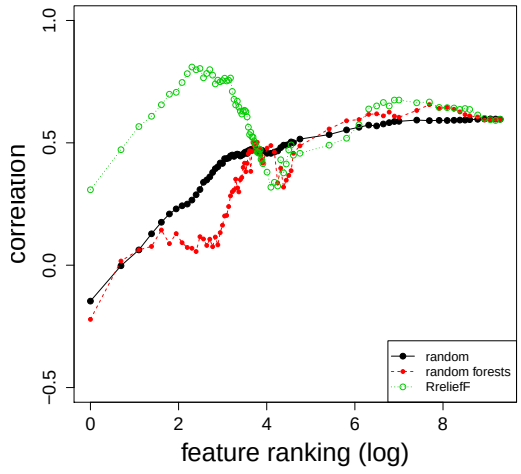
(b) RFA curves of the "mbDKFZ" data.



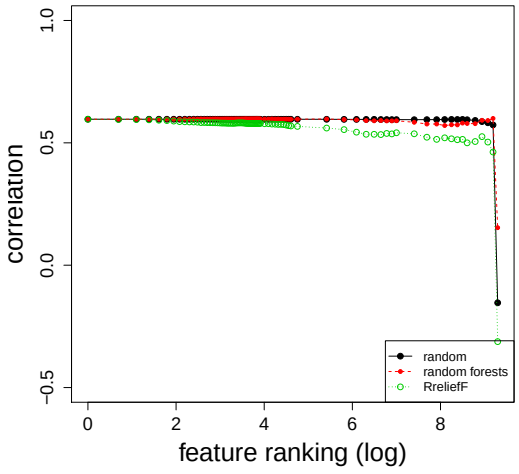
(c) FFA curves of the "nbCol251" data.



(d) RFA curves of the "nbCol251" data.

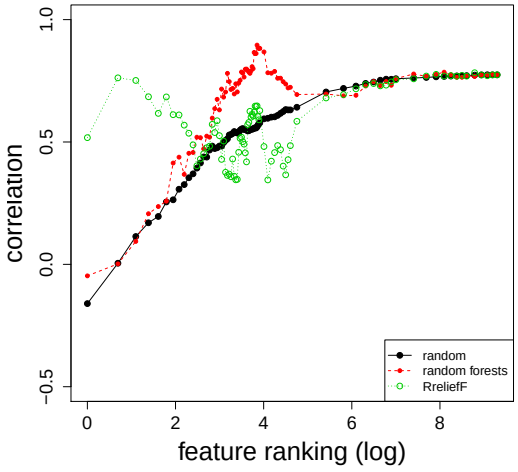


(e) FFA curves of the "nbEssen" data.

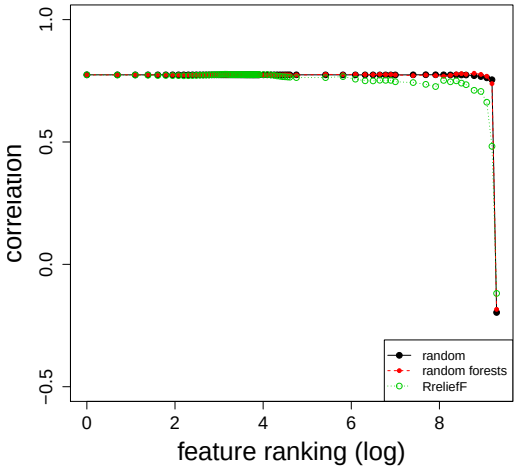


(f) RFA curves of the "nbEssen" data.

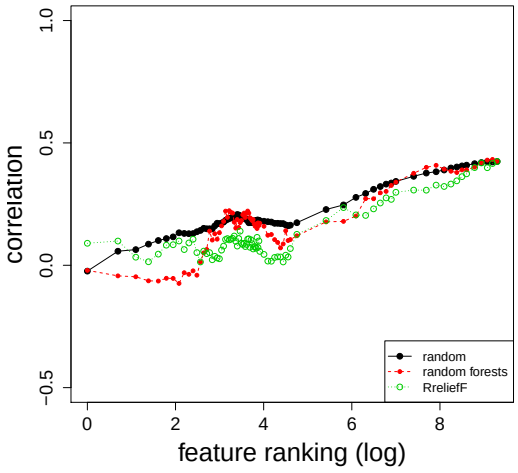
Figure 57: Comparison of the FFA (left) and RFA (right) of RReliefF, random forests and a random ranking. The feature rankings are obtained for different **embryonal tumor** (ET) datasets.



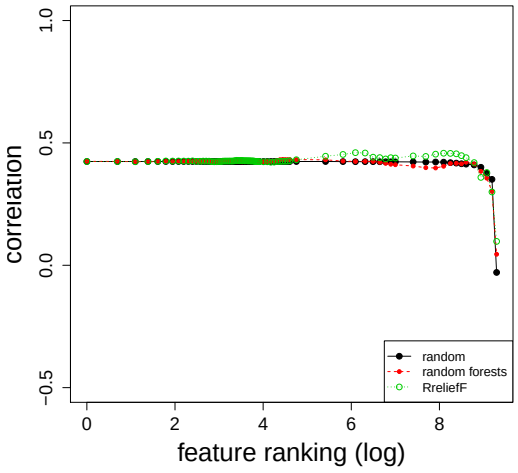
(a) FFA curves of the "rtCurie" data.



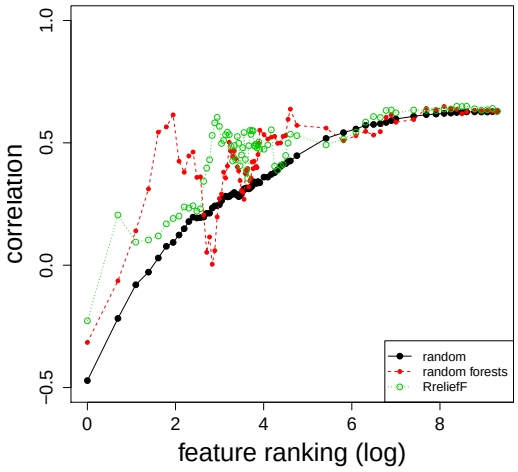
(b) RFA curves of the "rtCurie" data.



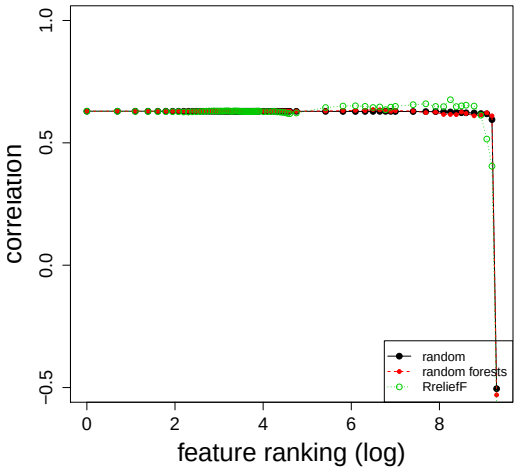
(c) FFA curves of the "wt10320" data.



(d) RFA curves of the "wt10320" data.

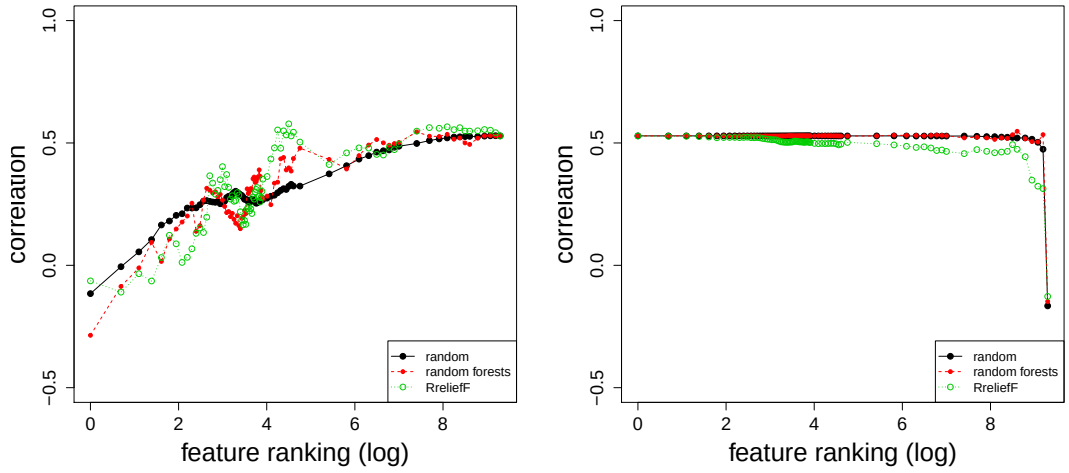


(e) FFA curves of the "wt11024" data.



(f) RFA curves of the "wt11024" data.

Figure 58: Comparison of the FFA (left) and RFA (right) of RReliefF, random forests and a random ranking. The feature rankings are obtained for different **embryonal tumor** (ET) datasets.



