# EVOLUTIONARY RISK-SENSITIVE FEATURE SELECTION FOR LEARNING TO RANK

DANIEL XAVIER DE SOUSA

# EVOLUTIONARY RISK-SENSITIVE FEATURE SELECTION FOR LEARNING TO RANK

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Doutor em Ciência da Computação.

ORIENTADOR: MARCOS ANDRÉ GONÇALVES
COORIENTADOR: THIERSON COUTO ROSA

Belo Horizonte - MG

Julho de 2018

DANIEL XAVIER DE SOUSA

# EVOLUTIONARY RISK-SENSITIVE FEATURE SELECTION FOR LEARNING TO RANK

Thesis presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Doctor in Computer Science.

ADVISOR: MARCOS ANDRÉ GONÇALVES
CO-ADVISOR: THIERSON COUTO ROSA

Belo Horizonte - MG

July 2018

Sousa, Daniel Xavier de

S725e Evolutionary Risk-Sensitive Feature Selection for Learning to Rank / Daniel Xavier de Sousa. — Belo Horizonte - MG, 2018
xxiv, 75 f. : il. ; 29cm

Tese (doutorado) — Federal University of Minas Gerais
Orientador: Marcos André Gonçalves

1. Computação - Teses. 2. Recuperação de Informação. 3. Aprendizado de ranqueamento. I. Orientador. II. Coorientador. III. Título.

CDU 519.6*73(043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

# FOLHA DE APROVAÇÃO

Evolutionary Risk-Sensitive Feature Selection for Learning to Rank

**DANIEL XAVIER DE SOUSA**

Tese defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. MARCOS ANDRÉ GONÇALVES - Orientador
Departamento de Ciência da Computação - UFMG

PROF. THIERSON COUTO ROSA - Coorientador
Instituto de Informática - UFG

PROFA. GISELE LOBO PAPPA
Departamento de Ciência da Computação - UFMG

PROF. RODRYGO LUIS TEODORO SANTOS
Departamento de Ciência da Computação - UFMG

PROF. EDLENO SILVA DE MOURA
Departamento de Ciência da Computação - UFAM

PROF. RICARDO DA SILVA TORRES
Instituto de Computação - UNICAMP

Belo Horizonte, 5 de outubro de 2018.

*Aos meus filhos, Manuela e Lucas Daniel, que tiveram que entender quando o papai ia brincar sozinho no computador.*

# Agradecimentos

É possível que em 20 anos toda a teoria apresentada nas seções seguintes se tornem ultrapassadas, e estas páginas percam sua importância. Então, me reservo o direito de colocar nessa seção as coisas que realmente terei vontade de ler quando esse tempo chegar.

De forma alguma a tese aqui apresentada foi elaborada exclusivamente no tempo do doutorado. Ela é parte de um processo de formação em pesquisa e computação, iniciado ainda na graduação, passando pelo mestrado e chegando no doutorado que se encerra. As conclusões aqui obtidas só foram possíveis, pois em vários momentos nesse processo de formação me foi dado o privilégio de cultivar o pensamento computacional, analisar situações e a liberdade de questionar. Isso posto, os agradecimentos aqui descritos em grande parte transcendem ao tempo do doutorado.

Agradeço a Deus e a Nossa Senhora. Eles foram meu sopro de esperança quando as previsões futuras não eram animadoras. Foram a força para lutar contra meus próprios defeitos.

Agradeço também aos meus pais, Maria Xavier de Sousa e Silésio José de Sousa, que me deram o melhor presente possível: a liberdade de seguir meus princípios e minha natureza, apoiados em condutas de corretude e amor. Como coerdeiros desse presente, agradeço pela amizade, carinho e amor dos meus irmãos Bruno Xavier de Sousa e Fellipe Xavier de Sousa.

Nesse processo de formação agradeço imensamente pelas pessoas que me instigaram, me desafiaram e mostraram o prazer do amor aristotélico ao se fazer pesquisa; meu orientador de iniciação científica e sempre parceiro Dr. Wellington Santos Martins, meu orientador do mestrado, Dr. Sérgio Lifschitz, e orientadores do doutorado, Dr. Marcos André Gonçalves e Dr. Thierson Couto Rosa. Um agradecimento especial ao professor Dr. Marcos André Gonçalves que me conduziu durante esse doutorado, e mostrou um olhar mais pragmático e simples para a ciência. Ao professor Dr. Thierson Couto Rosa, dono de um grande poder definir conceitos, que escutou diversas das minhas ideias fracassadas, mas nem por isso se fez ausente durante todo o processo de doutorado.

Agradeço aos meus amigos do Laboratório de Banco de Dados (LBD) da UFMG que se

fizeram presentes, sejam nos momentos de distração, saboreando o pão-de-queijo com café do ICEx, sejam nos momentos de discussão e tentativas de refutar hipóteses. São eles, Amir Khatibi, Cristiano da Silva, Clebson Sá, Daniel Hazan, Felipe Viegas, Guilherme Gomes, Rodrigo Silva, Reinaldo Fortes, Sérgio Canuto e Thiago Henrique. Faço um agradecimento especial ao amigo Sérgio Canuto, que dividiu comigo diversas madrugadas na submissão de artigos.

Agradeço também aos professores amigos do Instituto Federal de Goiás, Câmpus Anápolis, que foram importantes nessa conquista do doutorado, pois se mostraram parceiros ao acumular atividades me deixando com maior disponibilidade, ou mesmo me dando carona para o aeroporto nas viagens semanais. São eles, professores Alessandro Silva, Hugo Vinícius e Thiago Eduardo.

Aos meus amigos próximos Petras de Souza e Renato Lima Novais, que em diversas conversas me animaram, aconselharam e se fizeram verdadeiros parceiros.

Agradeço ao apoio financeiro da Fundação de Amparo à Pesquisa do Estado de Goiás (FAPEGO), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e Instituto Federal de Goiás.

A minha amada sogra, Abadia Batista, que foi o apoio necessário em diversas oportunidades, permitindo que eu ficasse tranquilo mesmo distante de casa.

Por último e não menos importante, agradeço a minha amada esposa Joane Batista de Sousa. Agradeço pelo seu amor, por ter feito do meu sonho, nosso sonho, da minha luta, nossa luta. Agradeço por me mostrar que o racional nem sempre é o mais importante, e por me ensinar o poder da alegria. Obrigado por ser um mar de amor que banha a mim e a nossos filhos, Manuela e Lucas Daniel.

*Apesar do aprendizado de máquina computar eventos e encontrar modelos para inferir o futuro, acredito que Deus bagunce qualquer padrão para evitar que uma vida seja um simples modelo matemático.*

(Daniel Xavier de Sousa)

# Resumo

Aprendizado de Ranqueamento (AR) é uma das principais linhas de pesquisa em Recuperação de Informação contudo, com o crescente aumento de dados e complexos algoritmos de aprendizado de máquina, o esforço para processar todas as subtarefas em AR tem crescido enormemente. Mais especificamente, após um algoritmo de ordenação fornecer um imenso subconjunto de documentos (às vezes gigabytes), existe um extenso trabalho na fase de AR para gerar novas meta-atributos no ato de execução da consulta e para executar algoritmos considerados estado da arte de aprendizado de máquina. Nesse contexto, a seleção de atributos (SA) tem se tornado importante alternativa para eliminar atributos não relevantes, pois, além de melhorar o tempo de execução dos AR, criando menos meta-atributos em tempo de execução da consulta e usando menos meta-atributos nos algoritmos de aprendizado de máquina para construir o modelo de ranqueamento, a SA também pode melhorar a efetividade com a ausência de atributos ruidosos e redundantes.

Por anos, porém, a literatura tem focado principalmente em efetividade e redução de atributos como os principais critérios objetivos para SA, no entanto, ao remover certos atributos pode se deteriorar a efetividade de modelos de aprendizado para algumas importantes e específicas consultas. De fato, nós temos notado, em nosso trabalho a otimização de somente a efetividade média como métrica pode deteriorar a acurácia de algumas consultas, enquanto melhora somente as consultas que são de mais alta performance.

Dessa forma, nesta tese nós propomos avaliar SA para AR com um objetivo adicional em mente, conhecido por sensibilidade ao risco, que em linhas gerais permite avaliar a robustez do modelo, garantindo boa efetividade entre as consultas e minimizando a perda de efetividade em consultas quando comparado a outros modelos de ranqueamento. Nós apresentamos novos objetivos uni e multicritério para otimizar SA, efetividade e sensibilidade ao risco, algumas vezes ao mesmo tempo. Para obter nossas metas, consideramos distintas medidas de sensibilidade ao risco, tais como $F_{RISK}$, $T_{RISK}$, e $G_{RISK}$[1]. Como resultado dessa

[1] As métricas $F_{RISK}$, $T_{RISK}$ e $G_{RISK}$ serão detalhadas mais adiante, mas de forma geral consideram diversas formas de comparar a robustez de um modelo em relação a um ou a um grupo de modelos de ranqueamento considerados baselines.

atuação, mostramos que sensibilidade ao risco é um critério objetivo crucial em SA para AR, promovendo, inclusive, resultados melhores do que quando usamos a efetividade como critério objetivo. Isso porque, diferente do valor médio utilizado para comparação, a sensibilidade ao risco avalia todas as consultas em relação a um ou a vários outros métodos de Recuperação de Informação, provendo mais rigor na comparação entre dois subconjuntos de atributos.

No intuito de avaliar nossa proposta de critérios objetivos para SA em AR, também propomos uma nova metodologia para explorar o espaço de busca com diversos objetivos, sugerindo extensões efetivas e eficientes do já bem conhecido algoritmo evolucionário SPEA2. Por efetividade, aplicamos uma comparação mais rigorosa para o conjunto de atributos, usando um teste estatístico pareado para aumentar a confiança no relacionamento de dominância. Por eficiência, introduzimos um algoritmo de aprendizado fraco como uma caixa-preta para melhorar a avaliação das diversas interações dos conjuntos de atributos nos procedimentos baseados em *wrapper*. Apesar de parecer contra intuitivo, conseguimos aprimorar o tempo de execução e a comparação dos atributos de forma mais acurada, melhorando a efetividade na seleção final do indivíduo para critérios multiobjectivos.

Nossos resultados experimentais mostram que a proposta multiobjetivo aperfeiçoa os métodos de SA estado da arte, considerando a combinação de efetividade e sensibilidade ao risco. Por exemplo, na coleção WEB10K conseguimos manter a efetividade e sensibilidade ao risco, reduzindo em até 35% dos atributos. Ainda, nós mostramos fortes evidências quanto ao benefício de usarmos aprendizado fraco como uma caixa-preta e a melhoria na seleção final do indivíduo a partir da Fronteira de Pareto, através do uso do teste pareado. Nesta tese, fornecemos, ademais, uma ampla análise da nossa metodologia e de seus impactos na redução de atributos, sensibilidade ao risco e efetividade em SA para AR.

# Abstract

Learning to Rank (L2R) is one of the main research lines in Information Retrieval. However with ever increasing data and more complex machine learning algorithms, the effort to process all sub-tasks in L2R has increased tremendously. More specifically, after a ranking algorithm provides a huge subset (sometimes gigabytes) of documents from query terms, there is an extensive work of L2R phase to generate meta-features on the fly and to process the time consuming state-of-the-art machine learning algorithms. In this context, *feature selection* (FS) becomes an important alternative to withdraw unimportant features. Besides improving the overall L2R execution time, FS can also try to improve the effectiveness with the absence of noisy and redundant features.

However, for years the literature has focused mostly on effectiveness and feature reduction as the main objective criteria for Feature Selection. But removing certain features may damage the effectiveness of the learned model for some specific but important queries. In fact, we have noted in our work that by optimizing only an average effectiveness and number of features as criteria in FS for L2R one can deteriorate the ranking effectiveness of some queries, providing less robust models.

Therefore, in this dissertation we propose to evaluate FS for L2R with an additional objective in mind, named **risk-sensitiveness**. We introduce the risk-sensitiveness to the FS for L2R, providing novel single and multi-objective criteria to optimize feature reduction, effectiveness and risk-sensitiveness, sometimes at the same time. To achieve our goal, we consider distinct risk-sensitive measures, such as $F_{RISK}$, $T_{RISK}$, and $G_{RISK}$. As results of this front, we show that risk-sensitiveness is a crucial objective criterion in FS for L2R, providing still better results than the effectiveness criterion. Mainly because more than an average value, risk-sensitiveness assesses the comparison of several queries against one or a set of Information Retrieval baselines, providing a larger comparison of two subsets of features.

In order to evaluate our new objective criteria for FS in L2R, we also propose a new methodology to explore the multi-objective search space, suggesting **effective** and **efficient** extensions of wrapper and a well-known Pareto Frontier algorithm, e.g. Strength Pareto

Evolutionary Algorithm (SPEA2). By effective, we mean a more strict comparison for sets of features, using a paired statistical test to increase the strength of the dominance relationship in the Pareto set. In case of the efficient extensions, we introduce a weak learner as a black-box in order to improve the evaluation of the wrapper strategy. Besides decreasing the time performance, this proposal also provides a more accurate comparison of features, improving the effectiveness of the final individual for the evolutionary process.

Our experimental results show that the proposal objective criteria outperforms the state-of-the-art FS methods concerning effective and risk-sensitive evaluation. For instance, for WEB10K dataset we allow a feature reduction of up 35% with same effective and risk-sensitive performance. Moreover, we show that the risk-sensitiveness criterion provided results more effective and robust than using only effectiveness. We show strong evidence towards the benefits of using weak learner as a black-box and the improvements of selecting the final individual from the Pareto set by using the paired statistical test. In this dissertation we also provide a thorough analysis of our methodology and its impact on feature reduction, risk-sensitiveness and effectiveness on FS for L2R.

# List of Figures

# List of Tables

# Contents

# Chapter 1

# Introduction

Learning to Rank (L2R) has established itself as an important research area in Information Retrieval (IR). This is because L2R is the central task in many important IR applications such as modern Web search engines, recommendation and question-answering systems [Liu, 2011]. In general, L2R applies machine learning algorithms to improve the ranking quality by using annotated information about the relevance of documents.

To obtain good results, L2R strategies usually rely on dense representations exploiting dozens of features, some of which are expensive to generate. In several scenarios, some of these features may introduce noise or may be redundant, increasing the cost of the learning process without bringing benefits or even harming the learned ranking model.

Thus, Feature Selection (FS) techniques have been examined in the L2R scenario [Naini and Altingovde, 2014; Pan et al., 2011; Geng et al., 2007] to improve processing time and increase effectiveness by removing noisy and redundant features. FS indeed may have a high positive impact on processing time in L2R [Geng et al., 2007; Naini and Altingovde, 2014; Chapelle et al., 2011]. In addition to the training time, there is also the cost of constructing the features (actually meta-features) as they are generated by several algorithms (e.g., BM25, PageRank) and some of them need to be computed at query time.

Nevertheless, effectiveness and cost (better summarized by the number of exploited features) are not the only objectives one may want to optimize in a L2R task. In fact, recently the **risk** of getting very poor effectiveness for a few queries with a learned model has gained much attention [Wang et al., 2012; Dinçer et al., 2014a; Collins-Thompson et al., 2014]. This interest in diminishing risk is due mainly to the fact that users tend to remember the few failures of a search engine very well rather than the many successful searches [Knijnenburg et al., 2012]. In fact, the authors in [Zhang et al., 2014] clearly show that improvements in ranking performance do not always correlate with risk reduction. This has motivated research in *risk-sensitive L2R* which considers the risk aspect of L2R models [Dinçer et al.,

2014b,a]. The goal of the risk-sensitive L2R task is to enhance the overall effectiveness of a ranking system while reducing the risk of performing poorer than a baseline ranking system for any given query.

## 1.1 Problem Statement

We claim that feature selection used with the intent specifically to enhance efficiency and effectiveness may be a problem to risk-sensitiveness in L2R. This happens because FS reduces the feature space when considering only overall effectiveness or cost as objectives. Thus, it is possible that the reduction of features may worsen the ranking of documents for a few queries (but important ones, such as medical searches), despite improving the ranking for many others. Therefore, there may be features that, despite not significantly improving the ranking effectiveness average, enhance the quality of few queries, providing a more robust performance.

Figure 1.1 provides evidence of the above claim. It shows different rankings[1] in $x$-axis for some features of the MLSR-WEB10K dataset[2]. The first ranking sorts the features considering effectiveness, measured in terms of NDCG@10. The other four rankings correspond to the same features using four different weights of the GeoRisk risk-sensitive function[3]. Each feature corresponds to a colored line in the figure, guided by the rank position of features (in $y$-axis) in each ranking.

Figure 1 shows that some features have an essential behavior on ranking effectiveness, but they are less important from a risk-sensitive perspective, whereas the opposite occurs with other features. In other words, Figure 1.1 illustrates an essential aspect of FS in L2R: the filtering of features considering only the optimization of effectiveness as a criterion may prune important features that would help to generate more robust (less risky) models. Hence, the problem proposed in this dissertation concerns the selection of features with risk-sensitiveness as a main objective criterion (without loss of effectiveness). We show the importance of setting risk as an explicit objective, as noisy features from an effectiveness perspective are not necessarily irrelevant or harmful in a risk-sensitive context. Furthermore, the selection of features when using effectiveness as a single objective criterion may incur in higher risk, mainly because the methods tend to optimize an average metric such as Mean Average Precision (MAP) or NDCG, despite potential losses in few points.

[1] For this case, we sort the features using effectiveness and risk-sensitive measures.

[2] MLSR-WEB10K is a public dataset, released by Microsoft with 10,000 queries and 136 features. Better described in section 4.1.

[3] GeoRisk provides a risk-sensitive evaluation of model performance, by comparing against a set of baselines. The weights ponder the degradation effect or negative variation of the evaluated model against a set of baselines. It is deeper described in section 2.3.

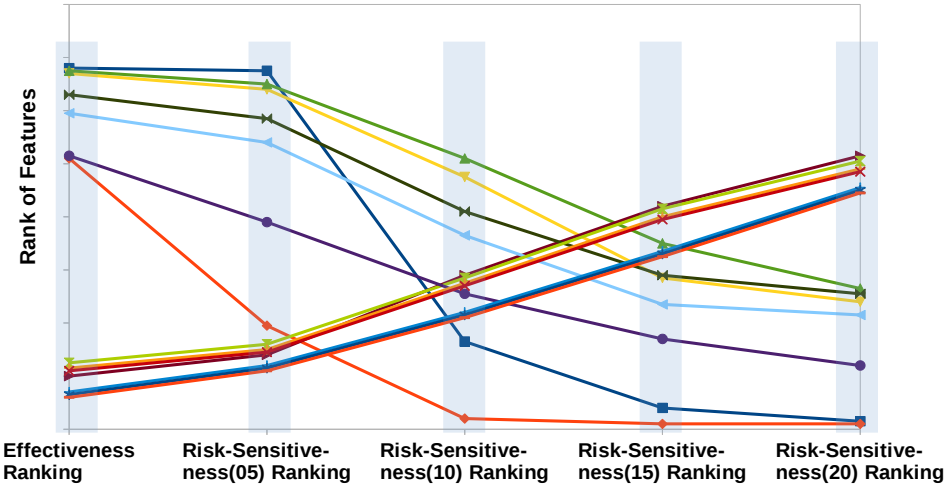



Figure 1.1: *The ranking of some features from MLSR-WEB10k when varying the measures over effectiveness (NDCG@10) and risk-sensitiveness (GeoRisk Dinçer et al. [2016]).*

## 1.2 Research Goals

The above observations have motivated us to address distinct objective criteria in FS for the L2R task. To the best of our knowledge, there are no studies that provide a thorough analysis of the impact of feature selection in both effectiveness and risk-sensitiveness in the L2R literature. Accordingly, the novel proposed objective criteria are: i) maximizing the ranking effectiveness; ii) minimizing the risk for most queries; and iii) reducing the feature space dimensionality, *all at the same time*. Accordingly, we analyze the impact of FS for L2R in these objectives considering them both individually (as single objectives), as well as combined as multi-objectives to be optimized.

By considering a robust and effective evaluation with FS, our dissertation aims to obtain a possibly smaller set of features that guarantees ranking effective and risk-sensitive performance. This is in contrast to existing FS for L2R approaches which goal is to drastically reduce the number of features in order to control the processing time.

We also propose a novel methodology to assess the impact on effectiveness and risk-sensitiveness when diverse (most of the times, conflicting) objective criteria are applied to the FS for the L2R task. Using an efficient and effective wrapper strategy, our proposed methodology explores diverse sets of features as a search space and uses an evolutionary search to select the best feature set according to single or multi-objective criteria. Wrapper strategies are traditionally recognized as time consuming approaches [Qinbao Song et al., 2013; Laporte et al., 2014]. To deal with this issue, in this dissertation we propose to exploit "cheap" weak-learners as black-boxes that make the process more scalable and less costly.

As a positive side effect, weak-learners also promote diversity in the solutions, as strong-learners are more agnostic to the employed set of features. In other words, we drive the exploration of the space of solutions using improvements in both wrapper and multi-objective optimization processing.

To investigate combinations of simultaneous objectives, our proposal uses a multi-objective criteria approach based on Pareto frontier optimization. There are several general-purpose multi-objective optimization methods that can be used in this case, e.g., the works in [Deb et al., 2002; Bandyopadhyay et al., 2008; Zitzler et al., 2001]. We have chosen Strength Pareto Evolutionary Algorithm (SPEA2) [Zitzler et al., 2001], which besides being the state-of-the-art in multi-objective optimization [Guardado et al., 2014], has already been successfully applied to several related problems [Dalip et al., 2014; Li et al., 2015]. In fact, Evolutionary Algorithms (EAs) are able to maximize non-continuous and non-differentiable IR evaluation measures [Wang et al., 2015], e.g. MAP, NDCG@k, and ERR@k. As a result, EAs are well suited to estimate the impact of the distinct proposed objective criteria and also to evaluate our statements, mainly due to their capability of obtaining non-linear ranking functions.

In this dissertation we perform a extended analysis, considering: i) several risk-sensitive measures, ii) a large number of objective combinations, and iii) L2R methods as black-boxes in order to provide our efficient wrapper-based FS. All these improvements result in a novel methodology for assessing several single and multi-objective criteria in FS for the L2R task. For instance, we have shown that a recently published risk-sensitive measure (i.e. GeoRisk [Dinçer et al., 2016]) has an important risk-sensitiveness impact in FS for L2R. In summary, in this dissertation we provide three novel contributions:

1. We open up a new perspective of Feature Selection for risk-sensitive L2R, which highlights the importance of considering risk as an explicit objective criterion. In this context, we are not only considering the average effectiveness obtained by a drastically reduced subset of features, but a subset which provides a risk-sensitive and effective performance;

2. To address the raised problem, we introduce single and multi-objective criteria to perform FS for L2R, considering three important objectives, *concomitantly*: feature dimensionality reduction, effectiveness and risk-sensitiveness. Some of these (conflicting) objective criteria were never evaluated in FS for L2R;

3. A novel efficient and effective evolutionary methodology to evaluate different objective criteria in FS for the L2R task. We apply weak-learners to decrease the execution

time while increasing diversity, and a paired test comparison over a multi-objective evolutionary search to provide an accurate set of features.

4. We provide a broad discussion of the proposed methodology and objectives, showing that, in FS for L2R, distinct goals (with feature reduction or accuracy) can be achieved by varying the objective criteria. Also, most previous works explored only small datasets, while here we consider large datasets, such as MSLR-WEB10K and YAHOO.

## 1.3 Research Questions

To better introduce the ideas in this work, we present the following research questions that guide our investigation.

*Q1 – How to combine different (possibly conflicting) optimization objectives in FS for L2R without being constrained to a particular L2R method?*

In this work we propose to evaluate several objectives (important requirements for FS) and their combinations. However, the challenge of optimizing distinct objectives, considering the possible conflicts among them, demands a multi-objective criteria method. For this task, our methodology applies the Pareto Frontier Set in an evolutionary search[4] as a wrapper FS strategy not constrained to any specific L2R method.

*Q2 – How to apply an efficient wrapper evolutionary FS algorithm over huge datasets, without loss of effectiveness?*

The computation of the fitness value for an individual is time consuming as it is necessary to construct a L2R model with a subset of features corresponding to the individual and to evaluate this model to derive the values for effectiveness and risk-sensitive measures. This has to be done for each individual in the population and is specially time consuming for some large datasets and state-of-the-art L2R algorithms. Hence, one of the key points in this work is the reduction of the searching time during the wrapper-based feature selection.

*Q3 – How to reduce the number of individuals in the Pareto frontier, while keeping individuals that maximize the objective?*

The literature shows that the Pareto frontier set can be large, especially when two objectives are conflicting. This can make the selection within the Pareto set very hard, decreasing the final performance. We here address this selection using a strict comparison over the individuals by the mean of an evolutionary search, using statistical hypothesis tests. As a

[4] In fact, we extend the well-known SPEA2, a general multi-objective optimization method.

result, our method provides a smaller Pareto set with only statistically superior individuals. This has an important impact on the accuracy of our methods, as we shall see.

*Q4 – How good in terms of risk-sensitiveness, effectiveness and feature reduction is the final individual produced by our methodology?*

Differently from all other works in the literature of FS for L2R, we here describe the performance of many objective criteria from a risk-sensitive perspective, and we show that risk-sensitiveness is an important objective criterion in FS. Moreover, we provide a full evaluation of many objective criteria over three dimensions, *concomitantly*: ranking performance, number of features, and risk-sensitiveness. Considering the several intents over FS for the L2R task, we provide clear demonstrations of results for the objectives and evaluated datasets.

*Q5 – How is the overfitting behavior concerning our proposed objective criteria and evolutionary FS methodology?*

For our experiments, we describe the performance of one selected individual from the Pareto Set without describing whether there are other better ones. Hence, we propose to evaluate the content of the Pareto set and to describe the overfitting behavior from each method, by showing the selected individual among all Pareto Set.

*Q6 – Are there groups of features which have a larger impact on risk-sensitiveness than effectiveness?*

Even though we combine multi-objective criteria and wrapper strategy to find a better feature interaction to build a model, we believe that it is possible to evaluate features which improve the risk-sensitiveness rather than effectiveness. In other words, we drive our attention to point out which features or group of features provide more impact on risk-sensitiveness. We show that there are some features which despite not being applied to optimize the effectiveness criterion, are used to support the robustness for some other queries.

*Q7 – What are the effects on the results variation of the proposed statistical test comparison and multi-objective criterion?*

In this work, we have proposed distinct strategies to improve risk-sensitiveness and effectiveness when performing FS in a multi-objective scenario. However, we now pay attention to the effect of each proposal in the experimental results, performing a $2^k$ Factorial Design [Jain, 1991] to discover the result variation obtained for each measure, i.e. risk-sensitiveness and effectiveness. As a result, we have observed that the statistical test also improves the risk-sensitiveness, as it performs a model comparison concerning all available

queries.

The rest of this dissertation is organized as follows. Section 2 presents background and related work. Section 3 describes our proposal. In Section 4 we describe our experimental evaluation, presenting the answers to our research questions. Finally, Section 5 summarizes our conclusions and next steps.

## 1.4 Publications

The main evaluations described in this dissertation are also presented in the following papers:

1. Sousa, D.; Canuto, S. ; Couto, T. ; Martins, W. ; Goncalves, M. . *Incorporating Risk-Sensitiveness into Feature Selection for Learning to Rank*. In: the 24th ACM International on Conference on Information and Knowledge Management, 2016, Indianapolis, EUA. CIKM, 2016.

2. Sousa, D.; Canuto, S. ; Couto, T. ; Martins, W. ; Goncalves, M. . *Risk-Sensitive Learning to Rank with Evolutionary Multi-Objective Feature Selection*. In: ACM Transactions on Information Systems, ACM TOIS. Waiting for the second stage of revision.