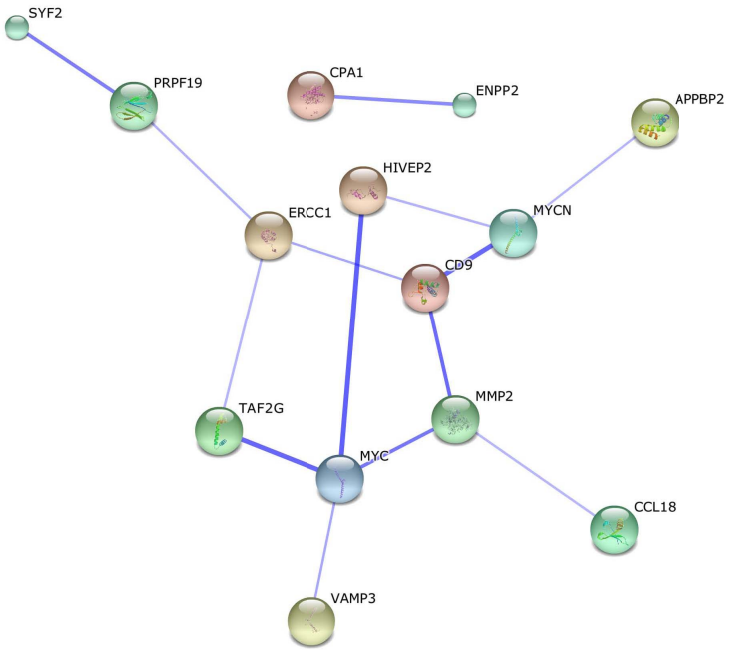
(a) FFA curves of the "wtETABM53" data.

(b) RFA curves of the "wtETABM53" data.

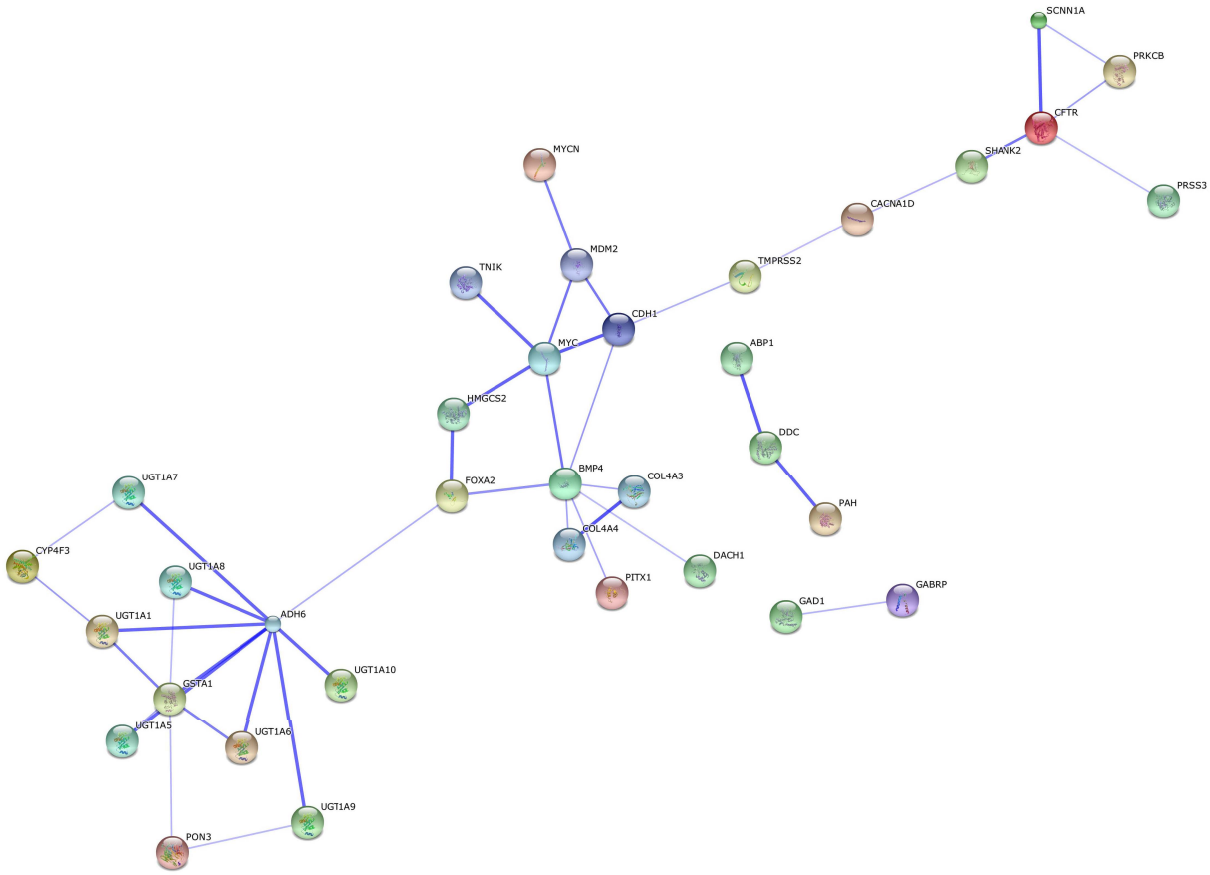
Figure 59: Comparison of the FFA (left) and RFA (right) of RReliefF, random forests and a random ranking. The feature rankings are obtained for different **embryonal tumor** (ET) datasets.

B.4 Gene Networks of Individual ET Datasets

We present the full set of gene networks constructed with the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING). We compare the networks constructed from the top-50 genes of each individual dataset, provided from the feature ranking of Random Forests and ReliefF.

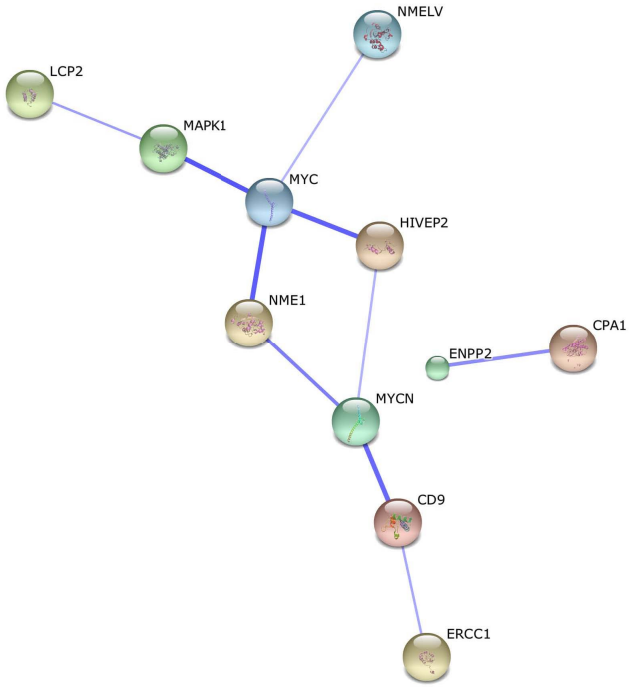


(a) Gene network constructed from the top-50 genes provided by RFs

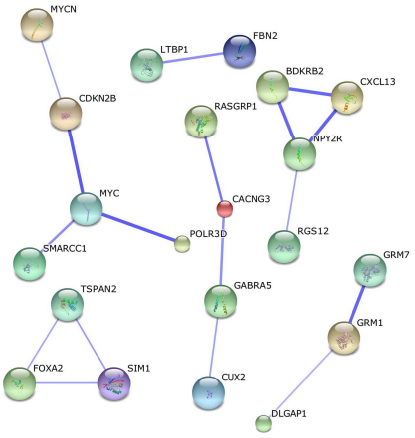


(b) Gene network constructed from the top-50 genes provided by ReliefF

Figure 60: Gene networks constructed from the top-50 genes of different feature ranking algorithms. The ranked gene list is induced for the “ews12102” dataset.

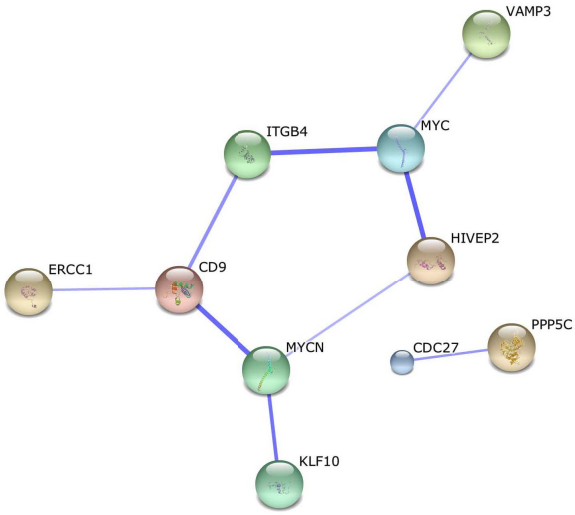


(a) Gene network constructed from the top-50 genes provided by RFs

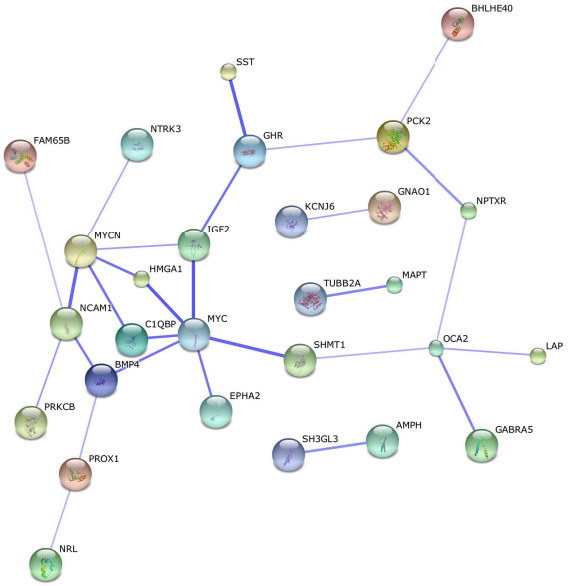


(b) Gene network constructed from the top-50 genes provided by ReliefF

Figure 61: Gene networks constructed from the top-50 genes of different feature ranking algorithms. The ranked gene list is induced for the "mb10327" dataset.

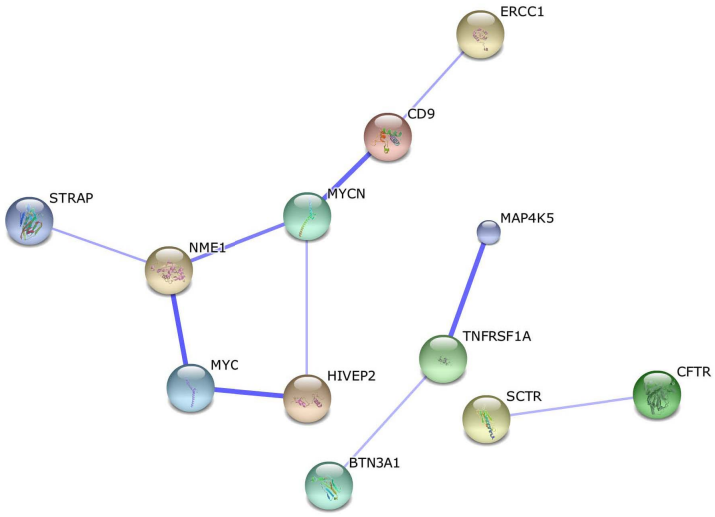


(a) Gene network constructed from the top-50 genes provided by RFs

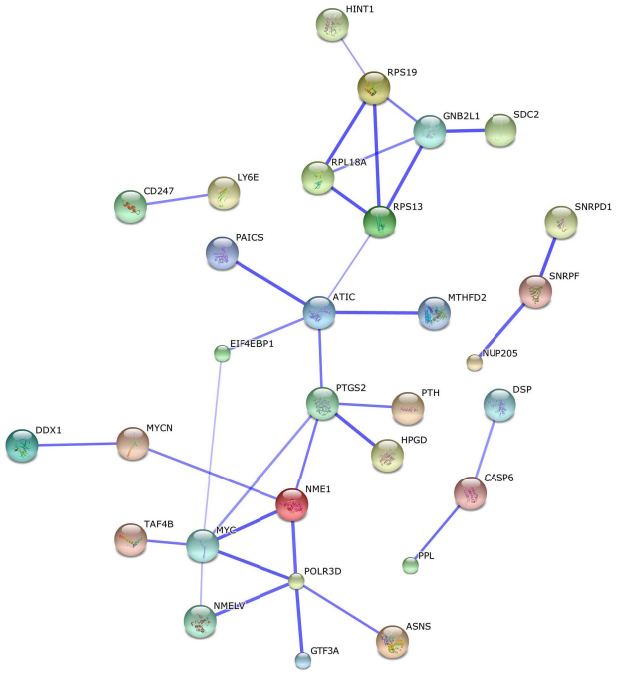


(b) Gene network constructed from the top-50 genes provided by ReliefF

Figure 62: Gene networks constructed from the top-50 genes of different feature ranking algorithms. The ranked gene list is induced for the “mb12992” dataset.

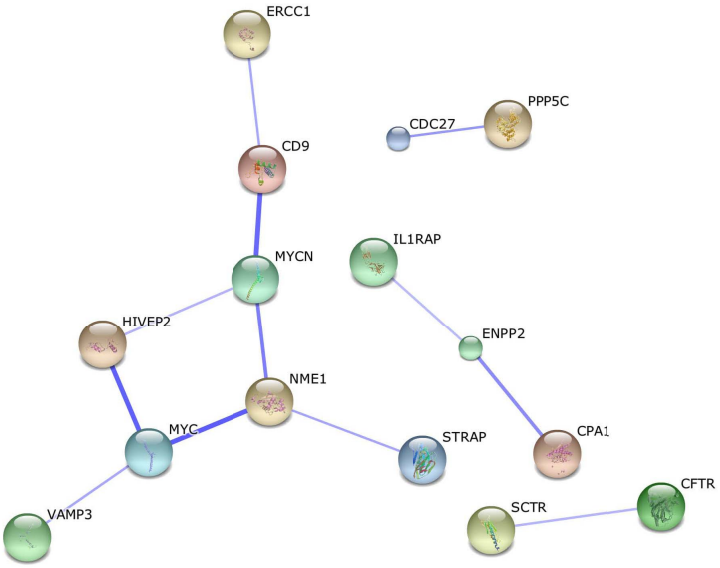


(a) Gene network constructed from the top-50 genes provided by RFs

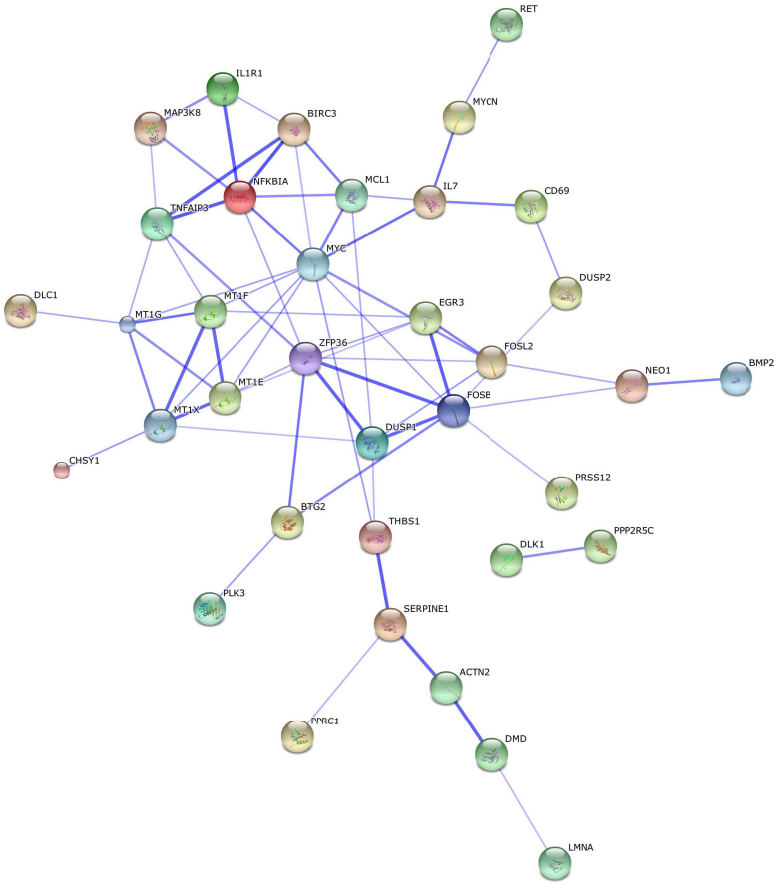


(b) Gene network constructed from the top-50 genes provided by ReliefF

Figure 63: Gene networks constructed from the top-50 genes of different feature ranking algorithms. The ranked gene list is induced for the “mbDKFZ” dataset.

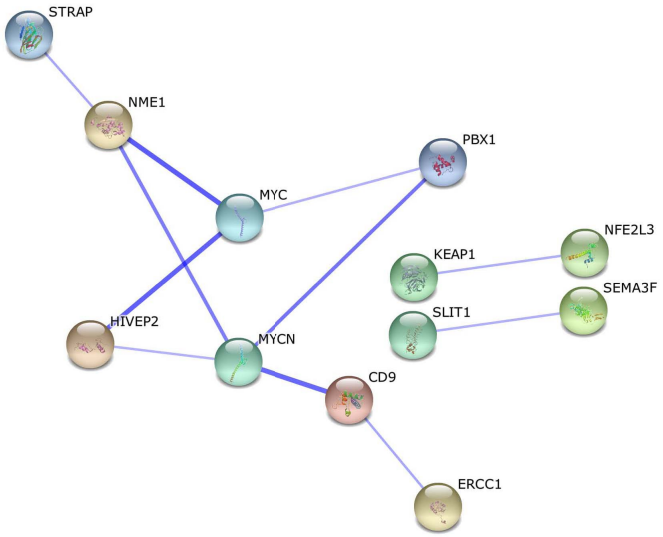


(a) Gene network constructed from the top-50 genes provided by RFs

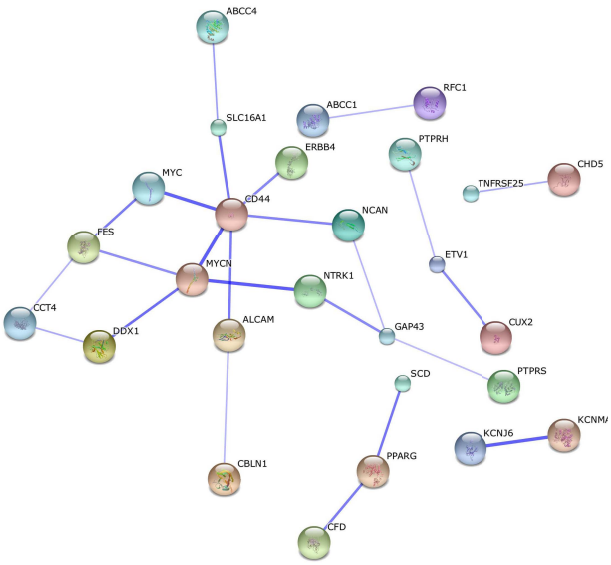


(b) Gene network constructed from the top-50 genes provided by ReliefF

Figure 64: Gene networks constructed from the top-50 genes of different feature ranking algorithms. The ranked gene list is induced for the "nbCol251" dataset.