# Chapter 2

# Background and Related Work

## 2.1 Impacts of the Feature-Space in the L2R Task

Besides being widely used in industry, Learning to Rank has been successfully applied to a variety of research areas, such as Question Answering [Severyn and Moschitti, 2015], Recommender Systems [Shi et al., 2010] and Document Retrieval Systems [Joachims, 2002].

As defined in [Liu, 2011], L2R learns a function $f : \mathbb{R}^m \rightarrow \mathbb{R}$ capable of maximizing a retrieval effectiveness measure when the documents are sorted by the values of $f$, where $\mathbb{R}^m$ is the vector space of $m$ features in $\mathbb{R}$, derived from documents and queries. In a desired scenario, we would apply L2R to all documents, but since this is not usually possible due to the huge amount of data available, a Web Search Engine executes the Ranking Pipeline, with two phases: i) the Ranking Phase, which first produces an initial ranking processing, and ii) the L2R Re-Ranking Phase, where the documents are re-ranked [Capannini et al., 2016].

The Ranking Phase aims at retrieving most of the relevant documents, also called the top-k set, trying to maximize the recall of the retrieval system. To provide the top-k as a set of relevant documents, the Ranking Phase performs a matching over user-query and documents, exploiting a fast base ranker (e.g. BM25 [Liu, 2011]).

Afterward, the L2R Re-Ranking Phase[1] uses the retrieved documents and a Machine Learning (ML) learned model in order to obtain a better re-ranking of the documents that maximizes the precision. The L2R Re-Ranking Phase uses the top-k documents obtained in the Ranking Phase to create a feature-space with relevance related features or meta-features[2]. We mean meta-features because they are derived from ranking algorithms such as term-

[1] This phase is also called Learning to Rank Framework in Liu [2011].

[2] The resulting set is formed by the union of the pairs (query, top-documents), forming the training and 'test' sets, in which each document is represented by the set of features and meta-features obtained in the Ranking Phase.

weighting scores (e.g., TF-IDF , Cosine Similarity, and Language Models [Liu, 2011]), that provide measures for the query document relationship. Notably, there is great effort to create these features – in some datasets up to 700 algorithms are executed.

With this feature-space, the L2R Re-Ranking Phase uses a supervised ML algorithm as a top ranker to re-rank the documents in top-k. Unlike the first phase, the top ranker maximizes precision, trying to put on top of the ranking the most relevant results. Such precision relies on still large training datasets with gigabytes of annotated query-document examples [Capannini et al., 2015]. In the case of ensemble ranking algorithms (the state-of-the-art algorithms in the L2R Phase [Mohan et al., 2011]), at learning time, an iterative and expensive process is performed over the whole training dataset. For instance, with the LambdaMART algorithm [Mohan et al., 2011] in each iteration, a regression tree is created over a huge training set.

In this context, the Ranking Pipeline (Initial Ranking and L2R Re-ranking Phases) is a complex process that can be expensive in scenarios in which the L2R model needs to be often updated. This update requirement has been previously defined as the Flexibility of Web Search Engines [Chapelle et al., 2011], where flexibility is the temporal ability of L2R Re-Ranking Phase to better adapt to the necessity of updating the training set in short time [Sousa et al., 2012; Chapelle et al., 2011]. Due to the frequent changes on the Web, ranking functions need to be re-learned repeatedly more often, and the Web Search Engine needs to provide fresh models, such as novel ground truth datasets [Sousa et al., 2012]. As a result, we have to consider that the Ranking Pipeline is a continuous task, and FS can reduce the time cost to provide models with recent changes to the users [Capannini et al., 2015]. In distinct scenarios, this temporal flexibility issues have been addressed in [Capannini et al., 2016] by evaluating some known L2R algorithms on effectiveness and time cost, learning the most effective ranker for a given time budget. We have exploited in [Sousa et al., 2012; Freitas et al., 2016] some approaches with many-core processor technology, i.e. Graphical Processing Units (GPU), to generate a fast Re-Ranking method based on Lazy Association Rules that builds on-demand L2R models, obtaining up to 508x of speed up against the serial version.

To sum up, the impacts of the feature-space in the L2R task are: i) in the time to create the features for the top-k documents obtained in the Ranking Phase, ii) in the training time of the top ranker to build the model, iii) in the test (or prediction time), to evaluate the model with the features of the new queries, and iv) in the flexibility of Web Search Engines. Accordingly, with the feature space dimension reduction one can improve time processing for new executions of the L2R Re-Ranking Phase. Besides that, in this dissertation we address these challenges with the advantage of taking risk into account.

## 2.2 Feature Selection in L2R

First of all, it is important to define that in this dissertation we do not consider dimensionality reduction as a method to build new and summarized features, such is the case of Principal Component Analysis or PCA Hastie et al. [2009]. For our work, we provide a reduced dimensionality by selecting existing features and using its subset to build more effective and risk-sensitive models.

For FS works, there are several exploiting L2R. They can be divided into three main strategies: embedded, filter, and wrapper. Embedded strategies select the features by trying to minimize the training error during the learning of a model. Thus, the objective function searches for the best minimal subset of features using a specific L2R error function. For instance, in [Lai et al., 2013] an embedded strategy called FSMRank is proposed. It attempts to minimize ranking errors while performing FS using a combination of importance and similarity measures. In [Laporte et al., 2014], the authors use sparse regularized SVMs to compute non-convex penalties, which leads to similar effectiveness score while reaching feature reduction. However, as asserted in [Laporte et al., 2014], these embedded solutions are designed for specific L2R algorithms, making them generally hard to adapt to other L2R alternative methods. This is an important limitation, as the area evolves with better L2R solutions.

In the filter approach, the selection strategy evaluates some quality measure (e.g. similarity and correlation) for the features without involving any learning algorithm. More usually, the user defines a predefined parameter $k$ of best features which are maintained and the others are filtered out. The selected features are then used to learn a final model by a L2R algorithm. For instance, in [Naini and Altingovde, 2014] the authors present several filter methods that select the most relevant and diverse features, applying diversification techniques such as Minimum Redundancy Maximum Relevance (mRMR) [Peng et al., 2005]. The work in [Shirzad and Keyvanpour, 2015] has further evaluated the mRMR concept, considering a non-linear feature selection method for L2R. They select a subset of $k$ features such that relevance and dissimilarity among the features are optimized. In these aforementioned works, the importance of a feature is computed one at a time, however as described in [Pan et al., 2011; Das, 2001], the worth of a feature depends on the set of other features it interacts with.

Wrapper strategies perform the selection of best features subsets based on the effectiveness of a "generic" L2R algorithm, known as a L2R black-box which is optimized by the learning procedure. The black-box computes the worthiness of a subset of features during the exploration of the search space. For instance, the authors in [Pan et al., 2011] propose a wrapper strategy based on Evolutionary Algorithms (EA), performing an evolutionary process by

eliminating "weak" features over the generations in order to reduce the dimensionality. As a result, they can achieve a reduced number of features with small losses in effectiveness. Furthermore, the work shows the importance of considering the interaction of features in order to apply the FS. However, the final number of features has to be set as a parameter, which is difficult to determine considering the large range of features. Besides, the main solution optimizes only one objective – relative feature importance.

Although wrapper solutions are adaptable to search for the best subset of features to a particular L2R algorithm or dataset, there are very few FS works using wrapper strategies for L2R, specially because they are time consuming. In fact, often the same (state-of-the-art) L2R algorithm used to select the feature subset as a black-box, is also applied to build the final model. The application of good learning methods (and complex ones) in both the training phase and to learn the final model supports the comments of authors in [Qinbao Song et al., 2013], who states that, despite providing the best performing feature subset, wrapper methods are very computationally intensive.

Our proposed methodology for FS using distinct single and multi-objective optimization makes use of wrapper solutions. However, instead of using a state-of-the-art L2R algorithm during the FS phase, we opted for a faster weak learning method. As described in Section 4.3, we consider as a weak learner an algorithm that performs relatively better than a random method and has a fast execution. This approach produces substantial improvements on the FS processing time and the quality of the selected features as we shall discuss in Sections 3.2.3 and 4.3. Furthermore, by improving the time performance, our methodology provides a feasible method to perform wrapper strategies in large datasets for the FS task.

Although our proposal of exploiting a fast weak learner in a wrapper approach has not been used for FS in L2R, as far as we know, it has already been used in classification arena. The work in [Das, 2001] shows experimentally that a feature subset that allows one algorithm to improve the accuracy should also contribute to a different algorithm to obtain a high accuracy performance, considering most real-world datasets in classification field. In our work, besides showing this novel evaluation with distinct L2R algorithms, we provide a possible explanation for this behavior, as described in Section 4.3.

Our wrapper strategy also differs from previous ones as it exploits a multi-objective Pareto-efficient method to evaluate several objectives. For this propose, there are several general-purpose multi-objective algorithms, such as NSGAII [Deb et al., 2002], AMOSA [Bandyopadhyay et al., 2008], and SPEA2 [Zitzler et al., 2001]. all of them applying Pareto set in distinct approaches to deal with objective-criteria. In our dissertation we adapted SPEA2 [Zitzler et al., 2001] to our experiments, which besides being the state-of-the-art in evolutionary multi-objective algorithms [Guardado et al., 2014], has been used such as a success in the literature [Li et al., 2015; Dalip et al., 2014; Guardado et al., 2014]. For instance,

in [Dalip et al., 2014] the authors use SPEA2 with two competing criteria: minimizing the number of features while maximizing effectiveness for the task of determining the quality of collaborative content on the Web. As a future work, we intend to verify the behavior of distinct multi-objective algorithms, evaluating the performance in L2R datasets. In any case, even with SPEA2, we are able to obtain relevant results, as we shall see.

In our dissertation, we execute the SPEA2 by combining the risk-sensitive evaluation as a criterion, which besides evaluating the robustness of a model, is used for the first time in the L2R literature as a feature selection objective-criterion.

## 2.3 Risk and Risk-Sensitive Evaluation

The wide diversity in effectiveness among queries when several IR systems are applied had already called the attention of IR researchers at the beginning of this century [Voorhees, 2005]. Over the years 2003 [Voorhees, 2003], 2004 [Voorhees, 2004], and 2005 [Voorhees, 2005] the Robust Retrieval Track (proposed by Text REtrieval Conference - TREC) investigated the difficulty of specific queries even with a high precision system on average. One important conclusion of TREC was that optimizing the standard average effective performance can harm some difficult queries, improving only the better-performing ones [Voorhees, 2005]. More currently, this variability of query precision for several IR systems has motivated the study of the risk-sensitiveness concept.

In [Wang et al., 2012] risk-sensitiveness is decomposed in *degradation* and *reward*, where the degradation (reward) of the model $M$ corresponds to the negative (positive) variation of queries evaluation regarding a specific IR system baseline. Suppose that we are given a set of training queries $Q_T$, and two ranking models: a baseline $B$ and a proposed model $M$. The degradation of the model $M$ corresponds to the average difference (or gain) in the effectiveness of the baseline $B$ against $M$ in all queries in $Q_T$. This definition of degradation was formally stated in [Wang et al., 2012] through the $F_{RISK}$ function defined in Eq. 2.1.

$$F_{RISK}(Q_T, M) = \frac{1}{|Q_T|} \sum_{g \in Q_T} max[0, B(g) - M(g)] \tag{2.1}$$

where $B(q)$ and $M(q)$ denote the effectiveness value of the baseline and the new model for a given query $q$, respectively. Note that the $F_{RISK}$ function uses the effectiveness of each query in $Q_T$, which can be measured by any commonly-used IR evaluation measure following the-higher-the-better values, such as AP, MRR, NDCG@$k$ [Liu, 2011]. The main goal for this function is to evaluate the difference between two models when assessing the same IR measures. Hence, as the value of function $F_{RISK}$ decreases, it improves the chance of having

a robust model, as there is less degradation of the model $M$ against the baseline model $B$. In this work we adopt the function $F_{RISK}$ as the definition of the degradation of a model $M$ with respect to a baseline $B$.

Degradation is a negative variation and an important concept in ranking systems. In [Knijnenburg et al., 2012], the authors argue that the few failures a search engine makes get more noticed by the users than the many successful searches. The same authors also performed an ample study on user experience in recommender systems, finding out how negative high-variance is for the users. Consequently, the minimization of degradation has attracted the attention of researchers as an important additional objective for a ranking solution [Wang et al., 2012; Dinçer et al., 2014b,a]. Furthermore, according to Wang et al. [2012] and Collins-Thompson et al. [2014], *robustness* is the ability of a ranking solution to minimize the degradation.

Contrary to degradation, the *reward* of a proposed method $M$ in relation to a baseline model $B$, is defined as the average gain in effectiveness of model $M$ against the baseline $B$ in all queries in $Q_T$ [Wang et al., 2012]. In [Wang et al., 2012] reward is formally stated by the function presented in Eq. 2.2:

$$F_{REWARD}(Q_T, M) = \frac{1}{|Q_T|} \sum_{q \in Q_T} max[0, M(Q) - B(Q)] \tag{2.2}$$

Reward and degradation can be combined in different ways to evaluate how much a method $M$ is sensitive to risk. The term *risk-sensitive task* was coined in the TREC 2013 Web track [Collins-Thompson et al., 2014] as the trade-off a system can achieve between effectiveness (overall gains across queries) and robustness, both regarding a baseline [Wang et al., 2012; Dinçer et al., 2014a]. In other words, a method is risk-sensitive if it can improve the ranking of most queries and does not decrease the ranking performance of other ones concerning a baseline ranking system (from now on referred to as *BS4R*, an acronym for Baseline System for Risk). Thus, the risk-sensitive task corresponds to a multi-objective optimization solution for the ranking problem which aims to maximize effectiveness and minimize the risk[3].

In [Wang et al., 2012], a measure to evaluate sensitiveness to risk is defined by means of the function $U_{RISK}$, which aggregates functions $F_{RISK}$ and $F_{REWARD}$ in a single *tradeoff function*. $U_{RISK}$ is the objective function that the proposal in [Wang et al., 2012] aims to maximize. Function $U_{RISK}$ is defined as:

$$U_{RISK}(Q_T, M) = F_{REWARD}(Q_T, M) - (1 + \alpha) F_{RISK}(Q_T, M) \tag{2.3}$$

[3]Minimizing the risk is equivalent to maximizing robustness.

The parameter $\alpha$ is the weight given to the degradation ($F_{RISK}$). Different values of $\alpha$ can significantly impact the risk-sensitive evaluation of the method, in [Wang et al., 2012; Dinçer et al., 2014a] its range varying between 0 and 20. As described in [Wang et al., 2012], $\alpha = 0$ provided similar interpretation to the IR evaluation measure used, without the risk-sensitive analysis.

The work described in [Dinçer et al., 2014a] extends the work in [Wang et al., 2012] by proposing a generalization of the $U_{RISK}$ function which is referred to as $T_{RISK}$.

$$T_{RISK}(Q_T, M) = \frac{U_{RISK}(Q_T, M)}{SE(U_{RISK}(Q_T, M))} \tag{2.4}$$

where $SE$ is the estimation of the $U_{RISK}$ standard error. We have applied the regular standard error of the mean to $U_{RISK}$, that it is, $\sigma(U_{RISK})/\sqrt{|Q_T|}$, where $|Q_T|$ means the cardinality of $Q_T$ and $\sigma$ the variance of values in $U_{RISK}$.

The proposed function, $T_{RISK}$, uses inferential hypothesis testing for evaluating a risk-sensitive task. The inferential techniques proposed in the paper enable us to: a) decide whether an observed level of risk for an IR system is statistically significant and b) determine the queries that individually lead to a significant level of risk.

On the other hand, the authors of [Dinçer et al., 2014b] study how the ranking method used as BS4R can affect the risk-sensitive evaluation. They show that the choice of an appropriate BS4R is of great importance in ensuring an unbiased risk-sensitive measurement of the performance of individual systems. In particular, the higher the correlation between any given system $M$ and the BS4R across queries, the higher the measured risk-sensitive scores of $M$ on average. This implies a bias in the estimation of the risks. The paper suggests some unbiased baselines, such as mean or maximum ranking performance over several ranking methods.

In the same vein, the work in [Dinçer et al., 2016] investigates the use of multiple BS4R in risk-sensitive evaluation, regarding not only mean and the variance of the observed losses and wins, but also the shape of the score distribution when using a set of ranking systems as risk-baseline. The authors claim that using a set of systems as BS4R is the proper way to know the difficulty of each query, avoiding queries that are badly predicted by a single system, but not for others. For this propose, the paper performs the Chi-square test statistics in order to calculate the expectation of the ranking effectiveness for each query using the overall performance of both current system and other risk-baseline systems. The function $Z_{RISK}$ is defined in [Dinçer et al., 2016] as:

$$Z_{RISK}(i) = \left[ \sum_{q \in Q_+} z_{iq} + (1+\alpha) \sum_{q \in Q_-} z_{iq} \right] \tag{2.5}$$

where

$$z_{iq} = \frac{x_{iq} - e_{iq}}{\sqrt{e_{iq}}}, e_{iq} = S_i \times \frac{T_q}{N}, \tag{2.6}$$

and $x_{iq}$ is the effectiveness of a query $q$ obtained with the corresponding system $i$. The element $i$ is defined as $i \in \{1,2,...,r\}$ for each system, where $r$ is the number of systems, and the element $q$ is defined as $q \in \{1,2,...,|Q_T|\}$, where $|Q_T|$ the max size of queries. Both $Q_+$ and $Q_-$ are sets of positive and negative $z_{iq}$, respectively. Let $S_i$ be the expected system performance for all queries in IR system $i$, $T_q$ the within-query IR system effectiveness for the query $q$, and $N = \sum_{i=1}^{r} \sum_{q=1}^{Q_T} x_{iq}$ the sum of all elements.

As the $Z_{RISK}$ computes the risk-sensitiveness regardless of the mean effectiveness of systems, it does not provide a comparative risk-sensitive evaluation of different systems. Accordingly, the same authors proposed a Geometric Mean of $Z_{RISK}$ for this purpose, called $G_{RISK}$ function:

$$G_{RISK}(S_i) = \sqrt{S_i/Q_T \times \Phi(Z_{Risk}(i)/Q_T)} \tag{2.7}$$

where $\Phi()$ is the cumulative distribution function of the Standard Normal Distribution. Basically, the $G_{RISK}$ provides a ranking systems comparison with a robustness perspective, evaluating each given query against a performance expectation. This expectation is obtained from the population of the observed ranking effectiveness of systems for a specific query.

All the aforementioned works aim to enhance the risk-sensitive task without considering FS. In this dissertation we propose evaluations over several objective criteria for FS strategies, considering a multi-objective Pareto efficient method to optimize, besides the trade-off between effectiveness and risk-sensitiveness.

# Chapter 3

# Feature Selection Proposal

## 3.1 Motivation

Our main motivation is to answer our first research question: *Q1 – How to combine different (potentially conflicting) optimization objectives in FS for L2R without being constrained to a particular L2R method?*

The usual main intent of FS is the reduction of features without harming effectiveness. However, focusing only on this objective may harm both the effectiveness of some queries and mainly the overall risk-sensitiveness. For instance, trying to minimize the number of features while optimizing the ranking performance may generate very specialized solutions (with low generalization), with few features fitting a group of queries that maximize an average metric such as Mean Average Precision (MAP) or NDCG. However, this reduced set may increase the risk of getting poor effectiveness for some other queries, as shown in Section 4.

Indeed, although the number of features is an important criterion in FS, we claim that there are other important ones which are capable of obtaining relevant results in effectiveness and risk-sensitiveness, while still reducing the feature space. In cases where the intent is a feature reduction without decreasing risk-sensitiveness, our experimental results (described in Section 4.5.2) show that the combination of risk-sensitiveness and effectiveness is better than effectiveness along with the number of features. For this reason, we propose a methodology to evaluate single and multi-objective criteria, considering, *at the same time*: i) effectiveness and risk-sensitiveness; ii) effectiveness, risk-sensitiveness, and feature reduction; or iii) risk-sensitiveness and feature reduction.

We exploit an evolutionary process that attempts to optimize some pre-defined objectives by varying the set of features to be used in the L2R model as a wrapper method. In

cases in which the number of features is not a direct objective, our methodology still tends to reduce the dimensionality while improving the other objectives (e.g. effectiveness and risk-sensitiveness) due to the elimination of noisy, redundant[1]. Section 3.2.3 provides further discussion on this subject. As we shall see in Section 4, this process allows us to obtain a good feature reduction, without harming the risk-sensitiveness for some objective combinations.

In Section 3.2 we present our evolutionary multi-objective proposal, which makes use of SPEA2 [Zitzler et al., 2001] to select a set of features. SPEA2 is the state-of-the-art in the evolutionary processing [Guardado et al., 2014], besides being able to optimize several proposed objective combinations at the same time, as reported in [Li et al., 2015].

## 3.2 Evolutionary Multi-Objective FS

SPEA2 is based on Genetic Algorithms [Srinivas and Patnaik, 1994] and thus uses an evolutionary approach to explore the solution space for multi-objective problems. In our case, this solution space corresponds to the power set of the set of features used in L2R, and a particular solution corresponds to a set of features, also referred to as an *individual*. In this process, each individual receives a *fitness value* that scores its worthiness based on its likelihood of surviving in the next generation. Once the fitness values have been computed for each individual in one generation, the best individuals are selected to take part in the breeding of the next one. These selected individuals are kept in an archive $A_g$ during generation $g$. Thus, in the process, the archive works like a bucket to keep the best individuals over the generations. On the other hand, the unfit individuals are eliminated during this evolutionary process. After many generations, surviving individuals (or their descendants) tend to be better than the eliminated ones, according to the fitness criteria. This process is summarized in Figure 3.1, where besides describing the overview of SPEA2 process, it also highlights the parts addressed by our work and explained in Sections 3.2.1, 3.2.2, and 3.2.3.

In more details, Algorithm 1 describes the original SPEA2. The algorithm takes as input the size $n$ of the population, the size $a$ of the archive, and the number $ng$ of generations. A population, $P_g = \{i_0, ..., i_n\}$, is the set of individuals in a generation $g$. In our case, each individual corresponds to a binary array (aka, a chromosome) in the feature space. A position in the array is defined as a *gene* and corresponds to a feature. It is 0 when the feature is absent in the individual and 1 otherwise. The algorithm first creates an empty archive $A_1$ and a population $P_1$ with $n$ individuals in lines 1 and 2, respectively.

[1]This idea is supported by the work in Li and Yang [2005], or high-risk features

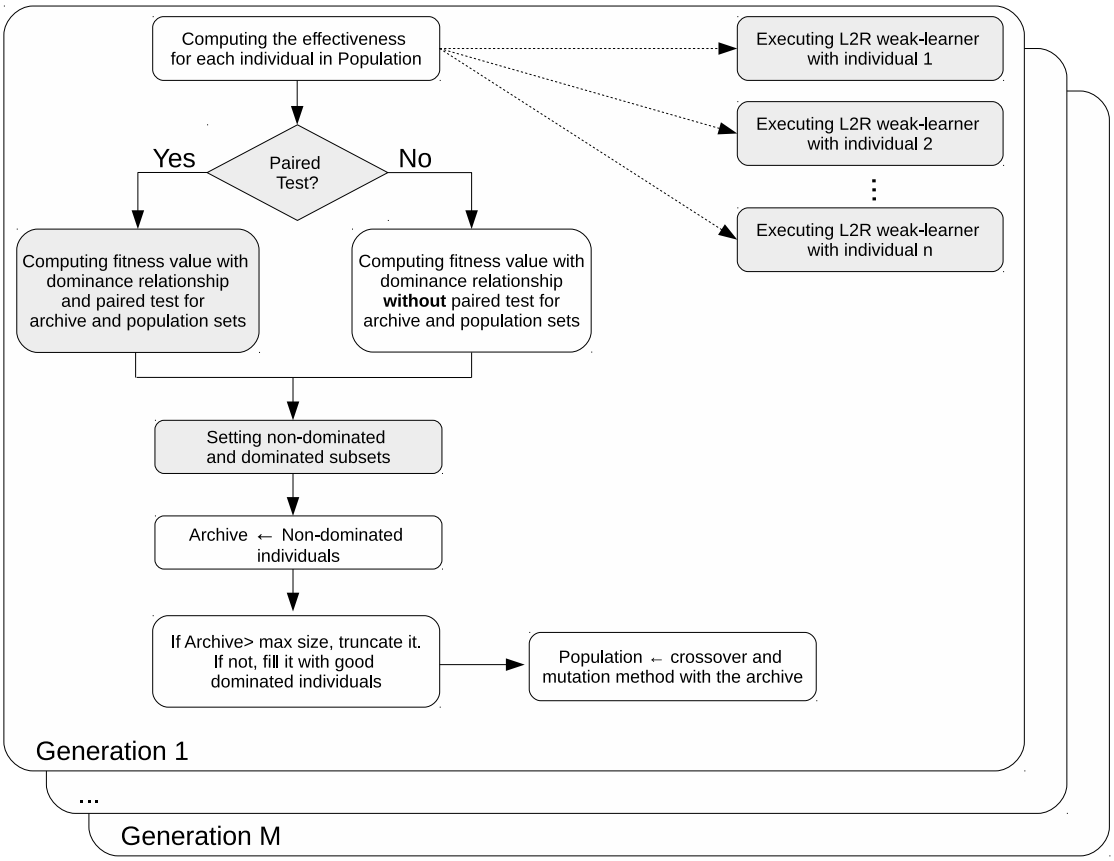



Figure 3.1: *The SPEA2 process highlighting the proposals addressed in this work, gray parts.*

Instead of using only randomly generated individuals for the initial population, we also include in $P_1$ some synthetic ones to explore the relevant search space regions faster. Thus, we include synthetic individuals generated from Random Forest Feature Importance algorithm, or RFFI[2]. Random Forest Feature Importance sorts the features according to feature impurity, and less impurity means higher position at each tree of the forest. In this case, each synthetic individual has a range of the best features ordered by RFFI. For instance, the first synthetic individual has 5 features from the top, the second one 10 features from the top, and so on.