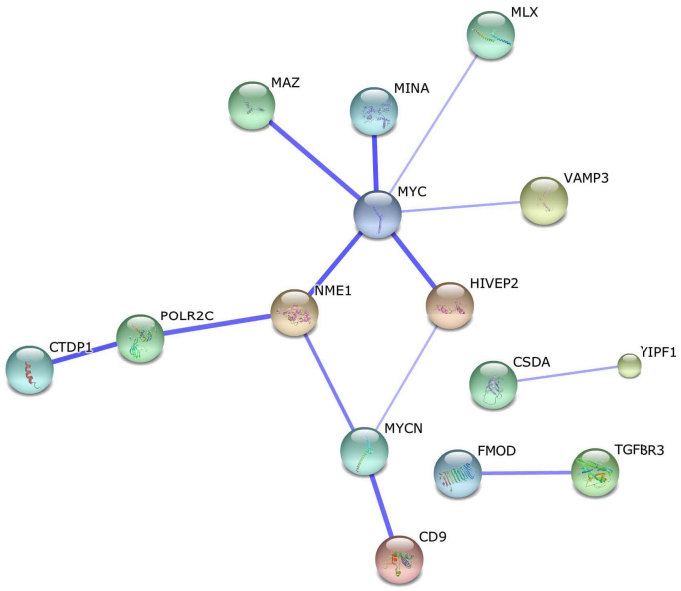


(a) Gene network constructed from the top-50 genes provided by RFs

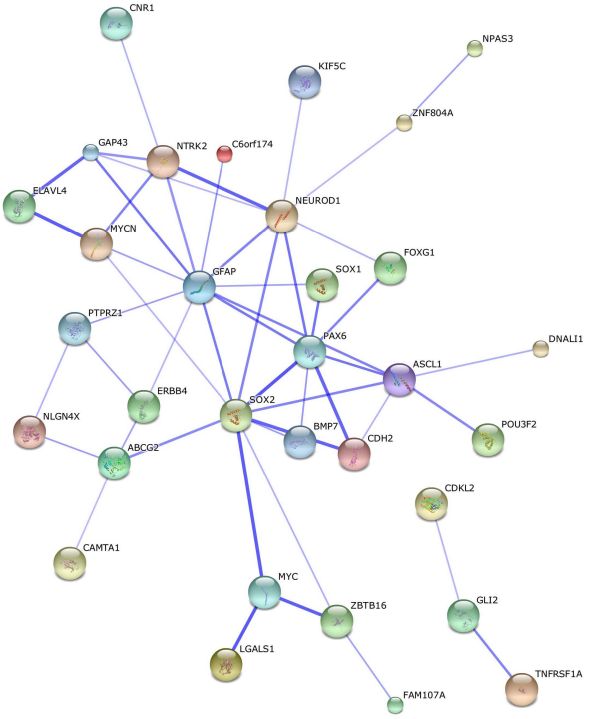


(b) Gene network constructed from the top-50 genes provided by ReliefF

Figure 65: Gene networks constructed from the top-50 genes of different feature ranking algorithms. The ranked gene list is induced for the “nbEssen” dataset.

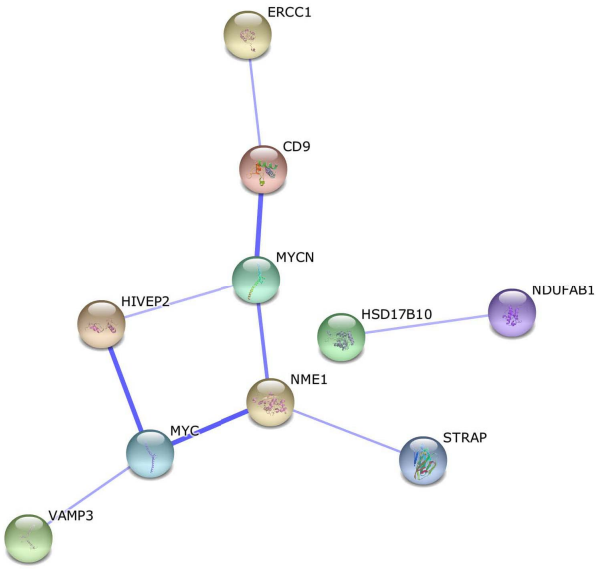


(a) Gene network constructed from the top-50 genes provided by RFs

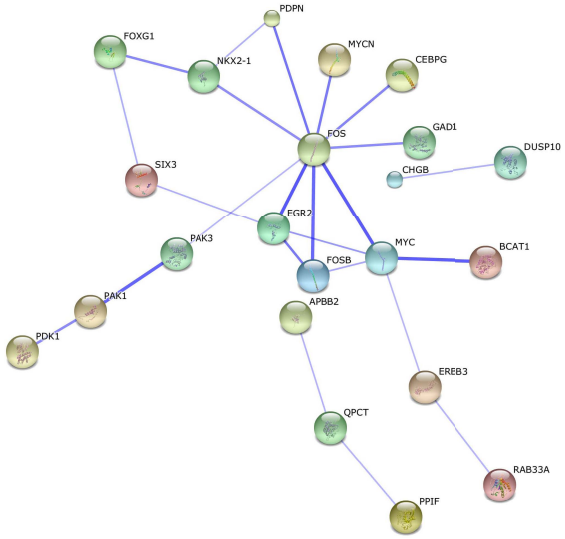


(b) Gene network constructed from the top-50 genes provided by ReliefF

Figure 66: Gene networks constructed from the top-50 genes of different feature ranking algorithms. The ranked gene list is induced for the "rtCurie" dataset.

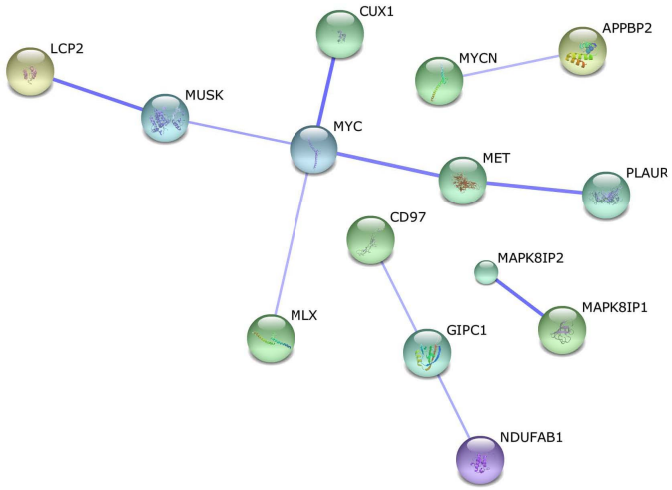


(a) Gene network constructed from the top-50 genes provided by RFs

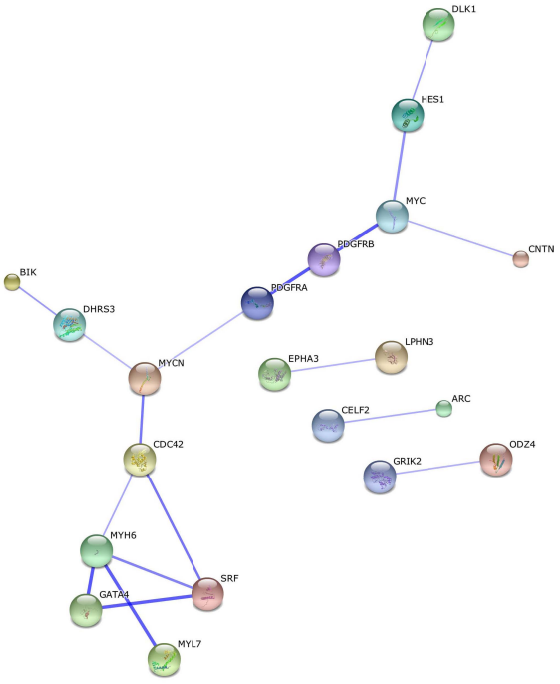


(b) Gene network constructed from the top-50 genes provided by ReliefF

Figure 67: Gene networks constructed from the top-50 genes of different feature ranking algorithms. The ranked gene list is induced for the “wt10320” dataset.

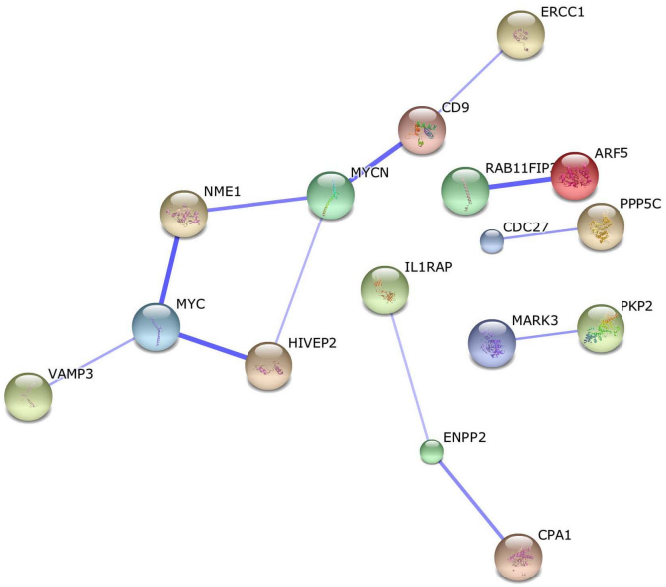


(a) Gene network constructed from the top-50 genes provided by RFs

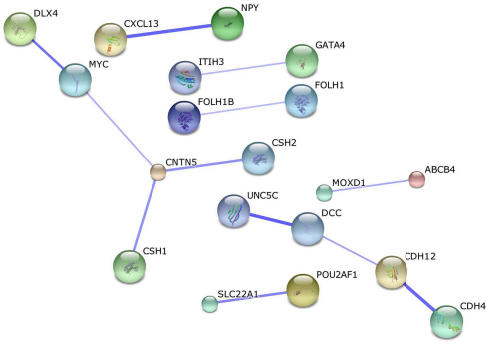


(b) Gene network constructed from the top-50 genes provided by ReliefF

Figure 68: Gene networks constructed from the top-50 genes of different feature ranking algorithms. The ranked gene list is induced for the "wt11024" dataset.



(a) Gene network constructed from the top-50 genes provided by RFs

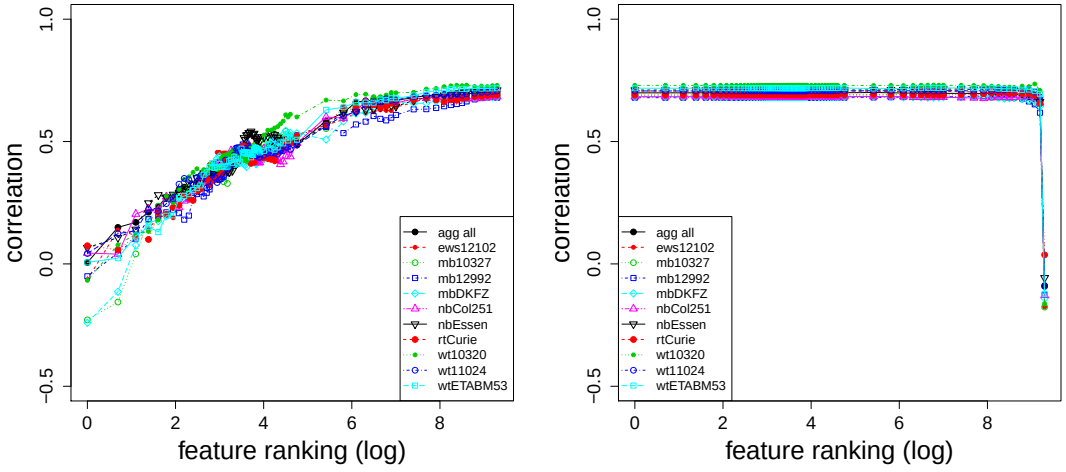


(b) Gene network constructed from the top-50 genes provided by ReliefF

Figure 69: Gene networks constructed from the top-50 genes of different feature ranking algorithms. The ranked gene list is induced for the "wtETABM53" dataset.

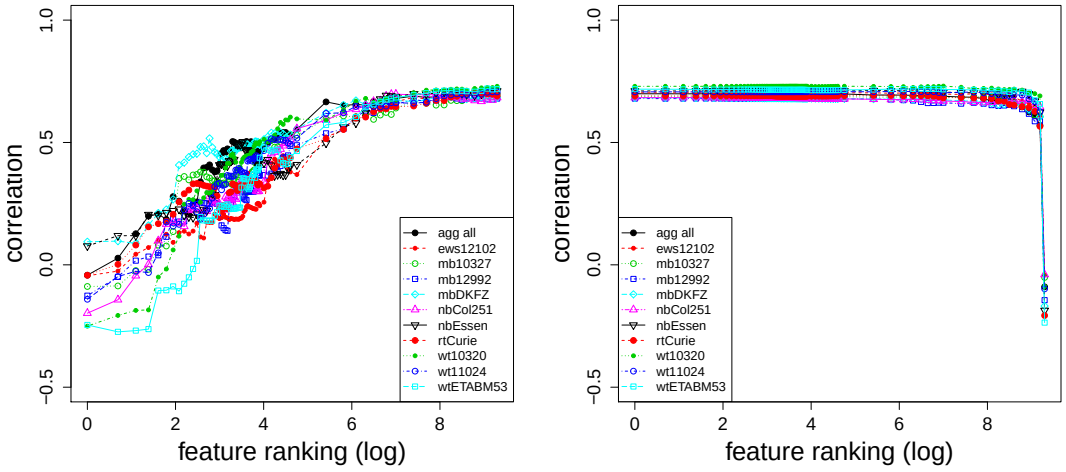
B.5 FFA and RFA curves of Aggregated ET Datasets

We present the FFA and RFA curves resulting from the aggregated analysis of a collection of 10 ET gene expression datasets. The purpose of the analysis is to discover a consesus set of genes for all ET entities, involved in the mechanism of tumour aggressiveness. Each figure contains a comparison of the FFA and RFA curves of the feature rankings produced from individual datasets, with the aggregated feature ranking produced for all of the ET dataset. We present the rankings produced by Random Forests and RReliefF. We also consider four aggregation functions for producing the consensus ranking: mean, median, min and max.



(a) FFA curves of random forests

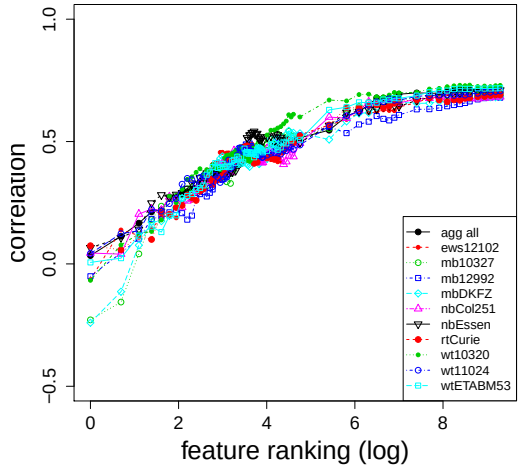
(b) RFA curves of random forests



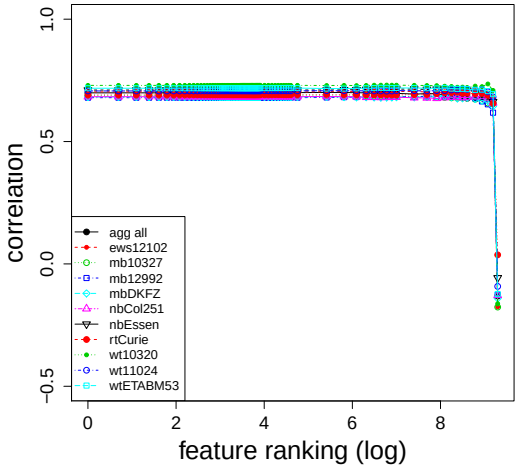
(c) FFA curves of reliefF

(d) RFA curves of reliefF

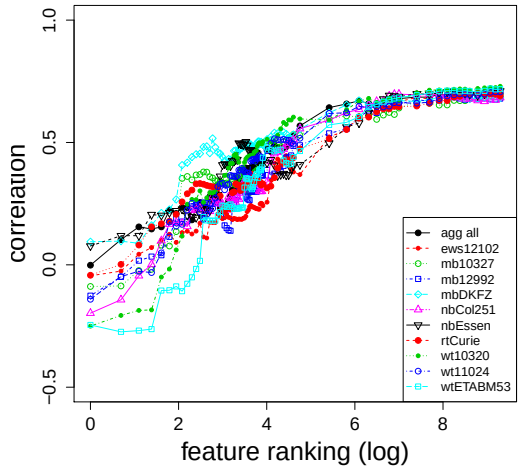
Figure 70: Comparison of the FFA (left) and RFA (right) of the aggregated feature ranking with the individual feature rankings for different **embryonal tumor** (ET) datasets. The aggregation is performed with the mean aggregation function