Once all individuals have been created, the fitness score for each one is computed (Line 4). When assigning scores to individuals, SPEA2 must consider the optimization of multiple objectives. Thus, the algorithm uses the *dominance relationship* among individuals to provide the fitness values. Let $x$ and $y$ be two potentially conflicting objectives. Let also $i$ and $j$ be two different individuals. Individual $i$ *dominates* $j$ ( denoted as $i \succ j$ ), if and only if $(x_i > x_j \wedge y_i \geq y_j) \vee (x_i \geq x_j \wedge y_i > y_j)$. In other words, $i$ dominates $j$ if $i$ is better than $j$ in one objective, and $i$ is not worse than $j$ in the other one. An individual $i$ is in the Pareto

[2]using the public library Scikit-Learn(http://scikit-learn.org/)

**Algorithm 1** The Original SPEA2 Algorithm.

**Require:** Population size $n$
**Require:** Size $a$ of Archive ($A_g$)
**Require:** Number of generations $t$
**Ensure:** $A_g$ close to Pareto frontier
Let $P_g$ = pop. of individuals $\{i_0, ..., i_n\}$ of generation $g$
Let $A_g$ = the best individuals of all generations until $g$
Let $D_g$ = dominated individuals of $P_g$ and $A_g$
Let $N_g$ = non-dominated individuals of $P_g$ and $A_g$
$A_1 \leftarrow \emptyset$
**Initialize** $P_1$ with random and synthetic individuals
**for** $g = 1$ to $t$ **do**
  **Compute** fitness(i), $i \in P_g \cup A_g$
  with $i \in P_g \cup A_g$ do:
    **Assign** $i$ to $D_g$ if fitness(i) $\geq 1$
    **Assign** $i$ to $N_g$ if fitness(i) $< 1$
  **Add** $N_g$ to $A_g$
  **if** $\mid A_g \mid > a$ **then**
    truncate($A_g$)
  **else if** $A_g < a$ **then**
    $k = a - \mid A_g \mid$
    Fill $A_g$ with the $k$ best individuals in $D_g$
  $P_{g+1} \leftarrow \emptyset$
  $A_{g+1} \leftarrow A_g$
  **while** $\mid P_{g+1} \mid -1 < n$ **do**
    **Select** two individuals $i_x$ and $i_y$ from $A_g$.
    ($new\_i_x, new\_i_y$) = *crossover*($i_x, i_y$)
    **Add** $new\_i_x$ and $new\_i_y$ to $P_{g+1}$
  **fol all** $i \in P_{g+1}$
    *random_mutate(i)*

frontier, when there is no other individual $j$ that dominates $i$. In this case, $i$ is said to be a *nondominated* individual. The *strength* $S(i)$ of an individual $i$ is defined as the number of individuals who are dominated by $i$, as described in Eq. 3.1.

$$S(i) = \mid \{ j \mid j \in P_g \cup A_g \wedge i \succ j \} \mid \tag{3.1}$$

The fitness score of $i$ is computed by Eq. 3.2.

$$fitness(i) = R(i) + Dens(i) \tag{3.2}$$

and

$$R(i) = \sum_{j \in (P_g \cup A_g) \wedge j \succ i} S(j). \tag{3.3}$$

where $R(i)$ sums the strength of the individuals who dominate $i$. Observe that $R(i)<R(j)$ means that the individual $j$ is worse than individual $i$, since the strength and number of individuals that dominate $j$ are greater than the strength and number of individuals that dominate $i$. Thus, the value of $fitness(i)$ is optimized by minimizing $R(i)$. When $R(i) = 0$, no individual dominates $i$, meaning that all individuals with $R(i) = 0$ are the best solutions, i.e., they belong to the Pareto frontier.

The term $Dens(i)$ in Eq. 3.2 is referred to as *density estimate* in [Zitzler et al., 2001]. It is used to break ties, i.e. individuals with similar dominance. In other words, if two individuals provide the same dominance relationship, the method selects the individual which is more distinct in the population, increasing the variability over the selected population. It is calculated according to Eq. 3.4:

$$Dens(i) = \frac{1}{\sigma_i^k + 2} \tag{3.4}$$

The value 2 in [Zitzler et al., 2001] is used to ensure that $Dens(i)$ is less than 1 and to keep the denominator greater than zero. Also, $\sigma_i^k$ is the distance between individual $i$ and its $k^{th}$ nearest individual using the K-nearest neighbor algorithm [Hastie et al., 2009] with the Euclidean Distance. The parameter $k$ is defined as $\sqrt{|\,A_g\,| + |\,P_g\,|}$.

Observe that $Dens(i)$ is assigned to promote a large variety of solutions, as it decreases when $i$ is farther from a dense region. In this sense, a higher priority is given to the more distinct individual, avoiding the search process to be trapped in a local optimal solution. In addition, an individual with tied $R(i)$ values but in a sparse region will have more chance of surviving to the next generation. This step helps to avoid overfitting, as the algorithms tend to diversity.

After computing the fitness for each individual, the algorithm populates the $D_g$ and $N_g$ sets, putting in $D_g$ the individuals which are dominated by other individuals (Line 6), and in $N_g$ all *nondominated* individuals (Line 7).

Lines 8-13 of Algorithm 1 define the elitism process, saving in the archive ($A_g$) all the *nondominated* individuals of the population. If the archive is full (Line 9), the algorithm removes the individual which is most similar to all other individuals in the archive. We use Euclidean Distance as the similarity measure among individuals. This removal is repeated until the size of the archive becomes equal to the limit $a$. This approach increases the diversity of genotype. If the archive is not full the algorithm fills the archive (Line 13) with the best individuals in $D_g$ (i.e. individuals that despite being dominated have small fitness).

After $A_g$ is full, the algorithm initializes the next generation archive ($A_{g+1}$) with $A_g$ (Line 15). Next, it creates a new population $P_{g+1}$, performing crossover and mutation on individuals of the current archive ($A_g$). Crossover is performed by using the Tournament

Selection method [Srinivas and Patnaik, 1994] (Line 17), which selects the individuals with highest fitness values, among a few sets of individuals chosen at random from $A_g$. Using the *Two Point Crossover* [Srinivas and Patnaik, 1994] method, the crossover is performed (Line 18) exchanging a continuous random sequence of genes between two selected individuals.

In Lines 20 and 21, the algorithm applies a random selection for mutation to each individual. The $random\_mutate(i)$ method flips a coin to perform the mutation for an individual $i$. In the positive case, it flips a coin again for each gene in the chromosome corresponding to $i$, following a Binomial Distribution with a previously defined parameter. In order to improve the variability over the search space, we selected this mutation method in order to set a low probability for a mutation process so that few individuals are mutated. However, when the mutation is performed, it produces a large modification in the chromosome.

After Algorithm 1 is completely executed, it ensures that a set of individuals are in or close to the Pareto frontier, which is a subset of the last archive. In order to select only one individual (as the definitive subset of features), we choose the individual which produces the model with the greatest effectiveness value in the training set.

As far as we know, the use of SPEA2 for selecting features of a L2R model has not been reported in the literature. Indeed, we extended SPEA2 regarding the following aspects: (i) the explicit definition of the exploited dominance relationships, (ii) the use of a statistical test to compute the dominance relationships and (iii) the use of a fast weak learning algorithm as a black-box L2R method to improve the wrapper strategy. We explain the demand for these extensions and how they were performed in the following subsections.

### 3.2.1 Dominance Relationships

In this section we describe in detail how we compute the fitness of an individual $i$. We compute the fitness of $i$ using Eq. 3.2, however, we use different definitions of the dominance relationship ($\succ$) according to the objectives we want to optimize. Since these objectives are: **effectiveness**, **risk-sensitiveness** and **feature reduction**, we need first to obtain a ranking model derived from the L2R black-box algorithm using the set of features forming individual $i$. Secondly, we need to compute measures of effectiveness, number of features and risk-sensitiveness to evaluate the learned model for $i$, according to the objectives considered.

The effectiveness of the model corresponding to an individual $i$ is computed using the values of a IR measure, e.g. MAP, MRR, or NDCG [Liu, 2011]. We refer to the effectiveness value of a model corresponding to an individual $i$ as $eff(i)$. The risk-sensitiveness of the model learned for $i$ is measured using three of the four measures presented in Section 2.3: $F_{RISK}(i)$, $T_{RISK}(i)$ and $G_{RISK}(i)$ (Eq.2.1, Eq.2.4 and Eq.2.7, respectively). We do not use the function $U_{RISK}$ in our evaluations because of its highly correlated results with $T_{RISK}$, as

shown in [Dinçer et al., 2014a]. Finally, the measure used to evaluate the objective of feature reduction is simply the counting of the number of features forming the individual $i$.

Given the model learned for an individual $i$ and the objective evaluation measures, we next give the definitions of dominance relationships we use to compute the fitness of individuals in the SPEA2 algorithm.

**Definition 1.** $i \overset{E\text{-}R}{\succ} j$ *if and only if* $(F_{RISK}(i) < F_{RISK}(j) \wedge eff(i) \geq eff(j)) \vee (F_{RISK}(i) \leq F_{RISK}(j) \wedge eff(i) > eff(j))$.

We use Definition 1 to determine whether an individual $i$ dominates individual $j$ (i.e., $i \succ j$) regarding ranking performance $eff()$ and degradation, $F_{RISK}()$. By using $eff()$ and $F_{RISK}()$ as independent objectives in Definition 1, we are improving the computation of risk-sensitive evaluation in comparison to those computed by $U_{RISK}$, $T_{RISK}$ and $G_{RISK}$, (Eq. 2.3, Eq. 2.4 and Eq. 2.7, respectively). This is because $F_{RISK}$ has no parameter $\alpha$ to be adjusted as is the case of the other three risk-sensitiveness measures. In addition, as $F_{RISK}()$ evaluates only the degradation of a model, it alone cannot be considered a risk-sensitive function. However, applying it with a ranking performance function, we provide an instance of risk-sensitive evaluation, since we are maximizing effectiveness and minimizing the risk of performing poorer than a baseline system.

**Definition 2.** $i \overset{E\text{-}G}{\succ} j$ *if and only if* $(G_{RISK}(i) > G_{RISK}(j) \wedge eff(i) \geq eff(j)) \vee (G_{RISK}(i) \geq G_{RISK}(j) \wedge eff(i) > eff(j))$.

In Definition 2 we combine ranking performance $eff()$ and risk-sensitiveness with multiple BS4R, using $G_{RISK}()$. For this definition, the ranking performance has a greater factor, as it participates on both objectives.

**Definition 3.** $i \overset{T}{\succ} j$ *if and only if* $T_{RISK}(i) > T_{RISK}(i)$

**Definition 4.** $i \overset{G}{\succ} j$ *if and only if* $G_{RISK}(i) > G_{RISK}(i)$

Definitions 3 and 4 explore the straight risk-sensitive objective criterion, $T_{RISK}$ and $G_{RISK}$, respectively. Noting that with only one objective criterion there is no Pareto frontier, and the SPEA2 algorithm becomes more similar to a classic single objective genetic algorithm.

**Definition 5.** $i \overset{E}{\succ} j$ *if and only if* $eff(i) > eff(j)$

We also evaluate effectiveness as a unique objective criterion, in Definition 5. As effectiveness is applied to many FS works in the literature, our main goal using Definition 5 is to evaluate its risk-sensitiveness in our Evolutionary Algorithm.

**Definition 6.** $i \overset{E\text{-}F}{\succ} j$ *if and only if* $(nFeat(i) < nFeat(j) \wedge eff(i) \geq eff(j)) \vee (nFeat(i) \leq nFeat(j) \wedge eff(i) > eff(j))\}$

**Definition 7.** $i \overset{T\text{-}F}{\succ} j$ *if and only if* $(nFeat(i) < nFeat(j) \wedge T_{RISK}(i) \geq T_{RISK}(j)) \vee (nFeat(i) \leq nFeat(j) \wedge T_{RISK}(i) > T_{RISK}(j))\}$

**Definition 8.** $i \overset{G\text{-}F}{\succ} j$ *if and only if* $(nFeat(i) < nFeat(j) \wedge G_{RISK}(i) \geq G_{RISK}(j)) \vee (nFeat(i) \leq nFeat(j) \wedge G_{RISK}(i) > G_{RISK}(j))\}$

As already described, we also optimize other objectives accepting a more drastically feature reduction. Thus, Definition 6 to 8 include number of features, effectiveness and risk-sensitive evaluation functions ($T_{RISK}$ and $G_{RISK}$). The $nFeat(i)$ corresponds to the number of features of individual $i$.

**Definition 9.** $i \overset{E\text{-}G\text{-}F}{\succ} j$ *if and only if*

$$\left[nFeat(i) < nFeat(j) \wedge G_{RISK}(i) \geq G_{RISK}(j) \wedge eff(i) \geq eff(j)\right] \vee$$
$$\left[G_{RISK}(i) > G_{RISK}(j) \wedge nFeat(i) \leq nFeat(j) \wedge eff(i) \geq eff(j)\right] \vee$$
$$\left[eff(i) > eff(j) \wedge G_{RISK}(i) \geq G_{RISK}(j) \wedge nFeat(i) \leq nFeat(j)\right]$$

**Definition 10.** $i \overset{E\text{-}R\text{-}F}{\succ} j$ *if and only if*

$$\left[nFeat(i) < nFeat(j) \wedge F_{RISK}(i) \leq F_{RISK}(j)) \wedge eff(i) \geq eff(j)\right] \vee$$
$$\left[F_{RISK}(i) < F_{RISK}(j) \wedge nFeat(i) \leq nFeat(j) \wedge eff(i) \geq eff(j)\right] \vee$$
$$\left[eff(i) > eff(j) \wedge F_{RISK}(i) \leq F_{RISK}(j) \wedge nFeat(i) \leq nFeat(j)\right]$$

To further evaluate our multi-objective approach, we also perform a combination of three objective criteria in Definition 9 (using $G_{RISK}$) and in Definition 10 (using $F_{RISK}$):

For each objective $O$ (except the number of features and $G_{RISK}$ function) we use a statistical significance test when comparing two individuals according to objective $O$. In the next section, we discuss the importance of this statistical test.

All aforementioned dominance relationships are defined with the intent of assessing the risk-sensitiveness and effectiveness of the proposed single and multi-objective criteria. However, our methodology is absolutely flexible to accommodate any other type of objective combination. As a suggestion, one application could perform feature selection in order to optimize two or three distinct effectiveness measures or more than two risk functions, without directly concerning effectiveness performance.

### 3.2.2 Using Paired Tests in the Dominance Relationships

One important issue when using the Pareto frontier is to select the final individual to learn the definitive model. This is because the obtained Pareto frontier is usually large, especially when the two objectives are conflicting [Wismans et al., 2011]. As a result, the task to evaluate and to select only one individual from several in Pareto set becomes difficult.[3]

To explain why the Pareto set increases, let us consider an example of two non-dominated individuals, $i$ and $j$. Suppose that regarding objective $x$, $i$ is a little greater than $j$, using a scalar value of some measure. Otherwise, taking into account another objective $y$ and also using a scalar value, $j$ can be substantially greater than $i$ regarding $y$. In this case, there is no dominance relationship between $i$ and $j$, as a result, both individuals are kept in Pareto frontier, increasing its size, even though there is a small difference between both individuals for objective $x$.

In this work we deal with this issue by considering a L2R idiosyncrasy, that is, the models learned for individuals $i$ and $j$ are used to generate rankings for each query of the training set. Thus, we can compare both models per query (with regard to the objective $x$) and use the training set as a sample for the evaluation. This allows us to perform a paired statistical test and consequently compare individuals $i$ and $j$ confidently. Using our aforementioned example, the individual $i$ will probably not be statistically different from $j$ on objective $x$, thus defining $i$ dominated by $j$, and consequently, keeping only $j$ as a non-dominated individual. As a result, when using a statistical test we are more strict to assign a difference between two models, providing an improved ranking of individuals and a reduced Pareto set, such as shown by the experiments described in Section 4.4.

In order to describe the influence of the statistical test, Table 3.1 lists all possible dominance relationships of individuals $i$ and $j$ without a statistical test comparison, for one and two objectives, and the possible changes when applying a paired statistical test. The absence of a statistical test is represented by Best Mean, which only uses a regular comparison of the best mean values. For instance, in the first line, the table shows that an equivalent relationship using Best Mean can be kept the same or changed towards $i \succ j$ or $j \succ i$, by using the paired test comparison. The first line also represents a conflict dominance for two individuals in two objectives, but applying statistical test there are three possibles changes. Differently for the remaining lines, where either the dominance relationship keeps the same, or changes towards an equivalent one, i.e, only maximum of one possible changes and the change drives to a tied relationship.

[3] Some works have already addressed this Pareto set size issue [Wismans et al., 2011; Tzeng and Tsaur, 1997]. In Wismans et al. [2011], for instance, some pruning and ranking methods are applied only over the optimal Pareto set, aiming to be used as an assist to the decision making process and providing a better compromise solution.

From Table 3.1, for two objective criteria, the paired test can impact more on the evolutionary comparison as it breaks the tied relationship. Alternatively, for all other options[4], the paired test only keeps the same dominance relationship or changes towards an equivalent one, despite assigning a more strict comparison. Hence, in the case of one objective criterion, we do not expect that the paired test should provide different results, as we do in more than one objective criteria. Section 4.4 describes the experimental results which support our claims.

| | Influence with Paired Statistical Test | | |
|---|---|---|---|
| Best Mean | $i \equiv j$ | $i \succ j$ | $j \succ i$ |
| $i \overset{OBJ_1 - OBJ_2}{\equiv} j$ | Yes | Yes | Yes |
| $i \overset{OBJ_1 - OBJ_2}{\succ} j$ | Yes | Yes | No |
| $j \overset{OBJ_1 - OBJ_2}{\succ} i$ | Yes | No | Yes |
| $i \overset{OBJ_1}{\equiv} j$ | Yes | No | No |
| $i \overset{OBJ_1}{\succ} j$ | Yes | Yes | No |
| $j \overset{OBJ_1}{\succ} i$ | Yes | No | Yes |

Table 3.1: *Consequence of using paired statistical test comparison in one and two objective criteria.*

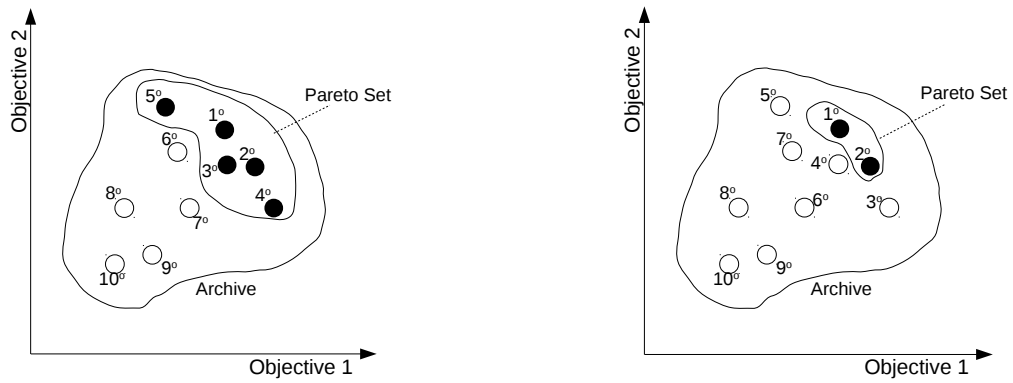
It is also worth observing that the reduction of a Pareto set does not reduce the variability over the evolutionary process. As described in Algorithm 1, the mutation and crossover processes are applied to all archive[5] and population sets, and not only to the Pareto frontier.

In order to illustrate this behavior, in Figure 3.2a the Pareto set is a subset of the Archive, and each individual (represented by circles) has the rank position considering a dominance relationship without a statistical test. Concerning the statistical test comparison performance in Figure 3.2b, the ranking positions of the individuals suffer a minor rearrangement and the Pareto set is diminished, as a consequence of more strict comparison for the dominance relationship, and without changing the size of the archive.

As our experiments show (see Section 4.4), in the cases of datasets with many queries, the significance test improves the comparison of individuals considerably, leading to a strict dominance evaluation. As a consequence, the Pareto frontiers are much smaller than those produced by the conventional method. Furthermore, only high-quality individuals remain in the final Pareto set, improving the selection of the final individual.

[4]Even though the dominance relationship is applied to two or more objective criteria, the dominance for one objective criterion can be interpreted as a relevance superiority.

[5]Remarking that the archive works as a bucket to keep the best individuals over the generations.




(a) Without Statistical Test (b) With Statistical Test

Figure 3.2: *The Pareto sets when using the statistical test over the evolutionary process.*

### 3.2.3 Using a Fast and Weak Learner Algorithm as a Black-Box

In this work our methodology to evaluate multi-objective criteria in FS makes use of a wrapper strategy. However, traditional wrapper strategies exploit the same L2R algorithm both as a black-box during FS and in the final ranking solution after the FS is performed. Usually, a state-of-the-art L2R algorithm is chosen, therefore, raising two important issues: i) as these algorithms are time consuming, wrapper approaches become infeasible in huge datasets and ii) a high-quality learner (i.e. with both low variance and low bias) is usually able to attenuate the presence of "bad" (i.e., noisy, redundant) features, predicting similar accuracy for different individuals. As a result, the state-of-the-art L2R algorithm could not explicitly indicate the inconvenience of "bad" features to the learning process of FS, driving the evolutionary exploration to bad regions in the search space.

Measuring an individual's fitness is crucial in our proposal, as it influences considerably the execution time of SPEA2 and the exploration of the search space. Thus, we claim that, for fitness processing, it is better to use a fast L2R method more sensitive to both, good and bad features, than to use a method which builds a highly effective model. In other words, it is interesting that the L2R algorithm used during the wrapper FS does not attenuate the effect of bad individuals, so that they can be filtered out more effectively from the Pareto frontier during the generation process.

Based on the above assumptions, during the SPEA2 execution we apply weak ranking methods (e.g. Regression Tree or Linear Regression) as the L2R black-box to build the model for each individual. Besides improving the comparison between individuals, they are much faster than most state-of-the-art L2R algorithms. In Section 4.3 we show the large differences between using a weak and a state-of-the-art learner as a black-box in our experiments. Nevertheless, after the final selection of features is done, we apply a more effective method, such as Random Forests [Breiman, 2001] or LambdaMART [Mohan et al., 2011],

all of them well-known very effective and strong L2R methods.

In order words, by using a weak-learner in an evolutionary process for FS, we claim that it is more discriminative to evaluate individuals over the evolving process for FS. Concerning the black-box as a *loss function* for feature selection, the function should optimize the prediction considering the best set of features.

It is worth observing that this exploitation of weaker learning methods to process a training subspace has been used in different machine learning contexts, such as classifier ensembles. Random Forests, for instance, exploits regression trees without pruning, in order to measure the information gain obtained by many training subspaces (aka, bagging). By combining all such biased measurements, it can achieve a model with lower variance [Hastie et al., 2009].

The original SPEA2 and regular wrapper strategies do not use the improvements proposed in this work, namely: (i) paired statistical significance test to measure meaningful differences and (2) a biased learning method as a black-box. These improvements and the novel multi-objective criteria are evaluated in the next section.

# Chapter 4

# Experimental Evaluations

In this chapter, we describe a set of experiments performed to answer our proposed research questions, assigning each subsection for a research question. However, we first describe the datasets, the BS4R, the FS baseline methods and the hyper-parameter settings.

## 4.1 Datasets and Evaluation Measures

We conduct our experiments on four well-known benchmark datasets: MSLR-WEB10K (from Microsoft Research[1]), Yahoo! Webscope dataset version 1 and set 2 (from YAHOO! Learning to Rank Challenge[2]), and LETOR[3] datasets: TD2003 and TD2004. For our evaluation, each dataset was divided into five folds for a 5-fold cross-validation procedure, with three folds for training, one for validation and one for testing. The details of WEB10K, YAHOO, TD2003, and TD2004 datasets are summarized in Table 4.1.

| | Queries | No. docs | No. features |
|---|---|---|---|
| WEB10K | 10,000 | 1,200,192 | 136 |
| YAHOO | 6,330 | 172,870 | 700 |
| TD2003 | 50 | 49,058 | 64 |
| TD2004 | 75 | 74,146 | 64 |

Table 4.1: *Characteristics of the datasets*

As defined in Section 2.3, risk-sensitiveness evaluates the robustness of a ranking solution relative to a defined Baseline System for Risk, or BS4R. When using $T_{RISK}$ and $F_{RISK}$ functions only one BS4R is applied. Conversely, the $G_{RISK}$ function uses more than

[1] http://research.microsoft.com/enus/projects/mslr/
[2] http://research.yahoo.com/Academic_Relations
[3] http://research.microsoft.com/enus/um/people/letor/

one BS4R to evaluate risk-sensitiveness. Furthermore, we have applied distinct BS4R approaches in the training and test phases to measure the risk-sensitive functions.

In the training phase or within the SPEA2 Algorithm, in order to measure the $T_{RISK}$ or $F_{RISK}$ functions for each individual, we use the full set of features combined with the black-box method as the BS4R. Consequently, all new individuals (features subsets) have the model trained with all features as a reference to risk-sensitive evaluation. However, to obtain the $G_{RISK}$ score for an individual, we use all remainder individuals from the generation as the set of BS4R. Recall that $G_{RISK}$ uses the shape of the score distribution of many BS4R to measure the risk-sensitiveness. Hence, all individuals from a population define the shape of robustness, and each individual is compared against the shape to evaluate its risk-sensitiveness. It is worth noting that, in our work $G_{RISK}$ changes the BS4R over the generations, differently from $T_{RISK}$ that uses the same BS4R for all generations. In both cases, all individuals are evaluated using the same machine learning approach. This fulfills the requirements defined in Dinçer et al. [2014b] to be a valid and an unbiased BS4R.

Considering the test set, in order to evaluate the risk-sensitive performance of the final model, we use the Mean, Max and BM25 performance as BS4R for $T_{RISK}$, as also applied in Dinçer et al. [2014b]; Wang et al. [2012]. BM25 is already available as a feature in the dataset, corresponding to the result when the method is applied to the whole document. With Mean and Max Baselines we evaluate the effectiveness (e.g. using NDCG@10) for each feature value as a score for document ranking, corresponding to the average effectiveness for all features for the Mean Baseline, and the highest effectiveness value for the Max Baseline. In the case of $G_{RISK}$, we use the full set of features combined with important L2R algorithms with different solving paradigms: ListNet, AdaRank, LambdaMART, Random Forest and MART. Additionally, to avoid overfitting when evaluating the $G_{RISK}$ of the selected features with Random Forest, for instance, RF is not used as a L2R algorithm to compose the set of BS4R methods.

To report the improvements of the selected features, we use $G_{RISK}$ and $T_{RISK}$ scores. Additionally, we use "*Wins*" and "*Losses* $> 20\%$", following [Dinçer et al., 2014a; Wang et al., 2012]. The measure *Wins* for a method $M$ counts the number of queries for which $M$ wins against the BS4R, ignoring ties. The measure "*Losses* $> 20\%$" (represented by "$L > 20\%$" in our tables) expresses the number of queries for which the relative loss in effectiveness of a method $M$ against the BS4R is higher than 20%. It is worth noting that "$L > 20\%$" has a "less is better" interpretation – which is shown in our tables using the symbol $\downarrow$.

We evaluate the effectiveness for queries of a dataset performing the average of the

NDCG@10 [Liu, 2011] over all queries[4]. Moreover, to ensure the relevance of the results, we assess the statistical significance of our measurements by means of a paired T-test [Sakai, 2014] with 95% confidence.

## 4.2 Hyper-Parameters Definitions and FS Baselines

For risk-sensitive measures ($T_{RISK}$ and $F_{RISK}$) only the $\alpha$ parameter[5] has to be previously defined and analyzed. However, besides the L2R and FS baselines, some "default" parameterization is also necessary for the SPEA2 evolutionary process. For SPEA2, we adopt a parameter setting with some values used in previous works [Laumanns et al., 2001; Pan et al., 2011; Zitzler et al., 2001]. Table 4.2 summarizes them. For instance, the work in Pan et al. [2011] indicates that having more individuals per generation is better than having more generations. Hence, we define population size as 75 (as used in Dalip et al. [2014]) and the number of generations as 30. In addition, as in Zitzler et al. [2001], a large archive size suggests a large elitism, thus we use an archive size of 150 (twice the population size). For the mutation and crossover parameters we follow the guidelines in Laumanns et al. [2001]: individual mutation probability = 0.2, gene mutation probability = 0.3, and crossover probability = 0.8. The *individual mutation probability* is the probability to perform a mutation in an individual, and the *gene mutation probability* is the probability to change a gene. In order to drive the search for better regions, we also insert synthetic individuals in the first population, selecting sets with 5%, 10%, 15%, up to 95% of best features evaluated by Random Forest Feature Importance. We apply these same parameters and settings for all evolutionary executions.

To compute $T_{RISK}$[6] and $G_{RISK}$ in the training phase (inside of SPEA2) we apply the following range of $\alpha$ values (considering Wang et al. [2012]): 1, 5, 10, 15, 20, 25, 30, 35, and 40. From the best evaluation performance on the validation set, we used $\alpha = 35$ for YAHOO and WEB10K, and $\alpha = 5$ for TD2003 and TD20004. On the other hand, to evaluate our method in test set we use the value 5 in $\alpha$ parameter, also suggested in [Dinçer et al., 2014a, 2016]. In fact, other values for $\alpha$ parameter were tested, but we did not find a statistical difference in effectiveness.

To evaluate the feature sets selected by the FS methods, we apply two well-known state-of-the-art L2R methods [Mohan et al., 2011]: Random Forest and LambdaMART. We

[4]We have tested with other metrics such as MAP and NDCG at other positions and the results were qualitatively the same.

[5]Remarking that using a parameter $\alpha$ the risk-sensitive measures apply a linear combination of *reward* and *degradation* values.

[6]As $U_{RISK}$ is a component of $T_{RISK}$, $T_{RISK}$ also uses the $\alpha$ parameter.

| Parameterization Description | | |
|---|---|---|
| SPEA2 | Population Size | 75 |
| | Generation Number | 30 |
| | Archive Size | 150 |
| | Mutation Probability | 0.2 |
| | Gene Mutation Probability | 0.3 |
| | Crossover Probability | 0.8 |
| Risk-sensitiveness | $\alpha$ Values | 1, 5, 10, ..,40 |
| Random Forest | Number of Trees | 100, 200, 300 |
| LambdaMART | Learning Rate | 0.025, 0.05, 0.075, 0.1 |
| | Number of Leaves | 10, 50, 80 |
| | Number of Trees | 100,200, 300, 500, 800 |
| BTFS | Features Rate | 0.02, 0.10, 0.25, 0.50 |
| | k | 10, 30, 50 |
| DivFS | Algorithms | MPT and MMR |
| | k | 0.20, 0.30, 0.40 |

Table 4.2: *Summary of the applied parameters.*

evaluate for Random Forest only the number of trees $\in \{100, 200, 300\}$[7]$\}$ on the validation set and left the remaining parameters with their default values (as in the Scikit-Learn[8]). For the LambdaMART algorithm, we chose the best performing parameters in the validation set for the learning rate, the number of leaves, and the number of trees. We evaluate the following values: *learning rate* $\in \{0.025, 0.05, 0.075, 0.1\}$, *number of leaves* $\in \{10, 50, 80\}$, and *number of trees* $\in \{100, 200, 300, 500, 800\}$. The remaining parameters of LambdaMART follow the QuickRank[9].

In order to evaluate our proposals against other Feature Selection methods, we consider instances of wrapper and filter strategies. Hence, we have used some recent FS methods that were cited in several papers, such as [Pan et al., 2011] (here called BTFS) such as a wrapper instance, and [Naini and Altingovde, 2014] (here called DivFS) such as a filter instance. To evaluate the BTFS method, we select the best parameters considering the best ranking performance over the validation set. We use the same elimination rate as in the original work: 0.02, 0.10, 0.25 and 0.50. The authors consider that using 30 features (the *k* parameter) of their datasets, which originally contained 419 and 367 features, was enough to obtain the same prediction as using the full dataset. Following the authors, we apply 30, 50, 10 and 10 features, respectively in WEB10K, YAHOO, TD2003, and TD2004. Concerning the evolutionary process in BTFS, we use the same parameters used in our solution, also described in

[7] As described in [Gomes et al., 2013], Random Forest is known to be robust to change in parameters, and few changes in the parameters are necessary to obtain good performance.

[8] http://scikit-learn.org/stable/

[9] http://quickrank.isti.cnr.it/

Table 4.2. To evaluate the DivFS method, we implement the best approaches found in the original paper, namely Modern Portfolio Theory and Maximal Marginal Relevance. In both cases, we use the target number of features considering the same rate used by the authors (also the $k$ parameter): 0.20, 0.30, and 0.40. We use the validation set only to evaluate the best parameters, and then apply the best values in the test set.

In the next sections we answer our research questions by means of an extensive set of experiments on the considered datasets. The first two questions are related to our Evolutionary algorithm extensions, and the third concerns our methodology to evaluate several single and multi-objective criteria.

## 4.3 Evaluating the Weak Learner as a Black-Box

*Q2 – How to apply an efficient wrapper evolutionary FS algorithm over huge datasets, without loss of effectiveness?*

As highlighted in Section 3.2.3, an evolutionary algorithm adapted to a wrapper strategy and using a state-of-the-art L2R method has serious time consuming issues, mainly due to the fitness processing. We address this by applying a weak learner as black-box to perform a faster execution and to improve the accuracy of feature selection in the SPEA2 process. We now provide evidence for this claim, assessing the performance of distinct L2R algorithms as black-boxes. Furthermore, we also use this section to describe the weak learners used to evaluate our multi-objective FS proposal, over the next sections.

Initially, we evaluate distinct L2R executions out of the evolutionary process, showing in Figure 4.1 and Table 4.3 the time and accuracy performances, respectively. The experiments execute all features on the training partition of a 5-fold cross-validation procedure, varying the L2R algorithms such as: Linear Regression, Regression Tree, Short Regression Tree, Random Forest, and LambdaMART. Particularly, for Short Regression Tree, we perform tuning of the leaves number on the validation set, applying it in training set to build a shorter tree. We also use the LambdaMART and Random Forest as instances of the state-of-the-art L2R algorithms. As expected, in Figure 4.1 both Random Forest and LambdaMART are the most time consuming, being up to 28x and 121x slower than the Linear Regression, respectively. Despite this, Random Forest and LambdaMART provide the best effective performance (in Table 4.3) against other L2R algorithms. In contrast, the faster ones provided the worst effectiveness performance, such as Linear Regression, Regression Tree and Short Regression Tree. In fact, the faster algorithms create simple models, without proper treatment for bad features. From both Figure 4.1 and Table 4.3 one can consider Linear Re-

gression, Regression Tree and Short Regression Tree as fast and weak learners, and Random Forest and LambdaMART as strong learners.

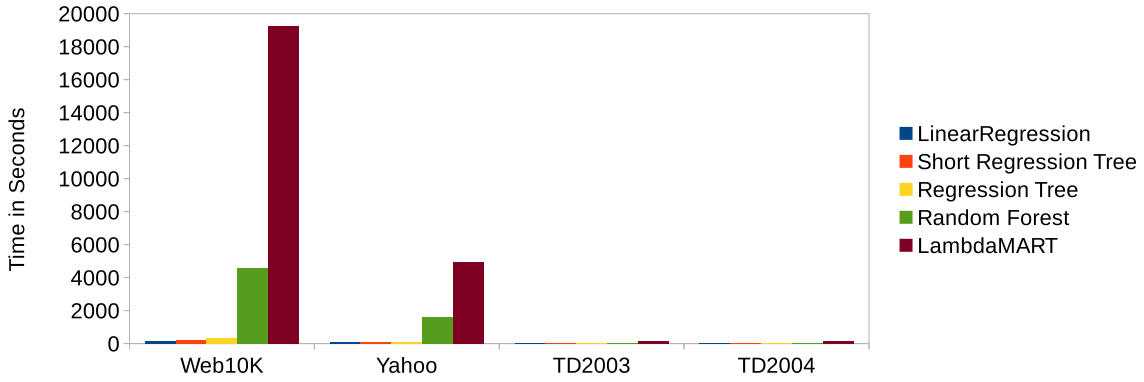



Figure 4.1: *The execution time (in seconds) of L2R algorithms using all features, applying a 5-fold cross-validation on the training set.*

| Datasets | Linear Regression | Short Regression Tree | Regression Tree | Random Forest | LambdaMART |
|---|---|---|---|---|---|
| WEB10k | 0.4042±0.0063 | 0.3896 ±0.0061 | 0.2717 ±0.0069 | 0.4243 ±0.0067 | 0.4482 ±0.0065 |
| YAHOO | 0.6889 ±0.0073 | 0.6645 ±0.0082 | 0.6000 ±0.0083 | 0.7014 ±0.0075 | 0.7160 ±0.0079 |
| TD2003 | 0.3236±0.0772 | 0.2291 ±0.0763 | 0.2276 ±0.0854 | 0.3487 ±0.0778 | 0.2644 ±0.0823 |
| TD2004 | 0.2917±0.0522 | 0.2501 ±0.0510 | 0.2038 ±0.0509 | 0.3393 ±0.0519 | 0.3119 ±0.0534 |

Table 4.3: *The effectiveness (NDCG@10) when processing 5-Fold in the training set with distinct algorithms.*

We now evaluate the aforementioned L2R algorithms as black-boxes in our evolutionary SPEA2 algorithm according to time and effectiveness. First, Figure 4.2 describes the time cost of our SPEA2 when varying only the black-box methods over the evolutionary processing, showing the execution time for 75 individuals and 30 generations in WEB10K dataset and without any tuning of L2R parameters. To avoid any comparison of objective-criteria and statistical test proposals, the experiment in the table uses for all execution the effectiveness as the objective-criterion, $\overset{E}{\succ}$, and the best effectiveness to compare the individuals. This experiment uses only one machine, in order to better describe the performance in serial execution, which is a Intel® i7-870, running at 2.93GHz, with 16GB RAM. In the case of LambdaMART we did not finish the execution due to the long processing time.

As one can observe in Figure 4.2, both Random Forest and LambdaMART are very time consuming as black-boxes. For instance, Random Forest spent up to 30x and 18x longer than Linear Regression and Regression Tree, respectively. The LambdaMART execution was terminated when missing over 30% of processing, despite that, it is 3x longer than the Random Forest. Note that the results in Figure 4.2 follow the individual execution in Figure

4.1. In fact, time performance is an improvement of using a fast learner in the evolutionary process. Depending on some parameters, such as the number of generations and individuals, the processing can be prohibitive for large datasets when a strong learner is used.

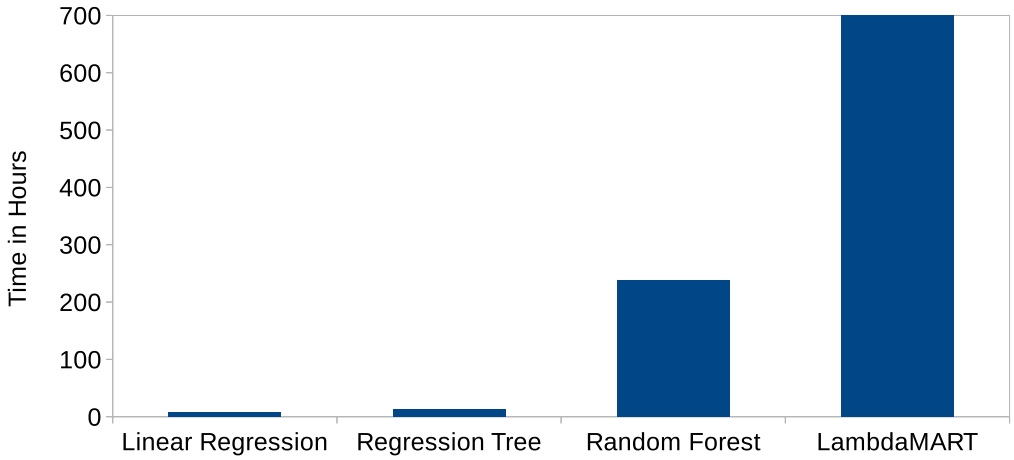



Figure 4.2: *The execution time (in hours) to process our wrapper evolutionary algorithm when varying the L2R algorithms as black-box in WEB10K dataset.*

Although the processing time is a requirement, it is absolutely important that the black-box algorithms provide accurate comparison over the individuals, guaranteeing an accurate feature selection. Hence, Table 4.4 presents the effectiveness of the final model when varying the aforementioned L2R algorithms as black-boxes. Due to the long processing time of LambdaMART, we do not provide its effectiveness results. In addition, as LambdaMART depends on parameter tuning to obtain an accurate evaluation of feature sets, its execution became an impossible experiment in evolutionary wrapper strategies. On the contrary, Random Forest is more robust to the parameter tuning [Gomes et al., 2013], and consequently it is more robust to evaluate several features sets over an evolutionary execution without changing the parameters in the training phase. The results in Table 4.4 correspond to NDCG@10 (with confidence intervals) using Random Forest with the final selected features subset (i.e., the selected individual). Note that besides distinct evaluation black-boxes, by using Random Forest as the final model, we are showing an important baseline, which is a case of performing same L2R algorithm (Random Forest) in both the black-box and in the final model.

Table 4.4 shows that even using a faster weak learner as a black-box the ranking performance of selected features is similar or even better than using Random Forest. As one can see, Regression Tree presented statistically similar results to those of Random Forest as black-box in all datasets. Linear Regression also obtained interesting results, outperforming the Short Regression Tree and providing results close to Regression Tree and Random Forest. Possibly, Short Regression Tree could not provide a proper evaluation of a set of fea-

| Datasets | Linear Regression | Short Regression Tree | Regression Tree | Random Forest |
|---|---|---|---|---|
| WEB10K | $0.4201^{RF}$ ±0.0068 | $0.4176^{RF}$ ±0.0069 | **0.4234** ±0.0069 | 0.4228 ±0.0067 |
| YAHOO | $0.7033^{RF}$ ±0.0085 | $0.6993^{RF}$ ±0.0086 | 0.7042 ±0.0085 | **0.7050** ±0.0085 |
| TD2003 | 0.3272 ±0.0771 | 0.3273 ±0.0765 | **0.3658** ±0.0858 | 0.3441 ±0.0775 |
| TD2004 | 0.3336 ±0.0527 | 0.3331 ±0.0505 | **0.3488** ±0.0519 | 0.3309 ±0.0518 |

Table 4.4: *NDCG@10 of selected features (with confidence intervals) when experimenting four L2R algorithms as black-boxes. All results for WEB10K and YAHOO are related to two folds only, due to the time cost of executing Random Forest as a black-box. The symbol "RF" shows that the results are* ***statistically distinct*** *against the Random Forest execution.*

tures without tuning of leafs number. These results in the table show that a weak learner such as a Regression Tree and Linear Regression may be used in place of a state-of-art L2R as black-box methods in the wrapper-based FS without affecting the quality of the final selected individual and improving the execution time.

The good results of Regression Tree as a black-box when using Random Forest as a final learner may be expected by the fact that the Regression Tree algorithm is also a weak-learner inside the Random Forest. Hence, finding a set of features which improves the performance of the Regression Tree should also improve the Random Forest performance. This strengthens our argument for using weak learners as black-boxes in the FS process, mainly when the main goal is to improve the time performance of wrapper strategies. Moreover, our experiments show strong evidence that regression models (i.e. Linear Regression and Regression Tree) as black-boxes provide important feature selection for ensemble of tree (i.e. Random Forest and LambdaMART[10]). In future, we intend to further evaluate whether other distinct algorithms (e.g. neural net or probabilistic models) also provide advantage as black-boxes models.

Concerning the intuition of weak learner quality in evolutionary search, one can understand as its capability to discriminate among the individuals during the evolutionary process and not only to find the individual that obtain the maximum performance. By discriminate, we mean the capacity to not reduce the importance of bad features. This can be observed more clearly in Figures 4.3 and 4.4, which show the SPEA2 when using effectiveness as objective criterion, $\overset{E}{\succ}$. The figures describe effectiveness (in NDCG@10) on TD2003 and TD2004 datasets, respectively, using a Random Forest and a Linear Regression[11] as black-boxes. The box-plots in the figures summarize the effectiveness (NDCG@10) of individuals

[10]Next section describe the results when using Random Forest and LambdaMART to evaluate the selected individuals.

[11]We describe these experiments using Linear Regression, because its results provide a better description of our weak learner intuitions.

in the archives from different generations of our evolutionary process. We select archives A1, A2, and A3 to represent, respectively, the first, middle and last generations. The first three boxes in the Figure represent the performance using Random Forest, and the last three, using Linear Regression. As the evolutionary process receives synthetic individuals from Random Forest Feature Importance algorithm[12], we perceive in Figure 4.3 and 4.4 that both experiments begin with relative good individuals. However, only the quality of individuals when using Linear Regression improves during the evolutionary process, and Random Forest could not discriminate the set of features over the process.

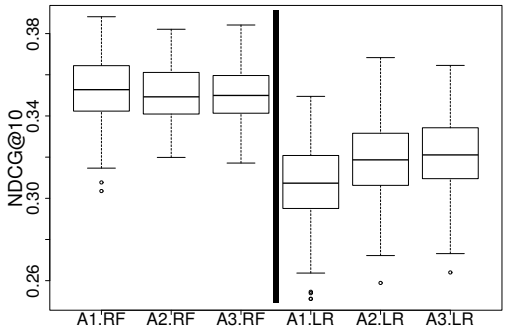



Figure 4.3: *The performance (NDCG@10) of SPEA2 using Random Forest (RF) and Linear Regression (LR) over generations for TD2003 dataset.*

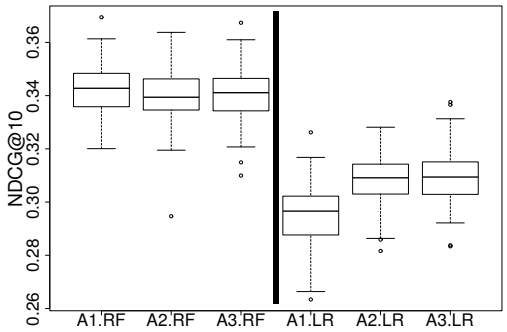



Figure 4.4: *The performance (NDCG@10) of SPEA2 using Random Forest (RF) and Linear Regression (LR) over generations for TD2004 dataset.*

In the figures, Random Forest average results are almost constant, while Linear Regression results show an increasing average curve. One possible reason is that a weak learner may not be able to attenuate the impact of noisy features as input, decreasing the overall effectiveness. We observe that there is higher variability in ranking performance when using

[12]For all experiments we select several rates of good features from the Random Forest Feature Importance algorithm, varying from 5%, 10%, 15% up to 95% of the best features.

Linear Regression, especially in Figure 4.3, showing that Linear Regression enables a more sensitive evaluation of the individuals.

To sum up and answering our Q2 research question, we show that weak learners, e.g. Linear Regression and Regression Tree, can be applied as a black-box in a wrapper strategy to perform FS on the L2R task. From now on, we perform and assess our wrapper evolutionary evaluation using Linear Regression and Regression Tree as black-boxes, assessing the selected individuals accuracy with Random Forest and LambdaMART algorithms.

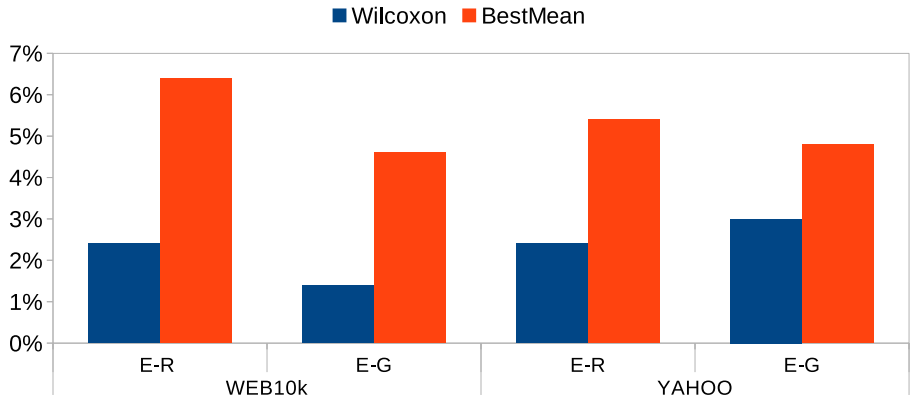
## 4.4 Evaluating the Paired Statistical Test for Pareto Set Selection

*Q3 – How to improve the selection of individuals inside of the Pareto frontier set, in order to provide a more effective subset of features?*

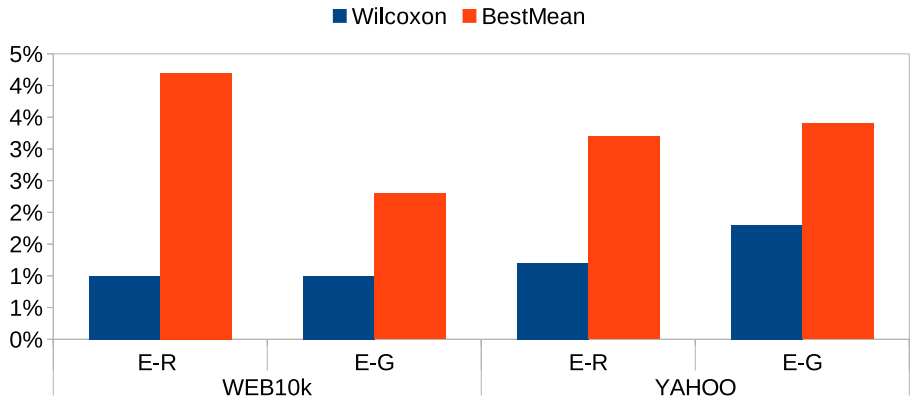
As a result of the SPEA2 process, a Pareto set is obtained with individuals that maximize all objective criteria. The size of this set can be large in cases where the objectives conflict with each other, increasing the difficulty of choosing only one individual. As described in Section 3.2.2, we deal with this issue by performing statistical tests to compare individuals in a multi-objective scenario, improving the ranking effectiveness and obtaining a smaller Pareto set.

To provide evidence for our claim, we begin by showing the reduced Pareto set due to statistical test comparisons, in Figure 4.5. Each column shows the percentage of Pareto set inside the last archive when applying $\overset{\text{E-R}}{\succ}$ and $\overset{\text{E-G}}{\succ}$ ($E-R$ and $E-G$, respectively) as objective criteria, and the weak learners such as Linear Regression and Regression Tree. The BestMean method performs a regular comparison with the best mean values of NDCG@10, without paired statistical tests, and the Wilcoxon (Wilcoxon Signed-rank test [Hsieh et al., 2008]) method corresponds to comparisons of the mean NDCG@10, but using a paired test to confirm when an individual is superior.[13] As expected, the experiment described in the figure allows us to observe that using statistical test comparison with multi-objective criteria provides a smaller Pareto set for all tests, through a more statistically strict comparison and breaking the tie relationship. We also note similar results when comparing both Wilcoxon and t-Test paired test methods. However, as Wilcoxon has been shown to be more robust against (or insensitive to) outliers [Hsieh et al., 2008], we use it in our experiments on the evolutionary processing, more specifically (in the training phase.).

[13]Reminding that the SPEA2 use all the archive and population to perform the cross-over and mutation process, as described in the Algorithm 1, and not only the Pareto set, described in Figure 4.5.




(a) Linear Regression as a black-box.




(b) Regression Tree as a black-box.

Figure 4.5: *Percentage of individuals remaining in the archive composing the Pareto Set in WEB10K and YAHOO datasets when using* $\overset{E\text{-}G}{\succ}$ *and* $\overset{E\text{-}R}{\succ}$ *objective criteria, Linear Regression and Regression Tree as weak-learners, and both method of fitness comparison: BestMean and Wilcoxon.*

In order to provide the real benefits of the smaller Pareto Set, we now show improvements on the effectiveness in Table 4.5, which describes the effectiveness (NDCG@10) of the final individual when a paired statistical test is applied to evaluate the fitness over the generations. The table describes the experimental results for BestMean, Wilcoxon and Wilcoxon-End, where BestMean and Wilcoxon were already explained, and Wilcoxon-End is our method to apply paired test only in the last generation. Our goal with Wilcoxon-End is to assess whether the statistical comparison in the last generation is enough to provide effective results. The results are assessed considering Linear Regression and Regression Tree as black-boxes. We apply in this experiment multi-objective and single-objective combinations, for instance, $\overset{E}{\succ}$, $\overset{E\text{-}R}{\succ}$, and $\overset{E\text{-}G}{\succ}$. Note that $\overset{E\text{-}R}{\succ}$ applies the statistical test in both objectives, effectiveness and $F_{RISK}$, as defined in 3.2.1.

As a very important result, Table 4.5 supports our claim that the statistical tests im-

prove the individual selection in the Pareto set. Considering only Wilcoxon and BestMean, the Wilcoxon method provides a better result for all multi-objective criteria, such as $\overset{\text{E-G}}{\succ}$ and $\overset{\text{E-R}}{\succ}$. Wilcoxon provides results statistically superior (the bold values) for all weak-learners, except for Regression Tree in WEB10K, which does not show statistically distinguishable values. Considering Wilcoxon and Wilcoxon-End methods, we can observe very close results, obtaining statistically distinguishable values only in three of eight multi-objective executions. Hence, showing that the statistical test can be applied only in the last generations, in order to improve the ranking of individuals and to reduce the Pareto set.

| | Linear Regression | | | Regression Tree | | |
|---|---|---|---|---|---|---|
| | WEB10K | | | | | |
| | $\overset{\text{E}}{\succ}$ | $\overset{\text{E-G}}{\succ}$ | $\overset{\text{E-R}}{\succ}$ | $\overset{\text{E}}{\succ}$ | $\overset{\text{E-G}}{\succ}$ | $\overset{\text{E-R}}{\succ}$ |
| Wilcoxon | 0.4212 | **0.4237**$^{b}$ | **0.4244**$^{be}$ | 0.4220 | 0.4238 | 0.4238 |
| BestMean | 0.4202 | 0.421 | 0.4205 | 0.4237 | 0.4232 | 0.4234 |
| Wilcoxon-End | 0.4201 | 0.4229 | 0.4205 | 0.4237 | 0.4237 | 0.4227 |
| | YAHOO | | | | | |
| | $\overset{\text{E}}{\succ}$ | $\overset{\text{E-G}}{\succ}$ | $\overset{\text{E-R}}{\succ}$ | $\overset{\text{E}}{\succ}$ | $\overset{\text{E-G}}{\succ}$ | $\overset{\text{E-R}}{\succ}$ |
| Wilcoxon | 0.7 | **0.7017**$^{b}$ | **0.7025**$^{be}$ | 0.6991 | **0.7019**$^{b}$ | **0.7027**$^{be}$ |
| BestMean | 0.7006 | 0.7005 | 0.7007 | 0.6994 | 0.6997 | 0.6994 |
| Wilcoxon-End | 0.7003 | 0.7022 | 0.7007 | 0.6994 | 0.7013 | 0.6994 |

Table 4.5: *Evaluating the statistical tests performance during the evolutionary search for WEB10k and YAHOO datasets. The letters b and e show* ***statistically difference*** *against BestMean and Wilconxon-End methods, respectively.*

Comparing $\overset{\text{E-G}}{\succ}$ and $\overset{\text{E-R}}{\succ}$ against $\overset{\text{E}}{\succ}$ in Table 4.5, we observe distinct results. As expected, the ranking effectiveness results for $\overset{\text{E}}{\succ}$ are not statistically distinguishable when the method varies over BestMean, Wilcoxon and Wilcoxon-End. In fact, as described in section 3.2.2, concerning a single-objective in an evolutionary process with a statistical test comparison, the ranking of the best individuals should not be very different in relation to the absence of the statistical test. On the other hand, in a multi-objective criteria, $\overset{\text{E-G}}{\succ}$ and $\overset{\text{E-R}}{\succ}$, the statistical test can break the conflict between two individuals, by equalizing two individuals for one objective and allowing a dominance relationship for the multi-objective comparison. This is a very interesting result, as it strengthens our assumptions that the statistical test avoids more tied comparison in a multi-objective scenario.