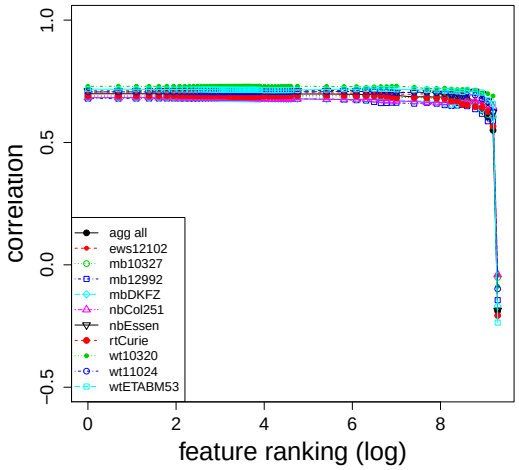
(a) FFA curves of random forests



(b) RFA curves of random forests

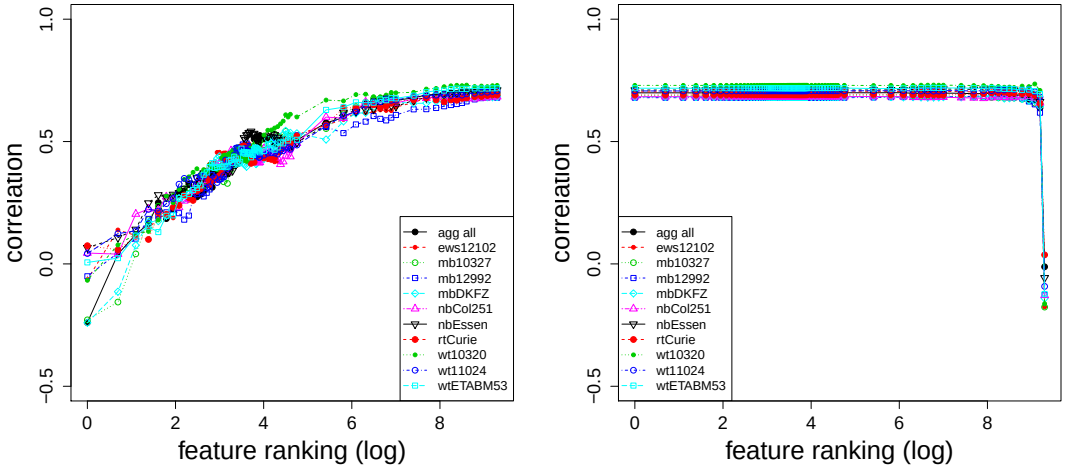


(c) FFA curves of reliefF



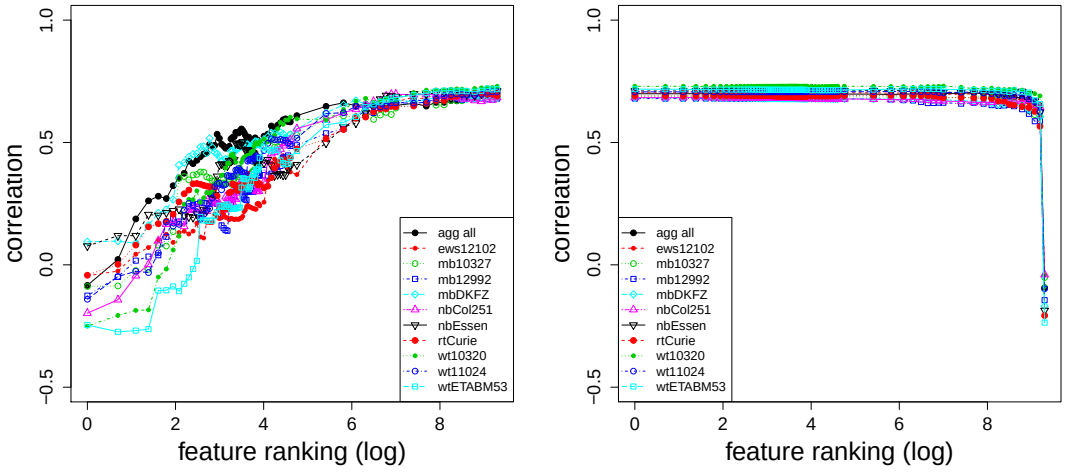
(d) RFA curves of reliefF

Figure 71: Comparison of the FFA (left) and RFA (right) of the aggregated feature ranking with the individual feature rankings for different **embryonal tumor** (ET) datasets. The aggregation is performed with the median aggregation function



(a) FFA curves of random forests

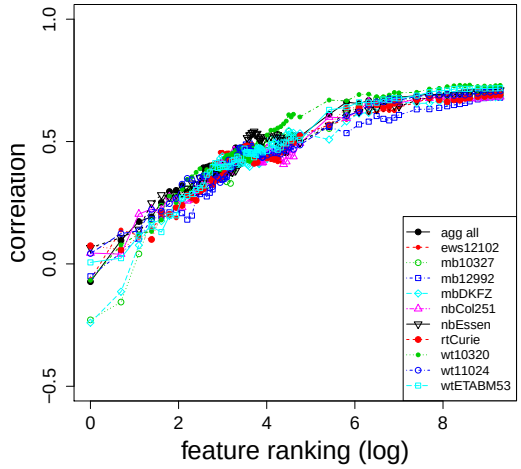
(b) RFA curves of random forests



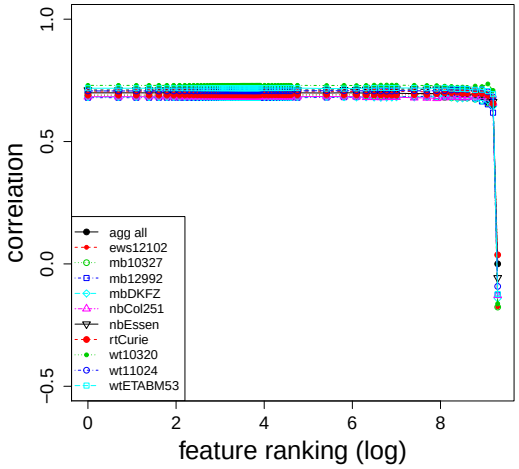
(c) FFA curves of reliefF

(d) RFA curves of reliefF

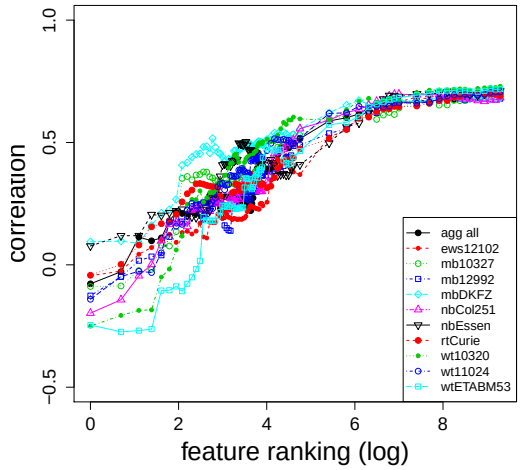
Figure 72: Comparison of the FFA (left) and RFA (right) of the aggregated feature ranking with the individual feature rankings for different **embryonal tumor** (ET) datasets. The aggregation is performed with the min aggregation function



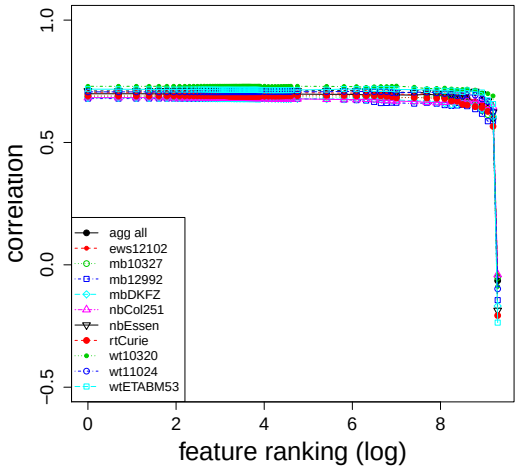
(a) FFA curves of random forests



(b) RFA curves of random forests



(c) FFA curves of reliefF

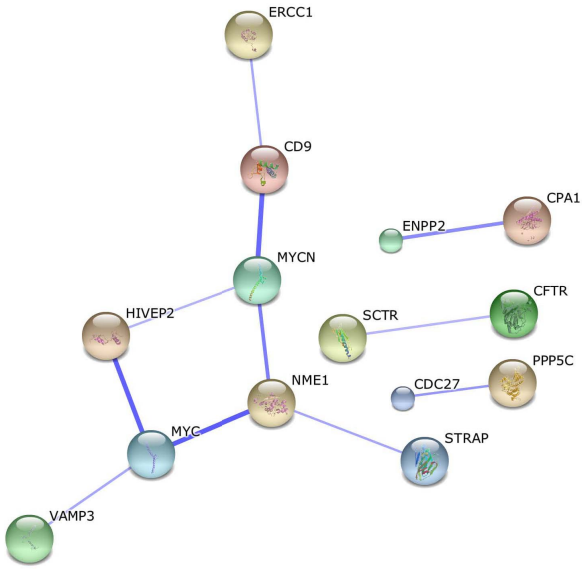


(d) RFA curves of reliefF

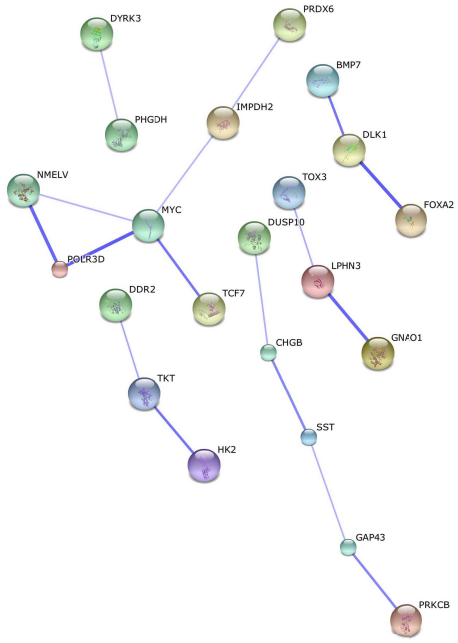
Figure 73: Comparison of the FFA (left) and RFA (right) of the aggregated feature ranking with the individual feature rankings for different **embryonal tumor** (ET) datasets. The aggregation is performed with the max aggregation function