We have also evaluated the Wilcoxon, BestMean and Wilcoxon-End methods on TD2003 and TD2004 datasets, in Figure 4.6. However, as there are few queries on these datasets, we note that there is no further improvement when a statistical test is performed. Despite that, the max absolute values (the bold ones) vary over Wilcoxon and Wilcoxon-

End methods, which provide evidence of the importance of our statistical test comparison proposal over individuals when considering multi-objective criteria.

| | Linear Regression | | | Regression Tree | | |
|---|---|---|---|---|---|---|
| | TD2003 | | | | | |
| | $\succ_E$ | $\succ_{E\text{-}G}$ | $\succ_{E\text{-}R}$ | $\succ_E$ | $\succ_{E\text{-}G}$ | $\succ_{E\text{-}R}$ |
| Wilcoxon | 0.3272 | 0.3327 | **0.3586** | **0.3658** | 0.3325 | 0.3495 |
| BestMea | 0.3339 | 0.3368 | 0.35 | 0.3447 | 0.3394 | **0.3884** |
| Wilcoxon-End | **0.3453** | **0.3605** | 0.3582 | 0.3502 | **0.3506** | 0.3604 |
| | TD2004 | | | | | |
| Wilcoxon | **0.3336** | **0.346** | **0.3546** | **0.3488** | **0.3474** | 0.3443 |
| BestMea | 0.3248 | 0.3432 | 0.3544 | 0.34 | 0.3407 | 0.3597 |
| Wilcoxon-End | 0.3236 | 0.3428 | 0.3474 | 0.3449 | 0.338 | **0.3625** |

Table 4.6: *Evaluating the statistical tests performance during the evolutionary search for TD2003 and TD2004 datasets.*

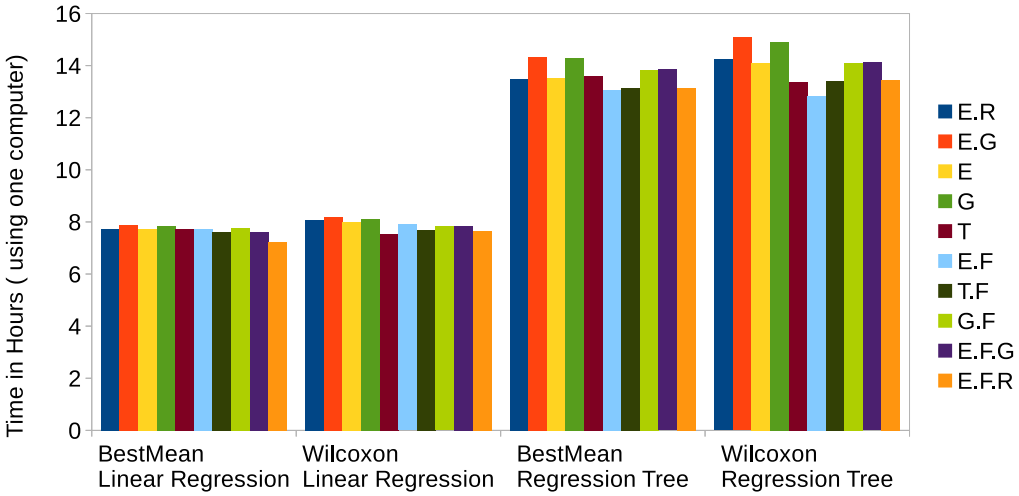



Figure 4.6: *The execution time (in hours) to process our individual comparison methods in the evolutionary algorithm when varying the objective criteria and weak learner as a black-box in WEB10K dataset.*

Regarding an efficiency evaluation, Figure 4.6 shows the time performance of BestMean and Wilcoxon methods applied with proposed multi-objective criteria in the WEB10K dataset. We do not show the Wilcoxon-End in the figure, because it has the similar time of BestMean. As Figure 4.6 shows, there is a little different when applying the statistical test in the SPEA2 process. Furthermore, Linear Regression is almost twice faster than Regression Tree. Although not described in the figure, 99% of processing time incurs due to L2R algorithm, processing the fitness values, which can change towards 98% (in average) if the paired test comparison is performed. This result confirms that SPEA2 has the time bounded

by the fitness processing (or the L2R method), and applying a paired test comparison does not increase significantly the processing time.

Answering our Q3 research question, a statistical paired comparison can improve the final selection of an individual in the Pareto set, considering a multi-objective criteria in FS for the L2R task.

## 4.5 A Multi-Objective FS Evaluation

*Q4 – What is the performance of risk-sensitiveness, effectiveness and feature reduction on the proposed objective criteria and methodology in FS for L2R?*

In this section we evaluated all proposed objective-criteria combinations for FS. First, we describe the general result in a summarized heatmap. Afterward, we present more details regarding risk-sensitiveness, ranking performance, and feature reduction.

The experimental results show that some of our objective-criteria for FS improve the risk-sensitiveness without decreasing the effectiveness, besides reducing the features space. This can be observed in Figure 4.7, which summarizes the effectiveness, risk-sensitiveness, and feature reduction performance when varying the objective criteria. The figure shows the number of statistical victories for each objective-criteria, which counts the victories of one objective-criteria against others, varying the datasets (WEB10K, YAHOO, TD2003 and TD2004), weak learners (Linear Regression and Regression Tree as black-boxes), and the BS4R (Max, Mean and BM25). In the figure, greater numbers mean more dark colors. In overall, $i \overset{\text{E-R}}{\succ} j$ and $i \overset{\text{E-G}}{\succ} j$ achieve a great victories number in risk-sensitiveness, but also ineffective performance, meanwhile obtaining a minor feature reduction. On the other hand, the effectiveness such as an objective-criterion, $i \overset{\text{E}}{\succ} j$, provided a larger feature reduction, however decreasing the effectiveness and risk-sensitiveness.

From Figure 4.7, we can observe that methods which applied a more drastically feature reduction could not improve the quality of effectiveness and risk-sensitiveness. On the other hand, the methods which focus in how to provide a more accurate model concerning to effectiveness and low-risk (e.g. $i \overset{\text{E-R}}{\succ} j$, $i \overset{\text{E-G}}{\succ} j$), could obtain a very interesting effective and robust performance, even when performing Feature Selection. Note that these methods are contrary to FS methods available in the literature, which the goal is to drastically reduce the amount of features in order to control the noise, redundancy and processing time.

Moreover, from Figure 4.7 we can see that using only ranking effectiveness as an objective, i.e, $\overset{\text{E}}{\succ}$, it is not possible to produce robust models, as empirically demonstrated

| | Effectiveness | Risk-sensitiveness | Feature Reduction |
|---|---|---|---|
| $\overset{\text{E-R}}{\succ}$ | 73 | 75 | 3 |
| $\overset{\text{E-G}}{\succ}$ | 62 | 69 | 13 |
| $\overset{\text{G}}{\succ}$ | 62 | 63 | 8 |
| $\overset{\text{E}}{\succ}$ | 54 | 56 | 27 |
| $\overset{\text{E-F-G}}{\succ}$ | 49 | 54 | 41 |
| $\overset{\text{T}}{\succ}$ | 46 | 53 | 32 |
| $\overset{\text{G-F}}{\succ}$ | 45 | 50 | 42 |
| DivFS | 27 | 35 | 55 |
| BTFS | 21 | 24 | 84 |
| $\overset{\text{T-F}}{\succ}$ | 10 | 9 | 71 |
| $\overset{\text{E-F-R}}{\succ}$ | 7 | 10 | 68 |
| $\overset{\text{E-F}}{\succ}$ | 4 | 7 | 74 |

Table 4.7: *Heatmap of our results for FS over effectiveness, risk-sensitiveness, and feature reduction.*

in most of the collections used in our experiments. In contrast, only by introducing a risk-sensitive measure one can produce more low-risk model.

In next sections we evaluate with more details our proposed objective criteria, in order to thoroughly describe our outcomes.

## 4.5.1 Risk-Sensitiveness Evaluation

### 4.5.1.1 Evaluation on WEB10K and YAHOO Datasets

We start evaluating the risk-sensitive performance of our proposed single and multi-objective criteria and our FS methodology in WEB10K and YAHOO datasets. Tables 4.8[14] and 4.9 present the experimental results for the WEB10K over selected sets of features trained with Random Forest. The first table shows results for the evolutionary algorithm with Linear Regression as the L2R black-box, while the second table contains results for Regression Trees. Each table is separated vertically by a line. Results at the right of this line correspond to methods that have feature reduction as one optimization objective, whereas results at left do not optimize this criterion.

For WEB10K, $\overset{\text{E-R}}{\succ}$ and $\overset{\text{E-G}}{\succ}$ are the only FS methods capable of consistently keeping their risk-sensitiveness evaluation for both Linear Regression (Table 4.8) and Regression

[14]For readability reasons, we did not include the confidence intervals from now on.

| WEB10K - Back-box: Linear Regression | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Full | E-R $\succ$ | E-G $\succ$ | E $\succ$ | G $\succ$ | T $\succ$ | E-F $\succ$ | T-F $\succ$ | G-F $\succ$ | E-F-G $\succ$ | E-F-R $\succ$ | DivFT | BTFS |
| $G_{RISK}$ | 0.419 | **0.4182** | 0.4177 | 0.4159 | 0.4167 | 0.4164 | 0.4129 | 0.4148 | 0.4161 | 0.416 | 0.4134 | 0.4134 | 0.4122 |
| Max BS4R | | | | | | | | | | | | | |
| $T_{RISK}$ | -94.7$^{e}$ | **-94.4**$^{e}$ | -94.6$^{e}$ | -95.8$^{f}$ | -95.2$^{fe}$ | -95.4$^{f}$ | -97.4$^{fe}$ | -96.3$^{fe}$ | -95.4$^{f}$ | -95.8$^{f}$ | -96.4$^{fe}$ | -96.4$^{fe}$ | -97.0$^{fe}$ |
| Win | 1758 | 1752 | **1774** | 1745 | 1708 | 1734 | 1640 | 1698 | 1720 | 1705 | 1683 | 1680 | 1674 |
| L>20%↓ | 5097 | **5097** | 5129 | 5201 | 5157 | 5180 | 5297 | 5248 | 5166 | 5208 | 5234 | 5266 | 5289 |
| Mean BS4R | | | | | | | | | | | | | |
| $T_{RISK}$ | 50.5$^{e}$ | **49.6**$^{fe}$ | 48.5$^{fe}$ | 47.3$^{f}$ | 48.3$^{f}$ | 48.3$^{f}$ | 45.8$^{fe}$ | 46.2$^{f}$ | 48.2$^{f}$ | 47.8$^{f}$ | 45.1$^{fe}$ | 43.8$^{fe}$ | 44.9$^{fe}$ |
| Win | 8402 | **8401** | 8389 | 8373 | 8377 | 8368 | 8339 | 8364 | 8381 | 8377 | 8343 | 8327 | 8316 |
| L>20%↓ | 768 | **792** | 803 | 813 | 810 | **792** | 826 | 806 | 804 | 800 | 847 | 839 | 835 |
| BM25 BS4R | | | | | | | | | | | | | |
| $T_{RISK}$ | 5.6$^{e}$ | **5.4**$^{e}$ | 4.3$^{fe}$ | 3.7$^{f}$ | 4.3$^{f}$ | 3.9$^{f}$ | 2$^{fe}$ | 2.4$^{fe}$ | 4.1$^{f}$ | 3.5$^{f}$ | 1.8$^{fe}$ | 1.8$^{fe}$ | 2.3$^{fe}$ |
| Win | 7287 | **7261** | 7254 | 7215 | 7228 | 7214 | 7134 | 7172 | 7233 | 7241 | 7193 | 7175 | 7157 |
| L>20%↓ | 1081 | **1127** | 1163 | 1173 | 1138 | 1147 | 1201 | 1193 | 1146 | 1143 | 1201 | 1189 | 1188 |

Table 4.8: *The risk-sensitive evaluation in WEB10K dataset, using the RF on selected features and the Linear Regression as a Black-Box. Bold represents the best values among FS methods. The superscript letters e and f appearing in results for $T_{RISK}$ represent results statistically distinguishable with the $\overset{E}{\succ}$ objective and the Full set of features, respectively.*

Tree (Table 4.9) similar to the "Full Set of Features"[15]. Also, in both tables, results for the $\overset{\text{E-R}}{\succ}$ and the $\overset{\text{E-G}}{\succ}$ objectives have the highest values for $T_{RISK}$ and $G_{RISK}$ among all the FS methods in all tests, and also the best results in almost all other risk-sensitive measures.

| WEB10K - Black-box: Regression Tree | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Full | E-R $\succ$ | E-G $\succ$ | E $\succ$ | G $\succ$ | T $\succ$ | E-F $\succ$ | T-F $\succ$ | G-F $\succ$ | E-F-G $\succ$ | E-F-R $\succ$ | DivFS | BTFS |
| $G_{RISK}$ | 0.419 | 0.417 | **0.418** | 0.411 | 0.417 | 0.415 | 0.41 | 0.41 | 0.417 | 0.417 | 0.405 | 0.413 | 0.412 |
| Max BS4R | | | | | | | | | | | | | |
| $T_{RISK}$ | -94.7$^{e}$ | -94.7 | **-94.6**$^{e}$ | -94.8$^{f}$ | -94.7$^{f}$ | -96.0$^{fe}$ | -98.9$^{fe}$ | -99.1$^{fe}$ | **-94.6**$^{f}$ | -95.1$^{f}$ | -100.4$^{fe}$ | -96.4$^{fe}$ | -97.0$^{fe}$ |
| Win | 1758 | 1747 | **1755** | 1745 | 1739 | 1714 | 1580 | 1536 | 1750 | **1755** | 1460 | 1680 | 1674 |
| L>20%↓ | 5097 | 5122 | **5116** | 5130 | 5154 | 5235 | 5460 | 5466 | 5137 | 5154 | 5508 | 5266 | 5289 |
| Mean BS4R | | | | | | | | | | | | | |
| $T_{RISK}$ | 50.5$^{e}$ | 48.4$^{f}$ | **48.9**$^{fe}$ | 48.1$^{f}$ | 48.4$^{f}$ | 46.7$^{fe}$ | 41.3$^{fe}$ | 41.6$^{fe}$ | 48.4$^{f}$ | 48.8$^{f}$ | 39.3$^{fe}$ | 43.8$^{fe}$ | 44.9$^{fe}$ |
| Win | 8402 | 8387 | **8415** | 8378 | 8385 | 8379 | 8254 | 8243 | 8375 | 8404 | 8216 | 8327 | 8316 |
| L>20%↓ | 768 | 786 | **771** | 799 | 795 | 793 | 872 | 858 | 794 | 795 | 891 | 839 | 835 |
| BM25 BS4R | | | | | | | | | | | | | |
| $T_{RISK}$ | 5.6$^{e}$ | **4.8**$^{fe}$ | **4.8**$^{fe}$ | 3.9$^{f}$ | 4.6$^{f}$ | 2.7$^{fe}$ | -2.4$^{fe}$ | -2.4!$^{fe}$ | 4.7$^{fe}$ | **4.8**$^{fe}$ | -3.4$^{fe}$ | 1.5$^{fe}$ | 2.1$^{fe}$ |
| Win | 7287 | 7251 | 7265 | 7246 | 7251 | 7199 | 7109 | 7067 | **7275** | 7258 | 7076 | 7175 | 7157 |
| L>20%↓ | 1081 | 1131 | 1128 | 1156 | 1140 | 1166 | 1297 | 1273 | 1122 | **1119** | 1308 | 1189 | 1188 |

Table 4.9: *The risk-sensitive evaluation in WEB10K dataset, using the RF on selected features and the Long Regression Tree as a Black-Box. Bold represents the best values among FS methods. The superscript letters e and f appearing in results for $T_{RISK}$ represent results statistically distinguishable with the $\overset{E}{\succ}$ objective and the Full set of features, respectively.*

Almost all $\overset{\text{E-R}}{\succ}$ and $\overset{\text{E-G}}{\succ}$ results are statistically superior to those for the $\overset{\text{E}}{\succ}$ objective criterion, which confirms our hypothesis that optimizing only effectiveness is not sufficient to have a good risk-sensitive performance. On the other hand, results for $\overset{\text{E-R}}{\succ}$ and $\overset{\text{E-G}}{\succ}$ are not inferior to those methods having risk-sensitiveness as the only objective to be optimized

[15] Most differences between the all-features method and $\overset{\text{E-R}}{\succ}$ and $\overset{\text{E-G}}{\succ}$ are not statistically significant

($\overset{G}{\succ}$ and $\overset{T}{\succ}$). This means that the multi-objective criteria *effectiveness* × *risk-sensitiveness* ($\overset{E\text{-}R}{\succ}$ and $\overset{E\text{-}G}{\succ}$) presented results comparable or even superior to those of methods trying to optimize each isolated objective. In fact, we note that by using multi-objective criteria, we are improving the evaluation of individuals through the evolutionary search, being more rigorous in evaluation of each features set.

It is also important to highlight that the methods that have feature reduction as an objective (even those also trying to optimize risk-sensitiveness) do not perform consistently better than $\overset{E\text{-}R}{\succ}$ and $\overset{E\text{-}G}{\succ}$. This confirms our initial claim that methods that optimize feature reduction may sacrifice risk-effectiveness.

Results for WEB10K dataset are especially important because this dataset is usually considered a reference in L2R. Therefore, results in WEB10K results confirm that it is feasible to find a robust solution by exploiting the $\overset{E\text{-}R}{\succ}$ or $\overset{E\text{-}G}{\succ}$ objectives criteria.

Now we turn our attention to Tables 4.10 and 4.11 that present results for the YAHOO dataset[16]. The method $\overset{E\text{-}R}{\succ}$ performed consistently better than all other FS methods, considering the two used black-boxes. However, method $\overset{G}{\succ}$ performed slightly better than $\overset{E\text{-}G}{\succ}$ for this collection, winning in some measures for some BS4R and tying up in others. Nevertheless, both $\overset{E\text{-}R}{\succ}$ and $\overset{E\text{-}G}{\succ}$ are among the three best FS methods for the YAHOO dataset.

| YAHOO - Black-box: Linear Regression | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Full | E-R $\succ$ | E-G $\succ$ | E $\succ$ | G $\succ$ | T $\succ$ | E-F $\succ$ | T-F $\succ$ | G-F $\succ$ | E-F-G $\succ$ | E-F-R $\succ$ | DivFS | BTFS |
| $G_{RISK}$ | 0.5489 | **0.5494** | 0.5491 | 0.5482 | **0.5494** | 0.5477 | 0.5454 | 0.5454 | 0.5477 | 0.5483 | 0.5463 | 0.5442 | 0.5469 |
| MAX BS4R | | | | | | | | | | | | | |
| $T_{RISK}$ | -49.6$^{e}$ | **-49.6**$^{e}$ | -49.8$^{e}$ | -50.0$^{f}$ | **-49.6**$^{e}$ | -50.1$^{f}$ | -50.4$^{fe}$ | -50.4$^{fe}$ | -50.0$^{f}$ | -50.0$^{f}$ | -50.4$^{fe}$ | -51.0$^{fe}$ | -50.2$^{f}$ |
| Win | 2173 | **2178** | 2160 | 2138 | **2178** | 2143 | 2144 | 2144 | 2150 | 2160 | 2132 | 2092 | 2122 |
| L>20%↓ | 2006 | **2020** | 2029 | 2027 | 2023 | 2026 | 2082 | 2082 | 2021 | 2034 | 2069 | 2073 | 2073 |
| Mean BS4R | | | | | | | | | | | | | |
| $T_{RISK}$ | -3.0$^{e}$ | **-3.0**$^{e}$ | -3.2$^{f}$ | -3.5$^{f}$ | **-3.0**$^{e}$ | -3.5$^{f}$ | -4.2$^{fe}$ | -4.2$^{fe}$ | -3.4$^{f}$ | -3.3$^{f}$ | -3.7$^{f}$ | -4.6$^{fe}$ | -3.6$^{f}$ |
| Win | 4901 | 4900 | 4902 | 4888 | **4909** | 4902 | 4860 | 4860 | 4905 | 4886 | 4872 | 4853 | 4890 |
| L>20%↓ | 778 | **784** | 785 | 792 | 785 | 800 | 807 | 807 | 793 | 795 | 805 | 820 | 795 |

Table 4.10: *The risk-sensitive evaluation in YAHOO dataset, using the RF on selected features and the Linear Regression as a Black-Box. As there is no public description of the features in YAHOO dataset, this table does not contain the BM25 BS4R.*

Furthermore, $\overset{E\text{-}R}{\succ}$, $\overset{E\text{-}G}{\succ}$, and $\overset{G}{\succ}$ methods are statistically superior to $\overset{E}{\succ}$ in almost all risk-sensitive measures and performed better than any method which optimizes feature reduction. For YAHOO, effectiveness as an objective alone is not sufficient to drive the search towards low-risk models. In addition, having a reduction of features as one of the objectives is not consistently good to derive risk-sensitive models.

[16]As there is no public description of the features in YAHOO dataset, Table 4.10 and 4.11 do not not contain the BM25 BS4R.

Besides that, taking only $\overset{G}{\succ}$, $\overset{G\text{-}F}{\succ}$ and $\overset{E\text{-}F\text{-}G}{\succ}$ methods in Tables 4.8 to 4.11, one can observe a robust behavior in our experiments, mainly when compared against the $\overset{T}{\succ}$, $\overset{T\text{-}F}{\succ}$ objective criteria. This clearly shows that $G_{RISK}$ explores a more robust search space, and it does this better than $T_{RISK}$. This behavior is explained by the strength of $G_{RISK}$, which captures the risk measure comparing the model to the shape of the score distribution with many BS4Rs.

More specifically, $T_{RISK}$ as a risk-sensitive objective criterion evaluates the individuals considering a static BS4R over the generations, only the full set of features is used as an immutable robust model. By using the $G_{RISK}$ this limitation is overcame. As all individuals of each generation are used as BS4R, the $G_{RISK}$ objective criterion provides a coevolutionary search, varying the fitness score for the same individual over the time and the fitness score is dependent of the individuals population. As as result, $G_{RISK}$ evaluates better the individuals of a population regarding risk-sensitiveness, as the BS4R change over the generation and the search adapts itself for better solutions.

The risk-sensitive results of the baselines methods (DivFS and BTFS) are among the worst in the WEB10K dataset. Nevertheless, for the YAHOO dataset this picture changes towards relatively better results for BTFS, when comparing to our multi-objective models that have feature reduction as one of the objectives to be optimized. In addition, most of these FS baselines are statistically inferior when compared to $\overset{E\text{-}F}{\succ}$ and $\overset{E\text{-}G}{\succ}$. This is to be expected as these FS approaches usually attempt to obtain the best effectiveness and feature reduction, disregarding risk. Additionally, as the number of features is an objective criterion in these methods, their results are similar to our evolutionary methods which include the number of features as an objective criterion.

| YAHOO - Black-box: Regression Tree | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Full | E-R $\succ$ | E-G $\succ$ | E $\succ$ | G $\succ$ | T $\succ$ | E-F $\succ$ | T-F $\succ$ | G-F $\succ$ | E-F-G $\succ$ | E-F-R $\succ$ | DivFS | BTFS |
| $G_{RISK}$ | 0.549 | **0.549** | **0.549** | 0.5472 | 0.5485 | 0.5444 | 0.544 | 0.544 | 0.5477 | 0.548 | 0.544 | 0.5442 | 0.5469 |
| Max BS4R | | | | | | | | | | | | | |
| $T_{RISK}$ | -49.6$^{e}$ | **-49.6**$^{e}$ | -49.8$^{e}$ | -50.4$^{f}$ | -49.8$^{fe}$ | -51.0$^{fe}$ | -51.2$^{fe}$ | -51.2$^{fe}$ | -50.0$^{f}$ | -49.9$^{fe}$ | -51.2$^{fe}$ | -51.0$^{fe}$ | -50.2$^{f}$ |
| Win | 2173 | **2184** | 2147 | 2108 | 2168 | 2088 | 2090 | 2090 | 2136 | 2152 | 2090 | 2092 | 2122 |
| L>20%↓ | 2006 | 2020 | 2041 | 2057 | **2017** | 2092 | 2100 | 2100 | 2048 | 2042 | 2100 | 2073 | 2073 |
| Mean BS4R | | | | | | | | | | | | | |
| $T_{RISK}$ | -3.0$^{e}$ | **-3.0**$^{e}$ | **-3.0**$^{e}$ | -3.7$^{f}$ | -3.2$^{e}$ | -4.2$^{fe}$ | -4.4$^{fe}$ | -4.4$^{fe}$ | -3.5$^{f}$ | -3.4$^{f}$ | -4.4$^{fe}$ | -4.6$^{fe}$ | -3.6$^{f}$ |
| Win | 4901 | 4906 | 4899 | 4876 | 4887 | 4863 | 4856 | 4856 | **4913** | 4903 | 4856 | 4853 | 4890 |
| L>20%↓ | 778 | **785** | 789 | 796 | 792 | 794 | 824 | 824 | 797 | 799 | 824 | 820 | 795 |

Table 4.11: *The risk-sensitive evaluation in YAHOO dataset, using the RF on selected features and the Regression Tree as a Black-Box. As there is no public description of the features in YAHOO dataset, this table does not contain the BM25 BS4R.*

#### 4.5.1.2 Evaluation on TD2003 and TD2004 Datasets

For completeness, we have also evaluated all methods with the TD2003 and TD2004 datasets, or TDs. Results for both collections with the two black-boxes (Linear Regression and Regression Tree) are presented in Tables 4.12 to 4.15. Since these collections are very small, it is hard to obtain consistent (i.e., statistically significant) results for them. Indeed, Gomes et al. [2013] has already demonstrated that it is hard to obtain statistical significance in both datasets because of their low number of queries. However, our goal with these datasets is to highlight the tendencies of ranking effectiveness and risk-sensitive evaluation.

| TD2003 - Black-Box: Linear Regression | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Full | $\overset{\text{E-R}}{\succ}$ | $\overset{\text{E-G}}{\succ}$ | $\overset{\text{E}}{\succ}$ | $\overset{\text{G}}{\succ}$ | $\overset{\text{T}}{\succ}$ | $\overset{\text{E-F}}{\succ}$ | $\overset{\text{T-F}}{\succ}$ | $\overset{\text{G-F}}{\succ}$ | $\overset{\text{E-F-G}}{\succ}$ | $\overset{\text{E-F-R}}{\succ}$ | DivFT | BTFS |
| $G_{RISK}$ | 0.379 | **0.373**$^{e}$ | 0.339 | 0.331 | 0.352 | 0.354 | 0.225 | 0.217 | 0.362 | 0.3598 | 0.2712$^{e}$ | 0.358 | 0.286$^{e}$ |
| Max BS4R | | | | | | | | | | | | | |
| $T_{RISK}$ | -6.3$^{e}$ | **-5.9**$^{e}$ | **-5.9**$^{e}$ | -6.5$^{f}$ | -6.1$^{e}$ | -7.1$^{f}$ | -8.4$^{fe}$ | -8.4$^{fe}$ | -7.1$^{e}$ | -6.2$^{e}$ | -8.2$^{fe}$ | -6.9$^{f}$ | -7.6$^{fe}$ |
| Win | 9 | **9** | **9** | 7 | 6 | 6 | 0 | 1 | 5 | 4 | 2 | 8 | 2 |
| L>20%↓ | 28 | **27** | 29 | 32 | 28 | 32 | 38 | 40 | 30 | 30 | 37 | 32 | 38 |
| Mean BS4R | | | | | | | | | | | | | |
| $T_{RISK}$ | 5.6$^{e}$ | **5.3**$^{e}$ | 3.7$^{f}$ | 3.7$^{f}$ | 4.0 | 5.3 | -1.5$^{fe}$ | -1.6$^{fe}$ | 5.2 | **5.3**$^{e}$ | 0.3$^{fe}$ | **5.3** | 0.3$^{fe}$ |
| Win | 41 | **40** | 39 | 37 | 38 | 38 | 27 | 25 | 39 | **40** | 30 | 39 | 34 |
| L>20%↓ | 5 | 8 | 7 | 10 | 9 | 7 | 18 | 20 | 7 | 6 | 14 | **4** | 12 |
| BM25 BS4R | | | | | | | | | | | | | |
| $T_{RISK}$ | 1.3$^{e}$ | 1.4$^{e}$ | -0.2$^{f}$ | -0.13$^{f}$ | 1.8$^{e}$ | 0.2$^{f}$ | -2.7$^{fe}$ | -2.7$^{fe}$ | 0.2 | **1.9**$^{e}$ | -1.5$^{fe}$ | -0.2$^{f}$ | -1.6$^{f}$ |
| Win | 34 | **34** | 31 | 28 | 30 | 30 | 20 | 19 | 32 | 34 | 23 | 32 | 26 |
| L>20%↓ | 7 | **6** | 9 | 8 | **6** | 8 | 16 | 15 | 8 | **6** | 11 | 9 | 12 |

Table 4.12: *The risk-sensitive evaluation in TD2003 dataset, using the RF on selected features and the Linear Regression as a Black-Box.*

| TD2003 - Black-Box: Regression Tree | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Full | $\overset{\text{E-R}}{\succ}$ | $\overset{\text{E-G}}{\succ}$ | $\overset{\text{E}}{\succ}$ | $\overset{\text{G}}{\succ}$ | $\overset{\text{T}}{\succ}$ | $\overset{\text{E-F}}{\succ}$ | $\overset{\text{T-F}}{\succ}$ | $\overset{\text{G-F}}{\succ}$ | $\overset{\text{E-F-G}}{\succ}$ | $\overset{\text{E-F-R}}{\succ}$ | DivFT | BTFS |
| $G_{RISK}$ | 0.380 | 0.362 | 0.367 | 0.364 | **0.380** | 0.352 | 0.230 | 0.230 | 0.310 | 0.336 | 0.250 | 0.358 | 0.286 |
| MAX BS4R | | | | | | | | | | | | | |
| $T_{RISK}$ | -6.3 | -6 | -6.7 | **-5.8** | **-5.8** | -6.3 | -9.0$^{fe}$ | -9.0$^{fe}$ | -6.7$^{fe}$ | -6.6$^{fe}$ | -8.5$^{fe}$ | -6.9 | -7.6$^{fe}$ |
| Win | 9 | **10** | 6 | 9 | 6 | 7 | 0 | 0 | 5 | 6 | 2 | 8 | 2 |
| L>20%↓ | 28 | 27 | 33 | 28 | **26** | 29 | 40 | 40 | 33 | 32 | 40 | 29 | 38 |
| MEAN BS4R | | | | | | | | | | | | | |
| $T_{RISK}$ | 5.6 | 4.6 | **6.6** | 4.6 | 5.6 | 4.3 | -1.8$^{fe}$ | -1.8$^{fe}$ | 2.1$^{fe}$ | 2.4 | -1.7$^{fe}$ | 5.4 | 0.3$^{fe}$ |
| Win | 41 | 38 | **43** | 39 | 39 | 40 | 26 | 26 | 34 | 38 | 27 | 39 | 34 |
| L>20%↓ | 5 | 8 | **4** | 8 | 7 | 7 | 20 | 20 | 9 | 9 | 20 | **4** | 12 |
| BM25 BS4R | | | | | | | | | | | | | |
| $T_{RISK}$ | 1.3 | 0.4 | 1.1 | 1.2 | **1.3** | 1.1 | -2.7$^{fe}$ | -2.7$^{fe}$ | -1.1$^{f}$ | -0.9 | -2.6$^{fe}$ | -0.2$^{f}$ | -1.6$^{fe}$ |
| Win | 34 | 32 | 31 | 31 | **33** | **33** | 19 | 19 | 27 | 31 | 20 | 32 | 26 |
| L>20%↓ | 7 | 8 | 6 | 6 | **5** | 6 | 16 | 16 | 10 | 9 | 15 | 9 | 12 |

Table 4.13: *The risk-sensitive evaluation in TD2003 dataset, using the RF on selected features and the Regression Tree as a Black-Box.*

Concerning the TDs experiments, when using Linear Regression as black-box the outcomes for methods $\overset{\text{E-R}}{\succ}$ and $\overset{\text{E-G}}{\succ}$ in TD2003 and TD2004 are among the best methods. When

| TD2004 - Black-Box: Linear Regression | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Full | E-R $\succ$ | E-G $\succ$ | E $\succ$ | G $\succ$ | T $\succ$ | E-F $\succ$ | T-F $\succ$ | G-F $\succ$ | E-F-G $\succ$ | E-F-R $\succ$ | DivFS | BTFS |
| $G_{RISK}$ | 0.364 | **0.364** | 0.354 | 0.346 | 0.343 | 0.361 | 0.298 | 0.338 | 0.354 | 0.346 | 0.297 | 0.346 | 0.334 |
| MAX BS4R | | | | | | | | | | | | | |
| $T_{RISK}$ | -9 | -8.7 | -8.8 | -9.4 | -9.9 | **-8.5** | -10.7$^{fe}$ | -10.6$^{f}$ | -8.72 | -9.7 | -10.9$^{fe}$ | -9.2 | -9.5$^{f}$ |
| Win | 14 | 14 | **16** | 9 | 11 | 14 | 5 | 7 | 14 | 9 | 7 | 13 | 11 |
| L>20%↓ | 47 | **45** | 48 | 50 | 51 | 45 | 55 | 52 | 47 | 48 | 56 | 51 | 51 |
| MEAN BS4R | | | | | | | | | | | | | |
| $T_{RISK}$ | 6.7 | **7.3**$^{e}$ | 6.4 | 4.6 | 6.2 | 5 | 0.3$^{fe}$ | 5.2 | 6.2 | 5.8 | 0.56$^{fe}$ | 3.9 | 4.9$^{f}$ |
| Win | 63 | **65** | 62 | 60 | 59 | 63 | 52 | 61 | 64 | 62 | 51 | 62 | 60 |
| L>20%↓ | 8 | **8** | 11 | 11 | 13 | 10 | 20 | 12 | 10 | 12 | 22 | 11 | 9 |
| BM25 BS4R | | | | | | | | | | | | | |
| $T_{RISK}$ | 0.1 | **0.3** | -0.1 | -0.6 | 0 | -0.2 | -2.3$^{fe}$ | -1.3$^{f}$ | **0.3** | 0.01 | -2.3$^{fe}$ | -1.0$^{f}$ | -0.8$^{f}$ |
| Win | 58 | **57** | 53 | 52 | 53 | 55 | 41 | 44 | 55 | 50 | 41 | 50 | 55 |
| L>20%↓ | 10 | 9 | 11 | 12 | 10 | 11 | 20 | 15 | **7** | 8 | 19 | 11 | 12 |

Table 4.14: *The risk-sensitive evaluation in TD2004 dataset, using the RF on selected features and the Linear Regression as a Black-Box.*

| TD2004 - Black-Box: Regression Tree | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Full | E-R $\succ$ | E-G $\succ$ | E $\succ$ | G $\succ$ | T $\succ$ | E-F $\succ$ | T-F $\succ$ | G-F $\succ$ | E-F-G $\succ$ | E-F-R $\succ$ | DivFS | BTFS |
| $G_{RISK}$ | 0.364 | 0.351 | 0.353 | 0.363 | **0.366** | 0.362 | 0.314 | 0.314 | 0.349 | 0.354 | 0.285 | 0.346 | 0.334 |
| MAX BS4R | | | | | | | | | | | | | |
| $T_{RISK}$ | -8.9 | -8.9 | **-8.7** | **-8.7** | -9.1 | -8.9 | -10.0!$^{e}$ | -10.0$^{fe}$ | -9.5 | -9.5 | -10.0$^{fe}$ | -9.2$^{fe}$ | -9.5$^{fe}$ |
| Win | 14 | 12 | 18 | **16** | 13 | 14 | 8 | 8 | 9 | 10 | 6 | 13 | 11 |
| L>20%↓ | 47 | 49 | 47 | 47 | 48 | **46** | 52 | 52 | 51 | 50 | 54 | 51 | 51 |
| MEAN BS4R | | | | | | | | | | | | | |
| $T_{RISK}$ | 6.7 | **6.3** | 5.3 | 5.3 | 8.2 | 6.3 | 1.6$^{fe}$ | 1.6$^{fe}$ | 6.1 | 6 | 0.2$^{fe}$ | 3.9 | 4.9$^{f}$ |
| Win | 63 | 61 | 61 | 63 | **66** | 62 | 55 | 55 | 61 | 60 | 50 | 62 | 60 |
| L>20%↓ | 8 | 11 | 11 | 10 | **7** | 9 | 17 | 17 | 9 | 10 | 20 | 11 | 9 |
| BM25 BS4R | | | | | | | | | | | | | |
| $T_{RISK}$ | 0.1 | -0.2 | -0.5 | -0.2 | **0.1** | -0.5 | -1.8$^{fe}$ | -1.8$^{fe}$ | -0.4 | -0.5 | -2.4$^{fe}$ | -1.0$^{fe}$ | -0.8$^{f}$ |
| Win | 58 | 52 | 54 | 52 | **57** | 53 | 46 | 46 | 51 | 50 | 41 | 50 | 55 |
| L>20%↓ | 10 | 10 | 12 | 11 | **8** | 13 | 15 | 15 | 11 | 12 | 19 | 11 | 12 |

Table 4.15: *The risk-sensitive evaluation in TD2004 dataset, using the RF on selected features and the Regression Tree as a Black-Box.*

using the Regression Tree as black-box, the experimental results for TD2003 show a better performance for $\overset{G}{\succ}$. In case of TD2004 using Regression Tree the results are not clear at all. Even with this lack of clarity for TD2004, one can observe that the methods that include the number of features as an objective criterion were outperformed by ones that do not include it, which was also assessed in TD2003.

Regarding the full set of features ("Full") comparison for TDs, we notice that the most of methods that include risk-sensitiveness as a criterion are statistically similar to "Full". However, in some cases $\overset{E\text{-}R}{\succ}$ outperforms "Full" with absolute values, such as for TD204 in Table 4.14 for risk-sensitive measures and for TD2003 dataset in Table 4.13.

We observe that there is a higher variance of results when changing the black-boxes for TD2003 and TD2004 datasets. In fact, when using Regression Tree as black-box there

is no consistency regarding the objective criteria which give the best risk-sensitiveness. One reason for that is the overfitting behavior of the Regression Tree without pruning, combined with the few available data in these datasets. This can limit the quality of the evolutionary search for some methods. However, one can observe a tendency of $\overset{\text{E-R}}{\succ}$, $\overset{\text{E-G}}{\succ}$, and $\overset{\text{G}}{\succ}$ towards better results. Again, giving evidence to our claim that using a risk measure as an objective criterion may be useful to avoid poor results in some queries.

### 4.5.2 Effectiveness Evaluation

Table 4.16 shows that in addition to risk-sensitiveness the objectives that optimize both effectiveness and risk-sensitiveness are also capable of obtaining good ranking performance, i.e. rank effectiveness. As we can see, the $\overset{\text{E-R}}{\succ}$ multi-objective method is the only one that obtained the best results against all evaluated FS methods, considering all datasets and black-boxes. In the majority of the results, the $\overset{\text{E-R}}{\succ}$ performance is statistically similar against using all features, except for WEB10K when using Regression Trees as a black-box method.

| Effectiveness Evaluation (NDCG@10) | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Full | $\overset{\text{E-R}}{\succ}$ | $\overset{\text{E-G}}{\succ}$ | $\overset{\text{E}}{\succ}$ | $\overset{\text{G}}{\succ}$ | $\overset{\text{T}}{\succ}$ | $\overset{\text{E-F}}{\succ}$ | $\overset{\text{T-F}}{\succ}$ | $\overset{\text{G-F}}{\succ}$ | $\overset{\text{E-F-G}}{\succ}$ | $\overset{\text{E-F-R}}{\succ}$ | DivFT | BTFS |
| Black-Box: Linear Regression | | | | | | | | | | | | | |
| WEB10K | $0.424^{e}$ | $\mathbf{0.424}^{e}$ | $0.424^{fe}$ | $0.421^{f}$ | $0.422^{f}$ | $0.422^{f}$ | $0.417^{fe}$ | $0.419^{fe}$ | $0.422^{f}$ | $0.421^{f}$ | $0.418^{fe}$ | $0.417^{fe}$ | $0.418^{fe}$ |
| YAHOO | $0.703^{e}$ | $\mathbf{0.703}^{e}$ | $0.702^{fe}$ | $0.700^{f}$ | $0.703^{e}$ | $0.700^{f}$ | $0.698^{fe}$ | $0.698^{fe}$ | $0.700^{f}$ | $0.701^{f}$ | $0.698^{fe}$ | $0.695^{fe}$ | $0.699^{f}$ |
| TD2003 | $0.363^{e}$ | $\mathbf{0.359}^{e}$ | $0.333^{f}$ | $0.327^{f}$ | $0.354^{fe}$ | $0.323^{f}$ | $0.203^{fe}$ | $0.202^{fe}$ | $0.335^{f}$ | $0.354^{e}$ | $0.262^{fe}$ | 0.344 | $0.262^{fe}$ |
| TD2004 | 0.351 | 0.354 | 0.346 | 0.334 | $0.329^{e}$ | **0.359** | $0.279^{fe}$ | $0.314^{f}$ | 0.356 | 0.333 | $0.289^{fe}$ | 0.329 | $0.313^{f}$ |
| Black-Box: Regression Tree | | | | | | | | | | | | | |
| WEB10K | $0.424^{e}$ | $0.423^{f}$ | $\mathbf{0.424}^{e}$ | $0.422^{f}$ | $0.423^{f}$ | $0.421^{fe}$ | $0.411^{fe}$ | $0.409^{fe}$ | $0.423^{f}$ | $0.423^{f}$ | $0.408^{fe}$ | $0.417^{fe}$ | $0.418^{fe}$ |
| YAHOO | $0.703^{e}$ | $\mathbf{0.703}^{e}$ | $0.702^{e}$ | $0.699^{f}$ | $0.702^{e}$ | $0.696^{fe}$ | $0.695^{fe}$ | $0.695^{fe}$ | $0.700^{f}$ | $0.701^{fe}$ | $0.695^{fe}$ | $0.695^{fe}$ | $0.699^{f}$ |
| TD2003 | 0.363 | 0.35 | 0.333 | **0.366** | 0.357 | 0.349 | $0.202^{fe}$ | $0.202^{fe}$ | $0.293^{fe}$ | $0.300^{fe}$ | $0.203^{fe}$ | 0.344 | $0.262^{fe}$ |
| TD2004 | 0.351 | 0.344 | 0.347 | **0.349** | 0.34 | **0.349** | $0.298^{fe}$ | $0.298^{fe}$ | 0.327 | 0.33 | $0.272^{fe}$ | $0.329^{fe}$ | $0.313^{fe}$ |

Table 4.16: *The NDCG@10 values in evaluated datasets, using the Random Forest model. Bold represents the best values for FS methods. The superscript letters e and f represent results statistically distinguishable with the* $\overset{E}{\succ}$ *objective and the Full set of features, respectively.*

It is worth noting that $\overset{\text{E-R}}{\succ}$, $\overset{\text{E-G}}{\succ}$, and $\overset{\text{G}}{\succ}$ outperformed the ranking performance of $\overset{\text{E}}{\succ}$ in the majority of the cases, which has effectiveness as the main objective criterion. In fact, the risk-sensitive computation includes as a component the gain of effectiveness against the BS4R. Therefore, in a multi-objective method such as $\overset{\text{E-R}}{\succ}$ and $\overset{\text{E-G}}{\succ}$, where there is the combination of effectiveness and risk-sensitiveness, there is also a tendency that the selected individual optimize the effectiveness more than methods which apply the effectiveness as the unique objective criterion. In case of $\overset{\text{G}}{\succ}$, each query effectiveness is compared against a set of BS4R, driving the evolutionary search for a region with less degradation of queries.

$\overset{\text{E-R}}{\succ}$, $\overset{\text{E-G}}{\succ}$, and $\overset{\text{G}}{\succ}$ are the best three objective criteria when evaluating effectiveness, with $\overset{\text{E-R}}{\succ}$ and $\overset{\text{E-G}}{\succ}$ being more often statistically indistinguishable to the full set of features. For the methods that include the number of features as an objective criterion, they usually end up at positions in the search spaces containing solutions that damage effectiveness. Considering the FS baselines, DivFS performed more consistently among different datasets, as it ties with full feature more often than the BTFS method.

### 4.5.3 Feature Reduction Evaluation

Figures 4.7 and 4.8 present the feature space reduction for all evaluated methods, when using Linear Regression and Regression Tree as black-boxes, respectively. As expected, a more drastic reduction is obtained when the number of features is included as an objective criterion, except for $\overset{\text{G-F}}{\succ}$ and $\overset{\text{E-F-G}}{\succ}$, where the $G_{RISK}$ function prevents the exploration of regions with less risk-sensitive evaluation, due to the coevolutionary search performance. As a result, $G_{RISK}$ also obtained less "risky" models (see Tables 4.8 to 4.11 ).

In particular, $\overset{\text{E-F}}{\succ}$, $\overset{\text{T-F}}{\succ}$, and BTFS, reduced the number of features dramatically (varying from almost 69% to 90% ), though with a resulting reduction in effectiveness and risk-sensitiveness. We note that, $\overset{\text{E-R}}{\succ}$ and $\overset{\text{E-G}}{\succ}$ could reach a significant reduction, over 13% and 38%, respectively, without degrading the effectiveness and risk-sensitiveness.

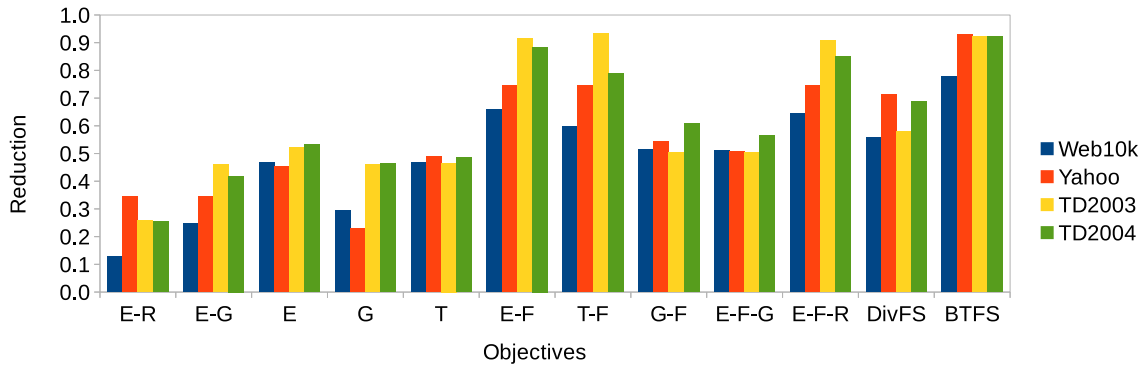



Figure 4.7: *Description of feature reduction for the FS methods, using Linear Regression as Black-Box.*

The results in the large datasets (WEB10K and YAHOO) show an important causality behavior between the number of features and risk-sensitiveness/effectiveness. The ranking method learned with Random Forest using all features ("Full Set of Features") presented the best values for almost all of the risk-sensitive and effective measures (Tables 4.8 to 4.16). Suggesting that more features mean more risk-sensitive or effective models. However, the methods combining effectiveness and risk-sensitiveness as objective criteria ($\overset{\text{E-R}}{\succ}$ and $\overset{\text{E-G}}{\succ}$)

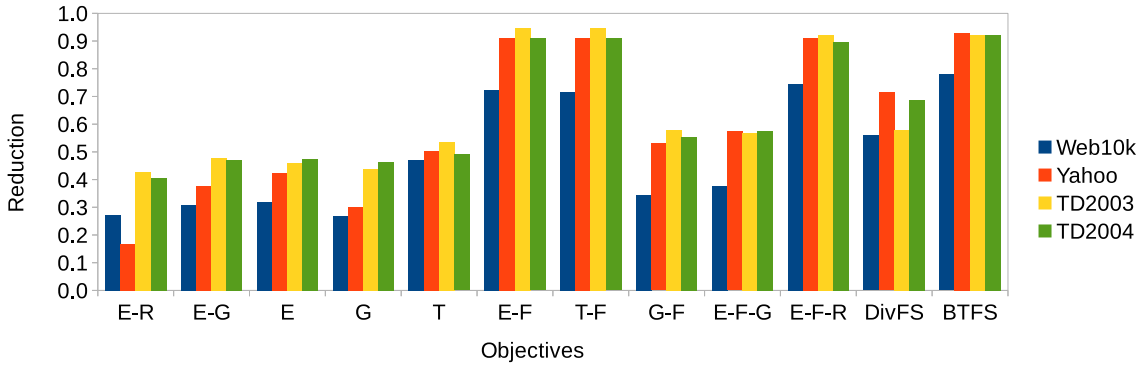



Figure 4.8: *Description of feature reduction for the FS methods, using Regression Tree as Black-Box.*

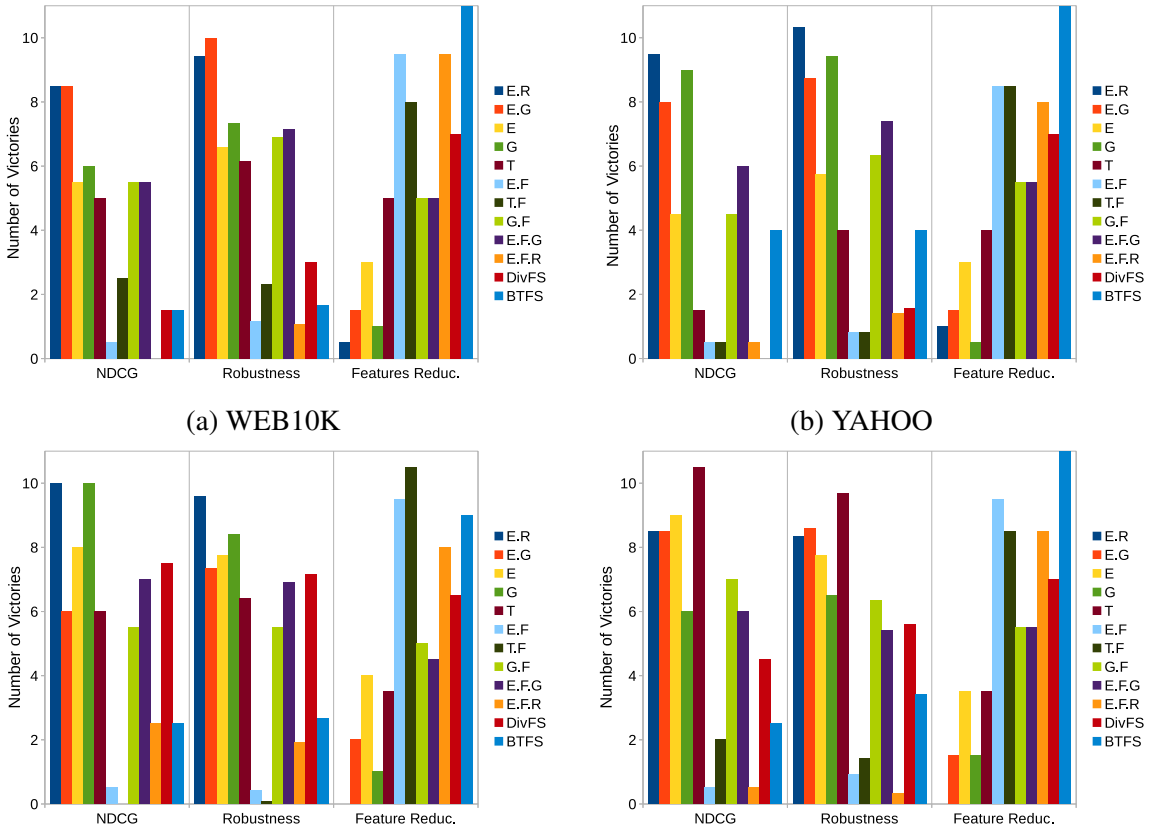
are capable of obtaining similar effectiveness and risk-sensitiveness by exploring the search space trying to find subsets of features that optimize the accuracy while eliminating unimportant subsets. The consequence is that the proposal multi-objective methods present results statistically similar to those obtained by the method using all the features ("Full Set of features") while still reducing the set of features, as shown in Figures 4.7 and 4.8.

In fact, we observe that the rate of noisy and redundant features are absolutely uncertain, and some datasets seem to have more of these kinds of features than others. Therefore, our experiments show that a proper approach with the ability to search for a random set of features that comply with some specific objectives, is an important contribution in FS for L2R.

### 4.5.4 Varying the Goals when Performing FS

The goal of our dissertation is not to learn the model with the smaller set of features, but to obtain a possibly little smaller set of features that guarantees ranking effectiveness and risk-sensitiveness performance through distinct multi-objectives methods. However, we also note that the best objective method depends on the main user goal. Hence, for the purpose of evaluating distinct aims, Figure 4.9 summarizes the aforementioned experiments for all objectives with the assessed datasets. Each chart in Figure 4.9 shows the number of victories for each objective against other when evaluation separately: effectiveness (NDCG@10), risk-sensitiveness, and feature reduction. The number of victories in the figure considers the average for evaluated two black-boxes. In case of risk-sensitiveness evaluation, the figure shows the best methods by computing the average victories of four measures: $T_{RISK}$, $G_{RISK}$, $L > 20\%$, and $Win$. In the figure, specifically for WEB10K and YAHOO, we consider a statistical difference (t-Test, with 95% confidence) to decide whether a method is better than other in the counting process. For TD2003 and TD2004 datasets we do not apply

paired statistical test, but only the absolute mean difference, as already commented due to the difficulty to obtain statistical significance differences in these collections [Gomes et al., 2013].



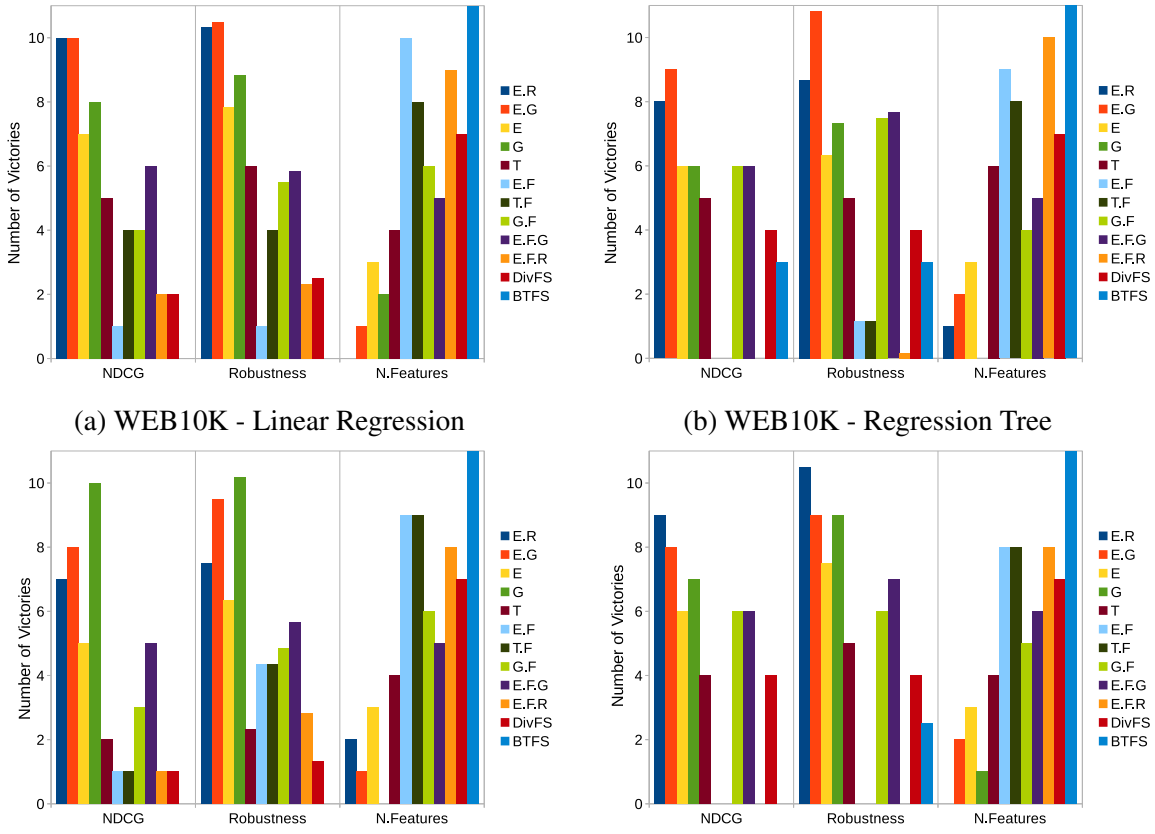

(a) WEB10K (b) YAHOO

(c) TD2003 (d) TD2004

Figure 4.9: *The average performance over two black-boxes, summarizing the victories with T-test (95% confidence).*

Figure 4.9 shows that $\overset{\text{E-R}}{\succ}$, $\overset{\text{E-G}}{\succ}$ and $\overset{\text{G}}{\succ}$ obtained the best trade-offs between effectiveness and risk-sensitiveness, while reducing some features. More often, $\overset{\text{E-R}}{\succ}$ and $\overset{\text{E-G}}{\succ}$ criteria appear in all experiments as the best results, concerning both effective and risk-sensitive evaluation. In addition, both produce the best results when compared to objectives that try to optimize only one of both criteria. The single objective $\overset{\text{G}}{\succ}$ also obtained a good trade-off between effectiveness and risk-sensitiveness, outperforming the $T_{RISK}$. As explained, this occurs because $T_{RISK}$ uses the same BS4R during the generations, hence decreasing the exploration over the search space, differently from $G_{RISK}$ function which exploits distinct BS4R over the evolutionary search.