B.6 Gene Networks of Aggregated ET Datasets

We present the full set of gene networks constructed with the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING). We compare the networks constructed from the top-50 genes provided by aggregating the gene lists from the individual entities. The base rankings used for aggregation are constructed from Random Forests or RReliefF. The aggregation is performed by using the mean, median, min or max aggregation function.

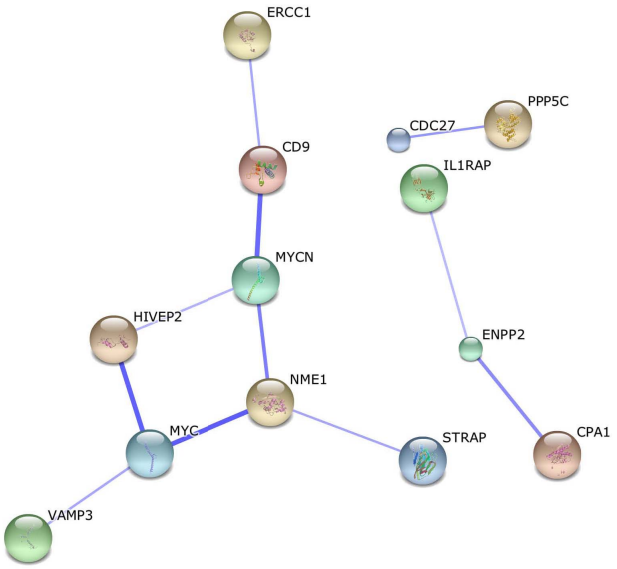


(a) Gene network constructed from the top-50 genes provided by the aggregation of RFs

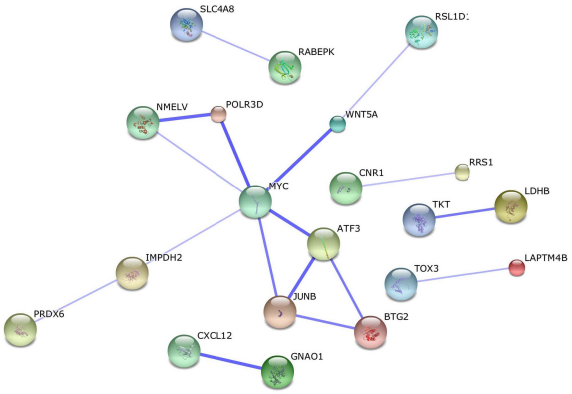


(b) Gene network constructed from the top-50 genes provided by the aggregation of reliefF

Figure 74: Gene networks constructed from the top-50 genes of the aggregated ranking of different feature ranking algorithms. The aggregation is performed with the mean aggregation function

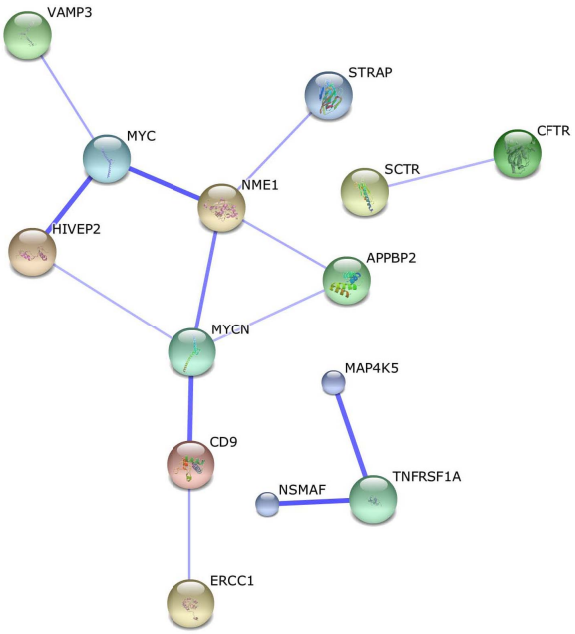


(a) Gene network constructed from the top-50 genes provided by the aggregation of RFs

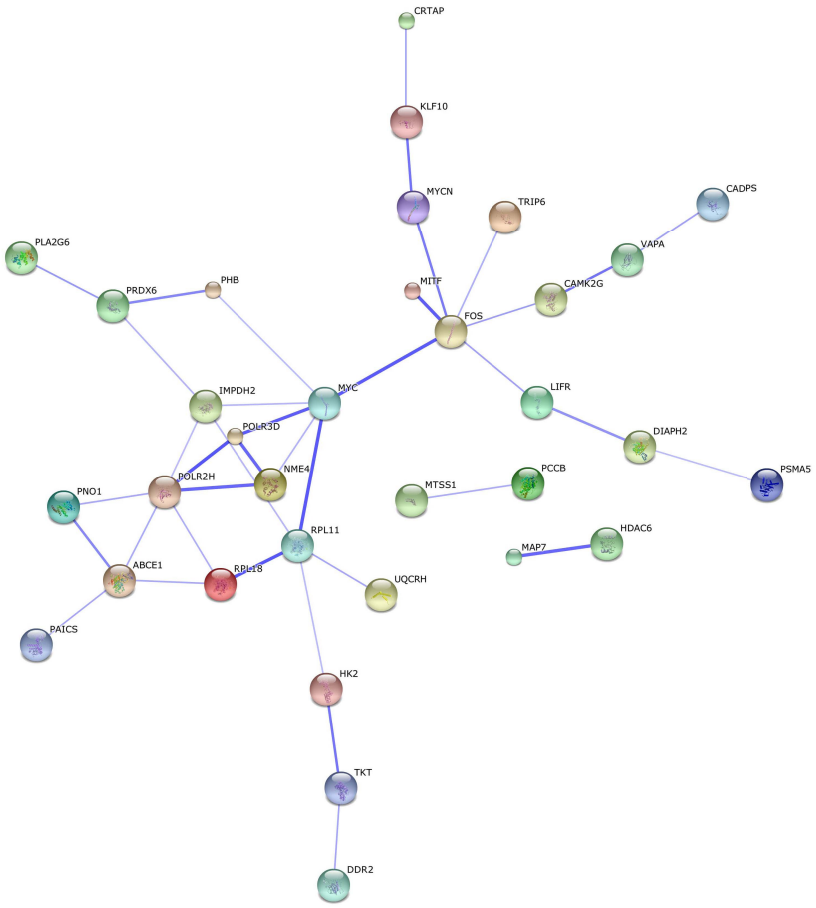


(b) Gene network constructed from the top-50 genes provided by the aggregation of reliefF

Figure 75: Gene networks constructed from the top-50 genes of the aggregated ranking of different feature ranking algorithms. The aggregation is performed with the median aggregation function

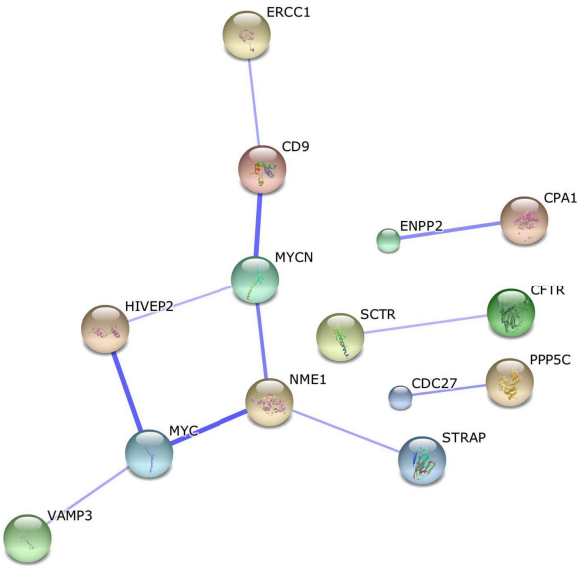


(a) Gene network constructed from the top-50 genes provided by the aggregation of RFs