Examining Figure 4.9 one can conclude that, if one of the goals in FS is feature reduction without large effectiveness losses and with low-risk, the best solution is the $\overset{\text{E-F-G}}{\succ}$




(a) WEB10K - Linear Regression

(b) WEB10K - Regression Tree

(c) YAHOO - Linear Regression

(d) YAHOO - Regression Tree

Figure 4.10: *LambdaMART executions.*

method. For this objective criteria, the reduction was around 55% of features, for all datasets, with a relatively good effective and risk-sensitive performance for both assessed black-boxes. The method $\overset{\text{E-F-R}}{\succ}$ performed inferior to $\overset{\text{E-F-G}}{\succ}$ in terms of effectiveness and risk-sensitiveness while keeping a comparable number of features in most collections. It is interesting to observe that both methods, while reducing more than 55% of features, performed consistently better in both – effectiveness and risk-sensitiveness – when compared to $\overset{\text{E}}{\succ}$ and all the other FS methods aiming at feature reduction in the two largest datasets (WEB10K and YAHOO).

To conclude, if the main goal is to obtain some reduction in dimensionality in a robust way and without effectiveness losses, the best options are $\overset{\text{E-R}}{\succ}$, $\overset{\text{E-G}}{\succ}$, $\overset{\text{G}}{\succ}$ with the dimensionality reductions varying from 13 to 46% (described in Tables 4.7 and 4.8 ).

The same conclusions can be observed when applying the LambdaMART algorithm to train the model with the selected feature subset, as shown in Figure 4.10. The figure summarizes the results for the main datasets, WEB10K and YAHOO, showing the performance of each black-box. In general, both methods $\overset{\text{E-R}}{\succ}$ and $\overset{\text{E-G}}{\succ}$ obtained the best results, with $\overset{\text{G}}{\succ}$ outperforming both $\overset{\text{E-R}}{\succ}$ and $\overset{\text{E-G}}{\succ}$ only in YAHOO with Linear Regression.

## 4.6 An Overfitting Evaluation

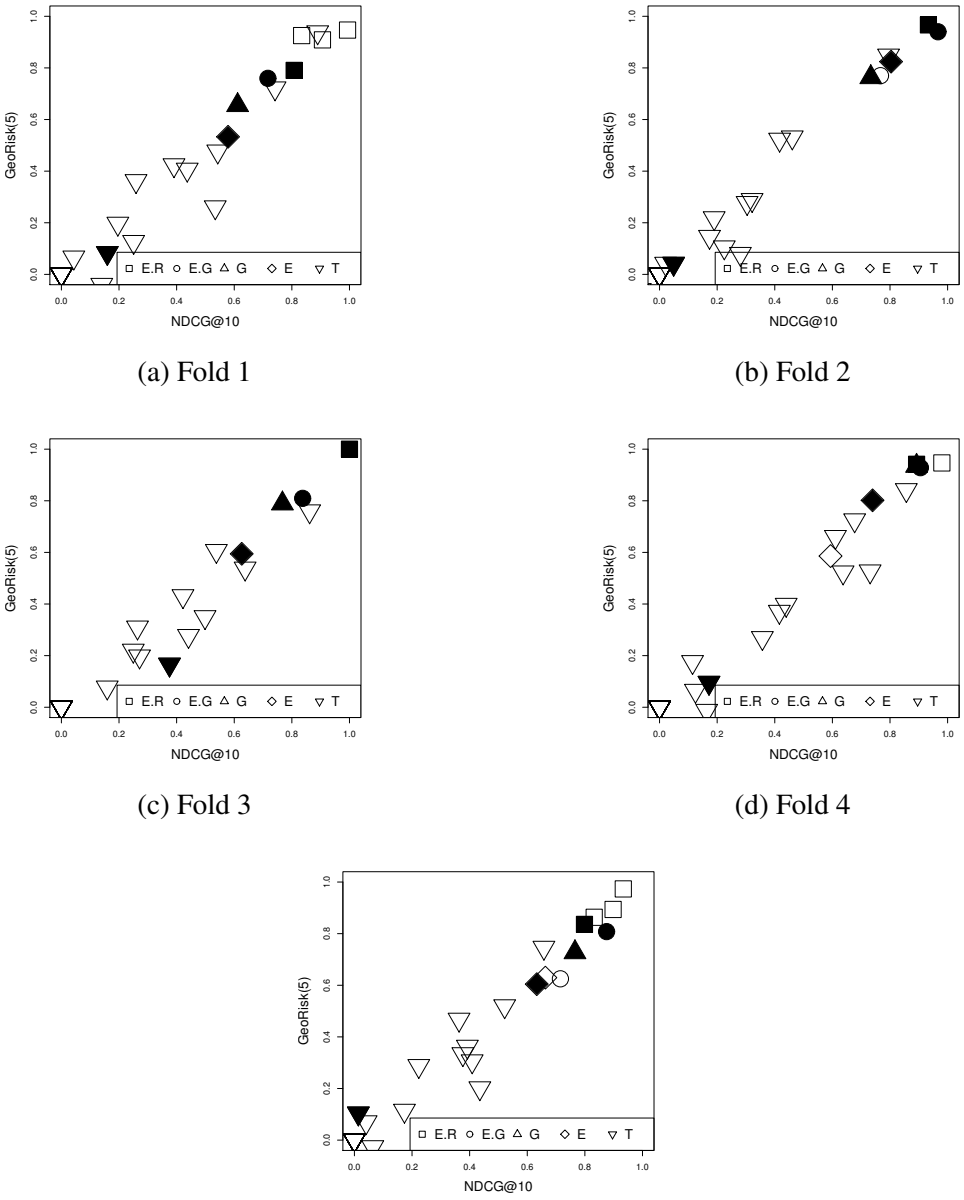
*Q5 – How is the overfitting behavior of proposed objective criteria and evolutionary FS methodology?*

In the previous sections we presented our improvements when selecting only one individual from the Pareto set. In this section, we evaluate all the content of Pareto set in order to describe the overfitting behavior of our proposed strategies and the arrangement of individuals over two dimensions, effectiveness and risk-sensitiveness.

Each individual from the Pareto set is presented as geometrical symbol in Figures 4.11 and 4.12, considering risk-sensitiveness ($G_{RISK}$) and effectiveness (NDCG@10). The Figures describe the results of FS methods observing each fold separately with normalized values for each axis. The geometrical symbols in Figures show the results of main objective criteria: Effectiveness-$F_{RISK}$ ($\overset{\text{E-R}}{\succ}$), Effectiveness-$G_{RISK}$ ($\overset{\text{E-G}}{\succ}$), Effectiveness ($\overset{\text{E}}{\succ}$), $G_{RISK}$ ($\overset{\text{G}}{\succ}$), and $T_{RISK}$ ($\overset{\text{T}}{\succ}$). We are not considering objective criteria which include the number of features for this experiment, because these objectives tend to downgrade effectiveness, as discussed in Section 4.3. In these experiments all geometric symbols describe the performance on the test set. In order to describe the overfitting behavior, we highlight the best individual in training set with filled ones, which are also those selected by the methods in the discussion of the previous sections. In the case of $\overset{\text{E}}{\succ}$, $\overset{\text{G}}{\succ}$, and $\overset{\text{T}}{\succ}$, as there is no Pareto set during the evolutionary process, we are showing the individuals which were statistically tied as the best ones in the last generation. For clarity purpose, we only describe results in WEB10K and YAHOO datasets with Linear Regression as black-box.

In Figure 4.11, for WEB10K, one can observe that both multi-objective optimizations, $\overset{\text{E-R}}{\succ}$ (square) and $\overset{\text{E-G}}{\succ}$ (circle), respectively, are closer to the top right corner in almost all folds, showing that these objectives are capable of finding the best individuals that maximize both objectives, and confirming observations made in Sections 4.5.2 and 4.5.1. On the other hand, even obtaining better results, Figure 4.11 shows an overfitting behavior for $\overset{\text{E-R}}{\succ}$ method, mainly on Folds 1 and 5, when the filled squares are below of the empty ones. For this experiment, $\overset{\text{E}}{\succ}$ and $\overset{\text{G}}{\succ}$ show a performance consistent with the findings of section 4.5.2 and 4.5.1, in which $\overset{\text{G}}{\succ}$ outperforms $\overset{\text{E}}{\succ}$.

In particular, for $\overset{\text{T}}{\succ}$ we observe that there are more individuals in the Figure (represented by triangle points down). This is because there are much more tied individuals in the last population for this objective criterion than for the other ones. Possibly due to the paired test over $T_{RISK}$, which should not provide statistical difference over the individuals, thus keeping more individuals which were not dominated by any other. As a result, $\overset{\text{T}}{\succ}$ presents




(a) Fold 1

(b) Fold 2

(c) Fold 3

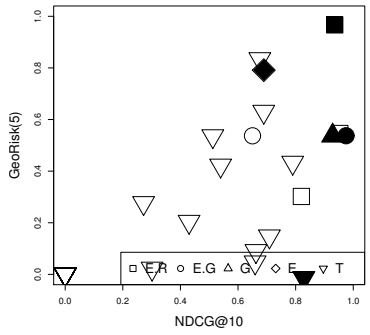
(d) Fold 4

(e) Fold 5

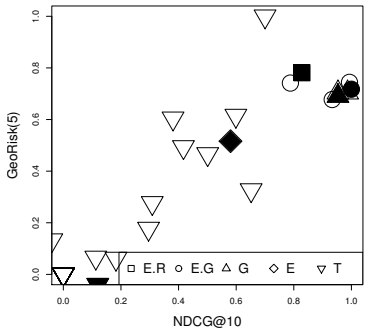
Figure 4.11: *Performance in effectiveness (NDCG@10) and risk-sensitiveness (GeorRisk) for individuals in Pareto frontier for Effectiveness-$F_{RISK}$ (E.R), Effectiveness-$G_{RISK}$ (E.G), Effectiveness (E), $G_{RISK}$ (G) and $T_{RISK}$ (T), on WEB10K dataset.*

a strong overfitting behavior, with many more empty triangles down when compared to the filled ones. In fact, we believe that this overfitting can decrease the evaluation of $\overset{\mathrm{T}}{\succ}$ for ranking and risk performance, as described in Sections 4.5.2 and 4.5.1.

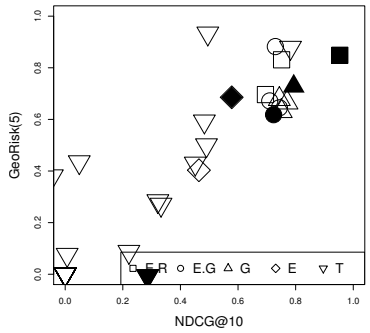
Figure 4.12 presents some similar results for YAHOO, and are also consistent with Sections 4.5.2 and 4.5.1. $\overset{\mathrm{E\text{-}R}}{\succ}$ obtained the best performance on Fold 1, and a better risk-



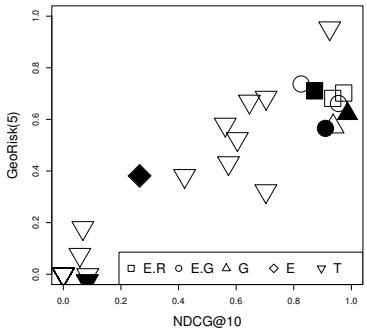

(a) Fold 1



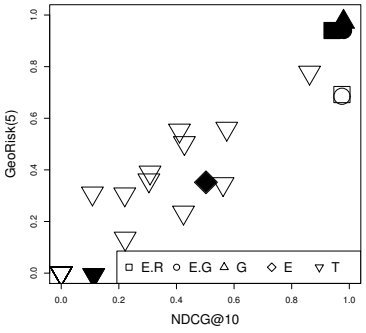

(b) Fold 2




(c) Fold 3




(d) Fold 4




(e) Fold 5

Figure 4.12: *Performance in effectiveness (NDCG@10) and risk-sensitiveness (GeorRisk) for individuals in Pareto frontier for Effectiveness-$F_{RISK}$ (E.R), Effectiveness-$G_{RISK}$ (E.G), Effectiveness (E), $G_{RISK}$ (G) and $T_{RISK}$ (T), on YAHOO dataset.*

sensitiveness performance on Folds 1, 2, 3, and 4. In Fold 5, the best three methods, $\overset{\text{E-R}}{\succ}$, $\overset{\text{E-G}}{\succ}$, and $\overset{\text{G}}{\succ}$ are practically tied. For this experiment, $\overset{\text{G}}{\succ}$ appears a little better than $\overset{\text{E-G}}{\succ}$, outperforming it on Folds 3 and 4. The overfitting can also be observed for YAHOO, especially for $\overset{\text{T}}{\succ}$, where the filled triangle points down are closer to the down left right corner than the empty ones.

To sum up and answering our Q6 research question, this experiment shows that $\overset{\text{E-R}}{\succ}$, $\overset{\text{E-G}}{\succ}$ and $\overset{\text{G}}{\succ}$ methods are capable of finding more often individuals at the top right corner. In addition, we could observe an overfitting tendency for some objectives, more often for $\overset{\text{T}}{\succ}$ and $\overset{\text{E-R}}{\succ}$. This suggests a future work to further evaluate the parameters settings or even make a more profound analysis of methods to select individuals from the Pareto set.

## 4.7 Describing Features with Greater Impact on Risk-Sensitiveness

*Q6 – Are there groups of features which have larger impact on risk-sensitiveness than effectiveness?*

Our work takes as an important concept the selection of features through a wrapper strategy, which assigns a fitness score regarding the interactions of features with a L2R algorithm. For instance, our experiments in section 4.5 provide a final set of features (from features interactions) which optimizes effectiveness and risk-sensitiveness. As a result, it is difficult to rank the features concerning their importance over risk-sensitiveness only. However, we now address this challenge through a search for algorithms of features which provide more effective or risk-sensitive performance, when concerning its interactions to build a L2R model.

In order to provide this quality evaluation, we had initially assessed the frequency of some features in the individuals in the Pareto Set in the last generation. However, we could not find a pattern of features in several individuals of the Pareto set. Due to the feature similarity in the datasets, similar individuals are composed of distinct features. Alternatively, increasing the abstraction level of the L2R feature space, one can observe that some features are similar in the sense that they are obtained by a same algorithm or measure applied to different parts of the documents. For instance, to provide two features, the same algorithm is applied to the title and the body of the document. Hence, we reduce the aforementioned issue for the following question: Are there groups of features, build by the same algorithm, which provides more impact on a specific objective criterion?

We address this question looking for groups of algorithms of features which provide more effectiveness or risk-sensitiveness performance. For this analysis, a group is a set of features obtained by the same algorithm, and each algorithm scores the documents to the queries in the training set regarding some parts of the documents. For instance, the “Covered Query Term” algorithm is a group with five features, as it counts the frequency of query terms

| | WEB10K | | | | | |
|---|---|---|---|---|---|---|
| | Linear Regression | | | Regression Tree | | |
| Groups | E-G $\succ$ | E-R $\succ$ | E $\succ$ | E-G $\succ$ | E-R $\succ$ | E $\succ$ |
| Covered query term number | ■ | ■ | ■ | ■ | ■ | ■ |
| Covered query term ratio | ■ | ■ | | ■ | ■ | |
| Stream length | ■ | ■ | ■ | ■ | ■ | ■ |
| IDF (Inverse document frequency) | ■ | ■ | | ■ | ■ | ■ |
| Sum of term frequency | ■ | ■ | ■ | ■ | ■ | ■ |
| Min of term frequency | ■ | ■ | ■ | ■ | ■ | ■ |
| Max of term frequency | ■ | ■ | | ■ | ■ | |
| Mean of term frequency | ■ | ■ | ■ | ■ | ■ | ■ |
| Variance of term frequency | ■ | ■ | | ■ | ■ | |
| Sum of stream length normalized term frequency | ■ | ■ | | ■ | ■ | |
| Min of stream length normalized term frequency | ■ | ■ | | ■ | ■ | |
| Max of stream length normalized term frequency | ■ | ■ | | ■ | ■ | ■ |
| Mean of stream length normalized term frequency | ■ | ■ | | ■ | ■ | ■ |
| Variance of stream length normalized term frequency | ■ | ■ | ■ | ■ | ■ | |
| Sum of TF*IDF | ■ | ■ | | ■ | ■ | |
| Min of TF*IDF | ■ | ■ | ■ | ■ | ■ | |
| Max of TF*IDF | ■ | ■ | ■ | ■ | ■ | ■ |
| Mean of TF*IDF | ■ | ■ | ■ | ■ | ■ | ■ |
| Variance of TF*IDF | ■ | ■ | | ■ | ■ | |
| Boolean model | ■ | ■ | ■ | ■ | ■ | |
| Vector space model | ■ | ■ | ■ | ■ | ■ | ■ |
| LMIR.ABS | ■ | ■ | | ■ | ■ | ■ |
| LMIR.DIR | ■ | ■ | | ■ | ■ | |
| LMIR.JM | ■ | ■ | | ■ | ■ | ■ |
| Number of slash in URL and length of URL | ■ | ■ | | ■ | ■ | ■ |
| Inlink, OutLink, pageRank, SiteRank | ■ | ■ | ■ | ■ | ■ | ■ |
| QualityScore and QualityScore2 | ■ | ■ | ■ | ■ | ■ | ■ |
| Query-url, url, and url dwel click count | ■ | ■ | ■ | ■ | ■ | ■ |

Table 4.17: *Algorithms as meta-features obtained when performing* $\overset{E\text{-}G}{\succ}$, $\overset{E\text{-}R}{\succ}$, *and* $\overset{E}{\succ}$ *objective criteria for WEB10K dataset.*

in the body, anchor, title, URL, and the whole document. Tables 4.17, 4.18 and 4.19 list the group used in our experiment, which are derived from MSLR-WEB10K (from Microsoft Research[17]) and LETOR[18] datasets descriptions. For this experiment, we have executed our SPEA2 algorithm with a space of groups (contrary to a space of features) in order to obtain the selected groups which optimize some objective-criteria. Concerning this, an individual

[17]http://research.microsoft.com/enus/projects/mslr/
[18]http://research.microsoft.com/enus/um/people/letor/

| | TD2003 | | | | | |
|---|---|---|---|---|---|---|
| | Linear Regression | | | Regression Tree | | |
| Groups | $\overset{\text{E-G}}{\succ}$ | $\overset{\text{E-R}}{\succ}$ | $\overset{\text{E}}{\succ}$ | $\overset{\text{E-G}}{\succ}$ | $\overset{\text{E-R}}{\succ}$ | $\overset{\text{E}}{\succ}$ |
| Covered query term number | ■ | ■ | ■ | | | |
| IDF (Inverse document frequency) | ■ | ■ | ■ | | ■ | ■ |
| TF*IDF | ■ | ■ | ■ | | ■ | ■ |
| Stream length | ■ | | | ■ | | |
| BM25 | ■ | ■ | ■ | ■ | ■ | ■ |
| LMIR.ABS | ■ | ■ | ■ | ■ | | |
| LMIR.DIR | | ■ | ■ | ■ | | |
| LMIR.JM | ■ | ■ | | ■ | ■ | ■ |
| Sitemap based term and core propagation | ■ | ■ | ■ | ■ | ■ | ■ |
| Hyperlink variations | ■ | ■ | ■ | ■ | ■ | ■ |
| HITS authority and hub | ■ | ■ | ■ | ■ | ■ | ■ |
| PageRank and HostRank | ■ | ■ | ■ | ■ | ■ | ■ |
| Topical: PageRank, HITS authority, and HITS hub | ■ | ■ | | ■ | ■ | |
| Inlink and outlink number | | | | | ■ | ■ |
| Number of: slash in URL, length of URL, and child page | | | ■ | ■ | ■ | ■ |
| Extracted title with: BM25, LMIR.ABS, LMIR.DIR and LIMIR.JM | ■ | ■ | ■ | ■ | ■ | ■ |

Table 4.18: Algorithms as meta-features obtained when performing $\overset{\text{E-G}}{\succ}$, $\overset{\text{E-R}}{\succ}$, and $\overset{\text{E}}{\succ}$ objective criteria for TD2003 dataset.

is an array which it is 0 when the group is absent, and 1 otherwise.

The results can be observed in Tables 4.17, 4.18 and 4.19, which paint in gray color the group of algorithms that appear in at least one individual in the Pareto Set and in more than two of the 5-folds[19]. In case of effectiveness objective ($\overset{\text{E}}{\succ}$), we are showing again the individuals which were statistically tied as the best ones in the last generation.

As we can see in Tables 4.17, 4.18 and 4.19, the risk-sensitiveness base criteria use more group of features over the individuals in the Pareto set than effectiveness criterion. In fact, in order to provide a robust model, more groups of features are used to improve the effectiveness for some queries over the individuals in the Pareto set. On the other hand, by using effectiveness as objective criterion more groups are removed from the final models, decreasing the effectiveness for some queries and the overall robustness.

We may say that the features that are in $\overset{\text{E-G}}{\succ}$ and $\overset{\text{E-R}}{\succ}$ but not in $\overset{\text{E}}{\succ}$ are necessarily improving the risk-sensitiveness performance. Specifically for effectiveness performance in the WEB10K dataset, groups such as “Sum of stream length normalized term frequency” and

[19]Reminding that we are using the 5-fold cross-validation for our experiments.

| | TD2004 | | | | | |
|---|---|---|---|---|---|---|
| | Linear Regression | | | Regression Tree | | |
| Groups | $\overset{\text{E-G}}{\succ}$ | $\overset{\text{E-R}}{\succ}$ | $\overset{\text{E}}{\succ}$ | $\overset{\text{E-G}}{\succ}$ | $\overset{\text{E-R}}{\succ}$ | $\overset{\text{E}}{\succ}$ |
| Covered query term number | ■ | | ■ | ■ | ■ | ■ |
| IDF (Inverse document frequency) | ■ | ■ | | ■ | ■ | ■ |
| TF*IDF | ■ | ■ | ■ | ■ | ■ | |
| Stream length | ■ | ■ | ■ | ■ | ■ | |
| BM25 | ■ | ■ | ■ | ■ | ■ | |
| LMIR.ABS | ■ | ■ | | ■ | ■ | ■ |
| LMIR.DIR | | ■ | | ■ | ■ | |
| LMIR.JM | | ■ | ■ | | | ■ |
| Sitemap based term and core propagation | ■ | ■ | ■ | ■ | ■ | ■ |
| Hyperlink variations | ■ | ■ | | ■ | ■ | ■ |
| HITS authority and hub | ■ | ■ | ■ | ■ | ■ | ■ |
| PageRank and HostRank | ■ | ■ | | ■ | ■ | ■ |
| Topical: PageRank, HITS authority, and HITS hub | | | | ■ | ■ | |
| Inlink and outlink number | ■ | ■ | | | ■ | ■ |
| Number of: slash in URL, length of URL, and child page | ■ | ■ | ■ | ■ | ■ | ■ |
| Extracted title with: BM25, LMIR.ABS, LMIR.DIR and LIMIR.JM | | ■ | ■ | ■ | | |

Table 4.19: *Groups of features obtained when performing* $\overset{\text{E-G}}{\succ}$, $\overset{\text{E-R}}{\succ}$, *and* $\overset{\text{E}}{\succ}$ *objective criteria for TD2004 dataset.*

"Variance of TF*IDF" are not applied in any of the weak learners, despite improving the risk-sensitiveness. In the case of TD2003 and TD2004 there are some distinctions between selected groups when varying the weak learner. For instance, "Inlink and outlink number" and "Topical: PageRank, HITS authority, and HITS hub", respectively to TD2003 and TD2004, are not necessarily improving the effectiveness or risk-sensitiveness when using the Linear Regression.

As a conclusion and answering our Q6 research question, we may say that there are groups of features which are not important to improve the overall effectivenesses, such as "Sum" and "Min of stream length normalized term frequency" in WEB10K. However, as these features are important to improve the effectiveness of same few queries, they are selected in case of risk-sensitiveness as a base criterion.

## 4.8 Assessing the Effect on the Results Variation of our Proposals

*Q7 – What are the effects on the results variation of the proposed statistical test comparison and multi-objective criterion?*

In our work, we provide distinct strategies to improve effectiveness and risk-sensitiveness while performing the feature selection. However, we now dedicate important attention to distinguish the effects of the strategies on the variation results. In other words, we assess the effects of our objective-criteria and paired statistical test proposals on the variation of NDCG@10 and $G_{RISK}$ results, called here as response variables.

In order to evaluate these effects, we perform a Factorial Design where each factor has two levels (or alternative values): presence or absence of the corresponding factor. In case of paired statistical test factor the first level considers its absence, that is, we use only the best mean without a statistical test, and the second one its presence, or the Wilcoxon method. For objective-criteria, the first level uses the effectiveness only, $\overset{\text{E}}{\succ}$, and the second one applies the combination of effectiveness and $G_{RISK}$ functions, $\overset{\text{E-G}}{\succ}$. We have chosen the $\overset{\text{E-G}}{\succ}$, because it provided improvements in effectiveness and risk-sensitiveness, besides increasing the feature reduction.

We have executed our SPEA2 process with aforementioned levels, assessing the response variable of the selected individual in the final Pareto set, and performing the comparison with the NCG@10 over the average with the 5-folds cross-validation. The $2^k$ Factorial Design executed here follows the definition in [Jain, 1991]. We separated our evaluation for each weak learner, as it is described in Figures 4.13, 4.14, 4.15, and 4.16.