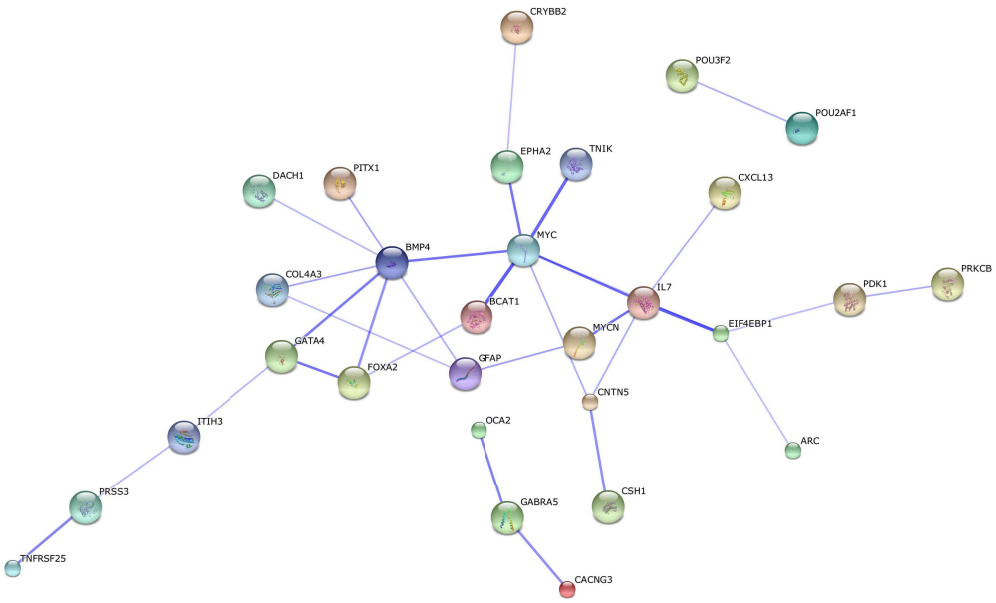


(b) Gene network constructed from the top-50 genes provided by the aggregation of reliefF

Figure 76: Gene networks constructed from the top-50 genes of the aggregated ranking of different feature ranking algorithms. The aggregation is performed with the min aggregation function



(a) Gene network constructed from the top-50 genes provided by the aggregation of RFs



(b) Gene network constructed from the top-50 genes provided by the aggregation of reliefF

Figure 77: Gene networks constructed from the top-50 genes of the aggregated ranking of different feature ranking algorithms. The aggregation is performed with the max aggregation function

C Bibliography

Journal Publications

- Slavkov, I.; Gjorgjioski, V.; Struyf, J.; and Džeroski, S. Finding explained groups of time-course gene expression profiles with predictive clustering trees. *Molecular BioSystems* **6**, 729–740 (2010). [JCR IF = 3.825]

- Lovrečić, L.; Slavkov, I.; Džeroski, S.; and Peterlin, B. ADP-ribosylation factor Guanine nucleotide-exchange factor 2 (ARFGEF2) : a new potential biomarker in Huntington's disease. *Journal of international medical research* **38**, 1653–1662 (2010). [JCR IF = 1.068]

- Slavkov, I.; Džeroski, S.; Peterlin, B.; and Lovrečić, L. Analysis of Huntington's disease gene expression profiles using constrained clustering. *Informatica medica slovenica* **11**, 43–51 (2006).

Book Chapters

- Slavkov, I.; and Džeroski, S. Analyzing gene expression data with predictive clustering trees. *Inductive databases and constraint-based data mining*. 389–406 (Springer, 2010).

Conference and Workshop Publications

- Džeroski, S.; Gjorgjioski, V.; Slavkov, I.; and Struyf, J. Analysis of time series data with predictive clustering trees. In *Selected, Revised and Invited papers from the 5th International Workshop on Knowledge Discovery in Inductive Databases /*, LNCS 4747. 63–80 (2007).

- Slavkov, I.; Džeroski, S.; Struyf, J.; and Loskovska, S. Constrained clustering of gene expression profiles. In: *Proceedings of the 8th International Multiconference Information Society* **A**, 212–215 (2005).

- Slavkov, I.; Džeroski, S.; Peterlin, B.; and Lovrečić, L. Analysis of Huntington's disease gene expression profiles using constrained clustering. In *Proceedings of the 1st Conference of Slovenian bioinformaticians*. 62–63 (2006).

- Džeroski, S.; Gjorgjioski, V.; Slavkov, I.; and Struyf, J. Analysis of time series data with predictive clustering trees. In: *Proceedings of the 5th International Workshop on Knowledge Discovery in Inductive Databases* . 47–58 (2007).

- Slavkov, I.; Pensa, R.; and Džeroski, S. Using bi-sets that characterize bi-partitions as features for classification : an application for microarray data analysis. In *Proceedings of the 9th International Multiconference Information Society* **A**, 60–63 (2006).

- van de Koppel, E.; Slavkov, I.; Astrahantseff, K.; Schramm, A.; Schulte, J.; Vandesompele, J.; de Jong, E.; Džeroski, S.; and Knobbe, A. Knowledge discovery in neuroblastoma-related biological data. In: *Proceedings of the 2nd Workshop in Data Mining in Functional Genomics and Proteomics at ECML/PKDD 2007* **A**, 45–56 (2007).

- Ristevski, B.; Džeroski, S.; Loskovska, S.; and Slavkov, I. A comparison of validation indices for evaluation of clustering results of DNA microarray data. In: *Proceedings of the 2nd International Conference on Bioinformatics and Biomedical Engineering.* 587–591 (2008).

- Slavkov, I.; Ženko, B.; and Džeroski, S. Evaluation method for feature rankings and their aggregations for biomarker discovery. In: *Proceedings of the 3rd International Workshop on Machine learning in systems biology.* 115–124 (2009).

- Slavkov, I.; Aleksovski, D.; Savage, N.; Walburg, K.; Ottenhoff, T.; and Džeroski, S. Discovering groups of gene with coordinated response to M. leprae infection. In: *Proceedings of the 4th International Workshop on Machine learning in systems biology.* 163–166 (2010).

- Slavkov, I.; Ženko, B.; and Džeroski, S. Evaluation method for feature rankings and their aggregations for biomarker discivery. In: *Proceedings of the 3rd International Workshop on Machine learning in systems biology (MLSB 2009) - JMLR proceedings* **8**, 122–135 (2009).

- Slavkov, I.; Džeroski, S.; Peterlin, B.; and Lovrečić, L. Analysis of gene expression profiles for Huntington's disease patients with predictive clustering trees. In: *Proceedings of the 2006 Congress of the Slovenian society for medical informatics*. 164–175 (2006).

- Slavkov, I.; and Džeroski, S. Making the right choice : an evaluation method for ranked lists of biomarkers. In: *Proceedings of the 9th International Conference on Systems Biology*. 6 (2008).

- Kocev, D.; Slavkov, I.; and Džeroski, S. More is better: ranking with multiple targets for biomarker discovery. In: *Proceedings of the 2nd International Workshop on Machine Learning in Systems Biology* **A**, 133 (2008).

- Slavkov, I.; and Džeroski, S. Making the right choice : an evaluation method for ranked lists of biomarkers. In: *Proceedings of the 3rd FEBS Advanced Lecture Course on Systems Biology*. 175 (2009).

- Lovrečić, L.; Krainc, D.; Džeroski, S.; Slavkov, I.; and Peterlin, B. Haemotranscriptomics in Huntington's disease. In: *From arrays to understanding diseases and pharmacogenomics of individual drug therapy*. 48 (2011).

D Biography

Ivica Slavkov was born on April 14th, 1982 in Skopje, Macedonia. He finished his primary and secondary education in Skopje, Macedonia. In 2000, he started his undergraduate studies at the Faculty of Electrical Engineering at the University "Sts. Cyril and Methodius" in Skopje, Macedonia. He finished his undergraduate studies in 2005 with the BSc thesis "Predictive clustering trees for microarray and patient record data analysis", under the supervision of Prof. Dr. Suzana Loškovska and co-supervision of Prof. Dr. Sašo Džeroski.

In the fall of 2005 he enrolled in the PhD programme entitled "New Media and E-science" at the Jožef Stefan International Postgraduate School in Ljubljana, Slovenia, under the supervision of prof. dr. Sašo Džeroski. From 2005 till present he works as a student at the Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana. During the period of 2005/2006 he also worked part-time for the University Medical Center Ljubljana, Department of Medical Genetics. During this period, he worked on the analysis of microarray data on patients with Huntington's disease. In the period between 2006 and 2009, he worked on the EU funded FP6 project "European Embryonal Tumor – Pipeline". His research on this project concerned with the development of methods for meta-analysis of data produced from different platforms and from different tumor entities. The focus of his work within the project is on biomarker discovery. In the period of 2010–2012 he worked on the EU funded FP7 project "Systems biology of phagosome formation and maturation - modulation by intracellular pathogens" (PHAGOSYS), where his work mainly concerned analysis of time series of microarray data, with predictive clustering methods. Ivica Slavkov also holds a scholarship for doctoral studies awarded by Slovene Human Resources and Scholarship Fund Ad futura since the fall of 2005 and a scholarship from the Department of Knowledge Technologies, Jožef Stefan Institute since the fall of 2006.