The figures show the results for our factorial designing experiment, where each bar represents the effect of our factors (or interaction of factors) over the variation to the response variable. The figures describe the results for WEB10K and YAHOO datasets and Linear Regression and Regression Tree as weak learners. As we can see, a relevant result, the paired test has a high effect for the risk-sensitiveness ($G_{RISK}$ measure, or GeoRisk in the figures), considering its isolated factor or even in its interaction with the objective criteria, which is the case of a greater factor interaction in the Regression Tree results. Nevertheless, the statistical paired test is a very interesting method to improve the robustness of the learning processing. In fact, it allows a robust comparison over all queries, as it performs the comparison for all sample of queries, considering the difference of rank effectiveness for each query and not only the overall average value. Furthermore, the objective-criteria and their interactions with paired test also perform an important risk-sensitive ( $G_{RISK}$ measure)

effect in our experiments.

Changing towards effectiveness as a variable response, NDCG@10, the objective-criteria and their interactions with paired statistical test describe a larger effect than the statistical test. More specifically, in WEB10K for Regression Tree, the interaction of factors provided a larger effect on NDCG@10. Differently of WEB10K in Linear Regression, where the interaction has a minor effect.

Ideally, we would like to evaluate the impact of the weak learner. However, as it would be necessary to perform a strong learner in order to guaranty the presence and absence of the factor, which is impracticable due to not scalable characteristic of strong learners as black-box, we are leaving this evaluation as future work. We intend to use some available architecture technology, such as Graphical Processing Unit, to improve the execution time of the strong learner.

As a conclusion and answering our Q7 research question, we observe that the statistical test also improves the risk-sensitive performance of our method, allowing a robust comparison over all queries. In addition, the objective-criteria improved the effective performance, besides their interactions with paired statistical test has shown an important effect on effectiveness and risk-sensitiveness.

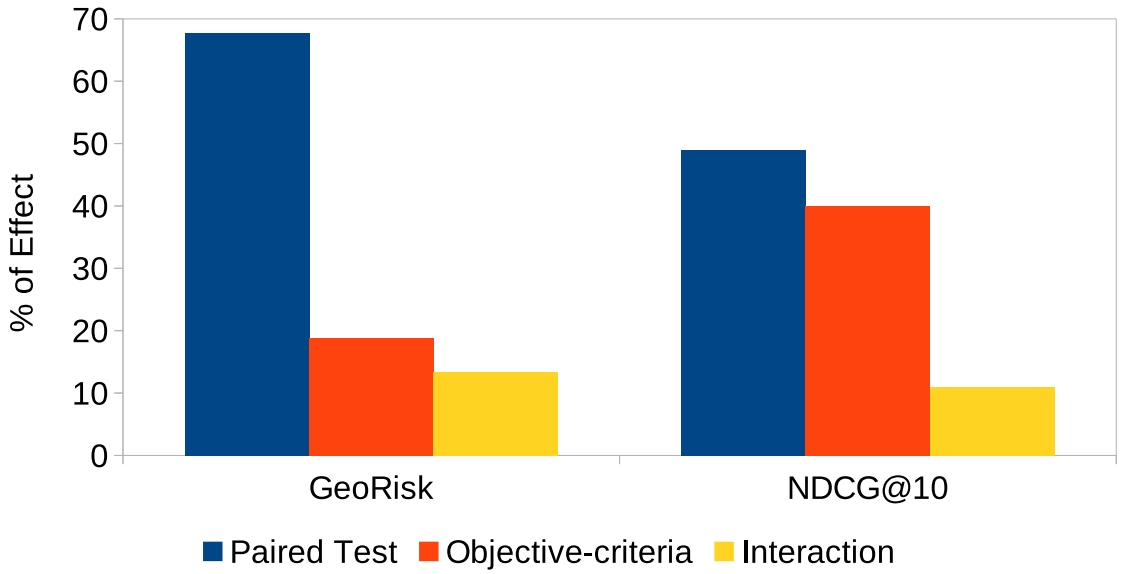



Figure 4.13: *The Factorial Design for Linear Regression with WEB10K*

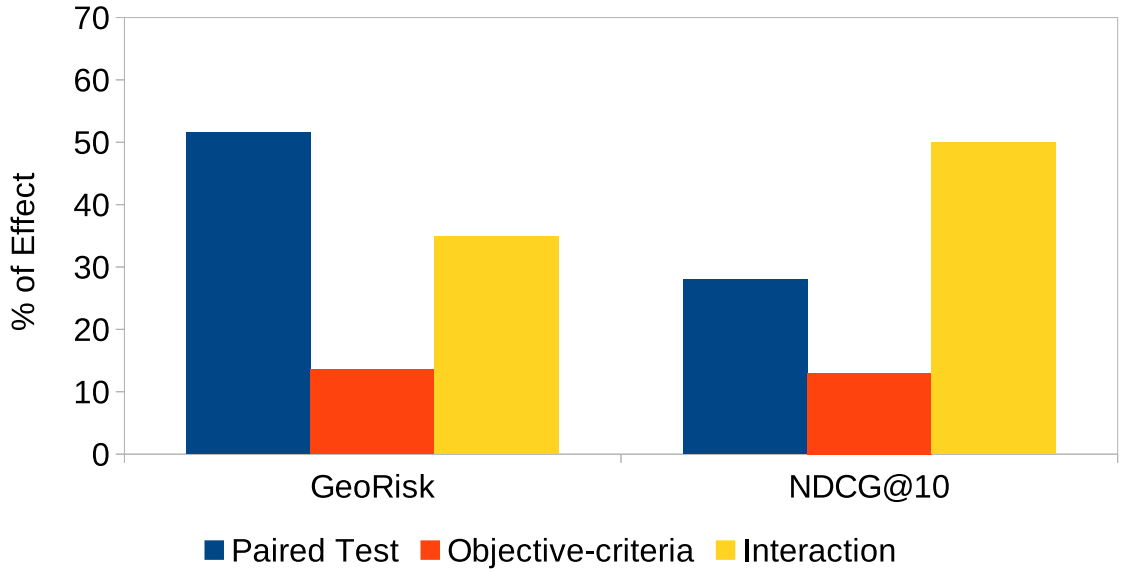



Figure 4.14: *The Factorial Design for Regression Tree with WEB10K*

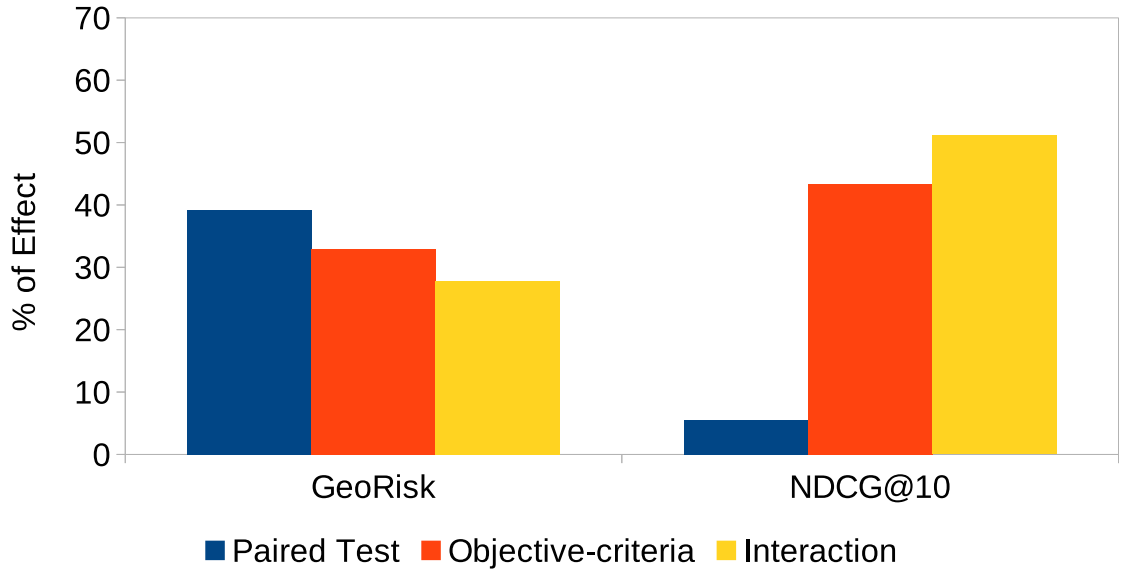



Figure 4.15: *The Factorial Design for Linear Regression with YAHOO*

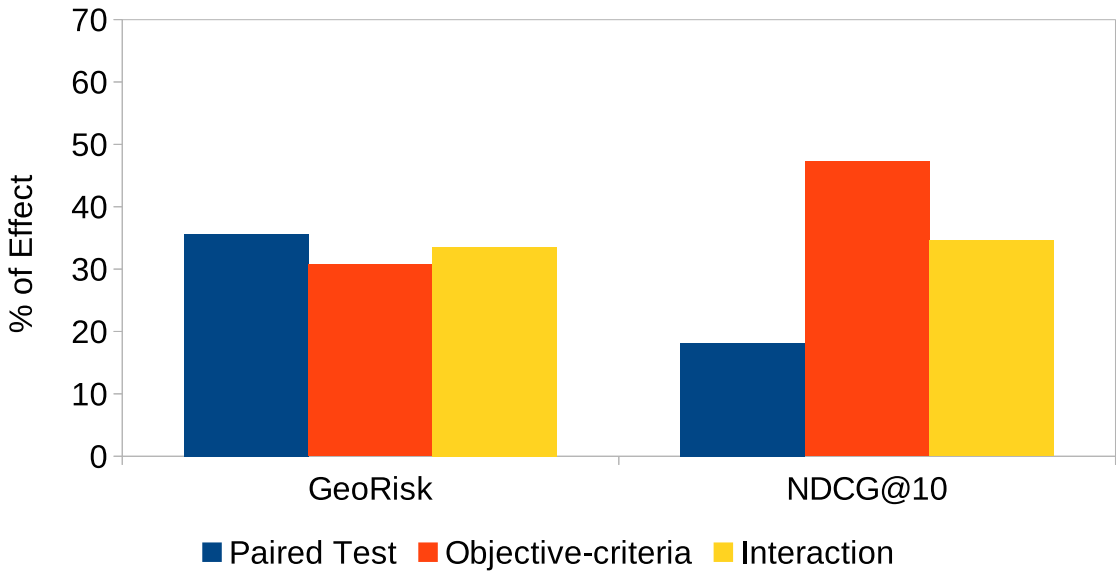



Figure 4.16: *The Factorial Design for Regression Tree with YAHOO*

# Chapter 5

# Conclusions and Future Work

In this chapter we present the summary of the results achieved and provide a description for future work.

## 5.1 Conclusions

This is the first dissertation that thoroughly investigated the impact of risk-sensitiveness in feature selection for Learning to Rank. In this context, it introduces single and multi-objective criteria that optimize risk-sensitiveness and effectiveness while performing the feature reduction. Furthermore, we also proposed a methodology based on multi-objective using the Pareto Frontier, improving the effectiveness and efficiency of the search space evaluation.

Before delving into the planned future work, in the following sections we outline each specific contribution.

### 5.1.1 A New Methodology to Evolutionary Algorithms

*Q1 – How to combine different optimization objectives in FS for L2R without being constrained to a particular L2R method?*

In this dissertation we perform our multi-objective criteria for FS in the L2R task using SPEA2 as a general multiobjective criteria, concerning the interaction of features on the wrapper strategy and without being attached to a particular L2R algorithm as a black-box. We noted that this strategy provides the flexibility to search for several regions in the feature space, even providing feature reduction without using number of feature as a objective criterion. As a result, we describe a new methodology to perform FS with wrapper strategy.

*Q2 – How to apply an efficient wrapper evolutionary FS algorithm over huge datasets, without loss of effectiveness?*

Our methodology extends the evolutionary wrapper algorithms by using weak learners as black-box. We show that weak learners, e.g. Linear Regression and Regression Tree, can be applied in a wrapper strategy to perform FS on the L2R task and by improving more than 120x the execution time without decreasing the effectiveness.

Furthermore, we note that a weak learner allows a more sensitive comparison to evaluate the individuals, as it assigns the fitness values penalizing the presence of bad features. In contrast to strong-learners, which can decrease the weight of bad features in time to build the model, providing similar effectiveness when comparing distinct sets of features (or individuals).

*Q3 – How to improve the selection of individuals inside of the Pareto frontier set, in order to provide a more effective subset of features?*

Our methodology also extends the works in Pareto set, as it applies strict comparison among individuals, by performing a paired statistical test to define the dominance relationship of the individual over the generations. As described in our experiments, this strategy reduces the conflict between individuals in multi-objective criteria, decreasing the Pareto set dimension and improving the effectiveness of the selected final individual. Moreover, we confirm our expectation that the paired test does not provide consistent influence when using only one objective criterion.

### 5.1.2 Risk-sensitive Feature Selection for Learning to Rank

*Q4 – What is the performance of risk-sensitiveness, effectiveness and feature reduction on the proposed objective criteria and methodology in FS for L2R?*

In this dissertation we stress the evaluation of several risk-sensitive measures with effectiveness and number of features as multi-objective to perform feature selection. As a result, we show that using effectiveness and risk-sensitiveness as objective criteria provide a better subset of features for L2R, which increases the effectiveness of most queries and also of some queries which are avoided in the absence of risk-sensitive measures. For instance, by using only the average of the effectiveness as an objective criterion, the training phase explores the search space with a minor set of features, which are specialized in the average of effective performance, and do not improve several queries.

In fact, we show that risk-sensitiveness is an important objective criterion in order to perform FS in L2R. In our experimental results, all methods which have only effectiveness

and/or number o features as an objective criterion could not outperform the methods which have risk-sensitiveness and effectiveness as criteria. This is an important contribution in the area, as effectiveness and feature reduction are commonly found in works of literature for FS in L2R. Moreover, we note that our FS objectives conclusions match the observations of the Robust Retrieval Track of TREC Voorhees [2005] in the beginning of this century, which optimizing the standard average effective performance can damage some difficult queries, improving only the better-performing ones.

Our work improves the feature-space by i) reducing the time performance to execute the L2R phases, making it more flexible to update the training set; ii) evaluating several feature selection objectives, for instance, reducing the feature dimensionality without damaging effectiveness; and mainly, iii) with the responsibility of not increasing the risk of obtaining bad predictions for some queries.

Furthermore, our experimental results show that it is possible to obtain a significant feature reduction without damaging risk-sensitiveness and effectiveness, and using a combination of multi-objective criteria is better than using a single one, even when the main goal is used as the objective.

*Q5 – How is the overfitting behavior of proposed objective criteria and evolutionary FS methodology?*

With respect to the overfitting, our experiment show that $\overset{\text{E-R}}{\succ}$, $\overset{\text{E-G}}{\succ}$ and $\overset{\text{G}}{\succ}$ methods are capable of finding more often individuals at the top right corner with less influence of overfitting. Differently, when using $\overset{\text{T}}{\succ}$ objective we could observe an overfitting tendency, which suggests a future work to further evaluate the parameters settings or even make a more profound analysis of methods to select individuals from the Pareto set.

*Q6 – Are there groups of features which have larger impact on risk-sensitiveness than effectiveness?*

We also evaluate the quality of specific features over effectiveness and risk-sensitiveness. We may say that there are groups of features which are not important to improve the overall effectivenesses, such as “Sum” and “Min of stream length normalized term frequency” in WEB10K. However, as these features are important to improve the effectiveness of same queries, they are selected in case of risk-sensitiveness as a base criterion. This is a very interesting result, as some features can now be considered important for the risk-sensitive perspective.

*Q7 – What are the effects on the results variation of the proposed statistical test com-*

*parison and multi-objective criterion?*

We observe that our proposal have distinct impact in results variation. For instance, we observe that the statistical test also improves the risk-sensitive performance when selection the feature, allowing a robust comparison over all queries. In addition, the objective-criteria proposed improved the effective performance, besides their interactions with paired statistical test has shown an important effect on effectiveness and risk-sensitiveness.

To sum up, from our experimental results we note that besides the features provide distinct important roles in the feature-space, e.g. low-risk and/or effectiveness, the rate of unimportant features (noisy and redundant) is absolutely uncertain in datasets, with some datasets having more of this kind of features than others. Hence, the task of selecting relevant sets of features becomes even more challenging. In this sense, our dissertation provides a relevant and a novel contribution in FS for L2R, by including the risk-sensitiveness as a criterion in FS, enhancing the selection of Pareto set individuals and the processing time of wrapper strategies

## 5.2 Future Work

The methodology described in this dissertation shows a multi-objective evolutionary execution for FS in L2R. In fact, there is room to improve our strategy for efficiency and effectiveness.

In terms of processing time, our experiments show that 99% of SPEA2 process is due to the black-box execution. Thus, even though the weak learner allows a better time cost, we consider as a second further step an evaluation of these weak learners in a parallel processing, such as a multi-core architecture in Graphic Processing Units (GPUs). Besides the weak learner can fit better in a GPU architecture than a strong-learner, with fewer iterations over the dataset, the computation of individuals in one generation is an *embarrassingly* parallel problem. In addition, we could assess the performance when increasing the number of individuals in the generations, exploring a larger search space.

Following our experiments in the factorial design in Section 4.8, we show that paired statistical test and the multi-objective criteria have a strong impact in the effective and risk-sensitive performance. However, we intend to extend this evaluation by considering other factors, such as evolutionary multi-objective algorithms and more L2R black-boxes. In our dissertation we use SPEA2 as a general-purpose multi-objective optimization method, however, we can evaluate if other sorts of general methods can improve the selection of best individuals or even finding the best individual with fewer generations. For instance, we expect to apply a factorial design with NSGAII [Deb et al., 2002] and AMOSA [Bandyopadhyay

et al., 2008] as alternative multi-objective methods. Besides that, we also intend to further execute our methodology with a strong learner in order to better evaluate the impact of the weak learner as a black-box. However, to evaluate a strong-learner algorithm, we need to adapt a fast implementation for an evolutionary environment in order to improve the time performance when executing many individuals of one generation.

In our work, we have explored improvements on the SPEA2 algorithm to obtain better performance in the context of a wrapper strategy to compute the fitness value. However, SPEA2 has many parameters to set up, the wrapper strategy is absolutely time consuming, and there is a hard work for tuning the Evolutionary algorithm settings, due to the huge search space ($2^n$, where n varies from 64 to 700 our evaluated datasets).Therefore, we intend to combine a filter strategy such as a wrapper one, by assessing the ranking effectiveness of a L2R model when using single features and assessing the use of both risk-sensitive and effectiveness evaluation as multi-objective criteria over the Pareto set. One possible application is to find features in the Pareto set, applying distinct objectives and their combinations, such as effectiveness and risk-sensitiveness of one feature and the similarity with other features. By combining a filter and a wrapper strategy, we intend to build a less complex model as well as to improve the processing time.

Another interesting future step is the prediction of features which improve some specific queries. As described in the Section 4.7, a smaller set of features was enough to provide a relevant effectiveness, while some extra features were applied to optimize the risk-sensitive performance. This happened because some queries need some specific features to improve their effectiveness. Therefore, besides to use a basic group of features, we may ask whether it is possible to select an extra group of features on-demand in order to improve the effectiveness of some specific queries. As a result, we can improve the overall effectiveness and the execution time, as we will apply the extra-features only when necessary. On way to perform this evaluation could be by the selection of sub-parts in the ensemble algorithms, as these parts are directly related to some features, e.g. the selection of regression tree in the Random Forest or LambdaMart algorithms.

Concerning the contributions of our work, in special the paired statistical test and the multi-objective optimization with risk-sensitive and effective performance, we can also evaluate their behavior in algorithms which build L2R models, without concerning FS task. In special, some recent works [Wang et al., 2015; Li et al., 2016] have applied evolutionary process with Genetic Algorithm and Genetic Programming to obtain an effective L2R model without applying the proposals suggested in this dissertation. In theory, our proposal could be adapted to them in order to also improve the effectiveness and robustness of the final model.

The dissertation presented here contributes to L2R, as a new methodology to execute

FS. However, we can also try to apply our methodology in many other scenarios, such as Recommender Systems and Question Answering. In case of Recommender Systems, we can use the risk-sensitiveness in multi-objective criteria to also improve the effectiveness when performing feature selection, or even we can use our paired statistical test to compare the models or the features in the training phase, considering the recommendations as a sample.

# Bibliography


Bandyopadhyay, S., Saha, S., Maulik, U., and Deb, K. (2008). A simulated annealing-based multiobjective optimization algorithm: Amosa. *Journal of Transactions on Evolutionary Computation*, 12(3):269--283.

Breiman, L. (2001). Random Forests. *Journal of Machine Learning*, 45(1):1--33.

Capannini, G., Dato, D., Lucchese, C., Mori, M., Nardini, F. M., Orlando, S., Perego, R., and Tonellotto, N. (2015). QuickRank: A C++ suite of learning to rank algorithms. *Proceedings of the 6th Italian Information Retrieval Workshop - WIIR*, pages 1--8.

Capannini, G., Lucchese, C., Nardini, F. M., Orlando, S., Perego, R., and Tonellotto, N. (2016). Quality versus efficiency in document scoring with learning-to-rank models. *Journal of Information Processing and Management: an International Journal*, 52(6):1161--1177.

Chapelle, O., Yi, C., and Liu, T.-Y. (2011). Future directions in learning to rank. *Proceedings on the 2010 International Conference on Yahoo! Learning to Rank Challenge - YLRC*, pages 129--136.

Collins-Thompson, K., Macdonald, C., Bennett, P., Diaz, F., and Voorhees, E. (2014). TREC 2013 Web Track Overview. *Proceedings of the 22nd Text REtrieval Conference (TREC 2013)*, pages 1--15.

Dalip, D. H., Lima, H., Gonçalves, M. A., Cristo, M., and Calado, P. (2014). Quality assessment of collaborative content with minimal information. *Proceeding of the 14th ACM/IEEE- Joint Conference on Digital Libraries - JCDL*, pages 201--210.

Das, S. (2001). Filters, wrappers and a boosting-based hybrid for feature selection. *Proceedings of the 18th International Conference on Machine Learning - ICML*, pages 74--81.

Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: Nsga-ii. *Journal of Transactions on Evolutionary Computation*, 6(2):182--197.

Dinçer, B. T., Macdonald, C., and Ounis, I. (2014a). Hypothesis testing for the risk-sensitive evaluation of retrieval systems. *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 23--32.

Dinçer, B. T., Macdonald, C., and Ounis, I. (2016). Risk-Sensitive Evaluation and Learning to Rank using Multiple Baselines. *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 483--492.

Dinçer, B. T., Ounis, I., and Macdonald, C. (2014b). Tackling Biased Baselines in the Risk-Sensitive Evaluation of Retrieval Systems. *Proceeding of the 36th European Conference on Information Retrieval - ECIR*, pages 26--38.

Freitas, M., Sousa, D., Martins, W., Couto, T., Silva, R., and Gonçalves, M. (2016). A Fast and Scalable Manycore Implementation for an On-Demand Learning to Rank Method. *Proceeding of the 17th Simpósio em Sistemas Computacionais de Alto Desempenho - WSCAD*, pages 1--12.

Geng, X., Liu, T.-Y., Qin, T., and Li, H. (2007). Feature selection for ranking. *Proceedings of the 30th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 407--414.

Gomes, G., Oliveira, V., Almeida, J., and Gonçalves, M. (2013). Is Learning to Rank Worth It? A Statistical Analysis of Learning to Rank Methods in the LETOR Benchmarks. *Journal of Information and Data Manager*, 1(1):57--66.

Guardado, J., Rivas-Davalos, F., Torres, J., Maximov, S., and Melgoza, E. (2014). An Encoding Technique for Multiobjective Evolutionary Algorithms Applied to Power Distribution System Reconfiguration. *Journal of The Scientific World - SWJ*, 2014(1).

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer, New York, NY, USA.

Hsieh, J. G., Lin, Y. L., and Jeng, J. H. (2008). Preliminary study on Wilcoxon learning machines. *Journal of IEEE Transactions on Neural Networks and Learning Systems*, 19(2):201--211.

Jain, R. (1991). *The art of computer systems performance analysis - techniques for experimental design, measurement, simulation, and modeling*. Wiley professional computing. Wiley, New York, USA.

Joachims, T. (2002). Optimizing search engines using clickthrough data. *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery in Data mining - KDD*, pages 133--142.

Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H., and Newell, C. (2012). Explaining the user experience of recommender systems. *Journal of User Modeling and User-Adapted Interaction*, 22(4-5):441--504.

Lai, H.-J., Pan, Y., Tang, Y., and Yu, R. (2013). FSMRank: feature selection algorithm for learning to rank. *Journal of IEEE Transaction Neural Networks and Learning Systems*, 24(6):940--52.

Laporte, L., Flamary, R., Canu, S., Dejean, S., and Mothe, J. (2014). Nonconvex regularizations for feature selection in ranking with sparse SVM. *Journal of IEEE Transactions on Neural Networks and Learning Systems*, abs/1507.00500(1):1118--1130.

Laumanns, M., Zitzler, E., and Thiele, L. (2001). On The Effects of Archiving, Elitism, and Density Based Selection in Evolutionary Multi-objective Optimization. *Proceedings of the First International Conference on Evolutionary Multi-Criterion Optimization - EMO*, pages 181--196.

Li, B., Li, J., Tang, K., and Yao, X. (2015). Many-Objective Evolutionary Algorithms: A Survey. *Journal of ACM Computing Surveys – CSUR*, 48(1):1--35.

Li, F. and Yang, Y. (2005). Using recursive classification to discover predictive features. *Proceedings of the 2005 ACM Symposium on Applied Computing - SAC*, pages 1054--1058.

Li, J., Liu, G., Yan, C., and Changjun, J. (2016). Robust Learning to Rank Based on Portfolio Theory and AMOSA Algorithm. *Journal of IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 47(6):1--12.

Liu, T.-Y. (2011). *Learning To Rank For Information Retrieval*. Springer, New York, USA.

Mohan, A., Chen, Z., and Weinberger, K. (2011). Web-search ranking with initialized gradient boosted regression trees. *Proceedings of the 2010 International Conference on Yahoo! Learning to Rank Challenge - YLRC*, pages 77--89.

Naini, K. D. and Altingovde, I. S. (2014). Exploiting Result Diversification Methods for Feature Selection in Learning to Rank. *Proceeding of the 36th European Conference on Information Retrieval - ECIR*, pages 455--461.

Pan, F., Converse, T., Ahn, D., Salvetti, F., and Donato, G. (2011). Greedy and randomized feature selection for web search ranking. *Proceeding of the 11th IEEE International Conference on Computer and Information Technology - CIT*, pages 436--442.

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *Journal of IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226--1238.

Qinbao Song, Jingjie Ni, and Guangtao Wang (2013). A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data. *Journal of IEEE Transactions on Knowledge and Data Engineering*, 25(1):1--14.

Sakai, T. (2014). Statistical reform in information retrieval? *Newsletter ACM SIGIR Forum*, 48(1):3--12.

Severyn, A. and Moschitti, A. (2015). Learning to Rank Short Text Pairs with Convolutional Deep Neural. *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR*, pages 373--382.

Shi, Y., Larson, M., and Hanjalic, A. (2010). List-wise learning to rank with matrix factorization for collaborative filtering. *Proceedings of the 4th ACM Conference on Recommender Systems - RecSys*, pages 269--272.

Shirzad, M. and Keyvanpour, M. (2015). A feature selection method based on minimum redundancy maximum relevance for learning to rank. *Proceeding of the 7th IEEE Artificial Intelligence and Robotics Conference - IRANOPEN*, pages 1--7.

Sousa, D., Couto, T., Martins, W., Silva, R., and Gonçalves, M. (2012). Improving on-demand learning to rank through parallelism. *Proceedings of the 13th international conference on Web Information Systems Engineering - WISE*, pages 526--537.

Srinivas, M. and Patnaik, L. M. (1994). Genetic Algorithms: A Survey. *Journal of Computer*, 27(6):17--26.

Tzeng, G.-H. and Tsaur, S.-H. (1997). Application of multiple criteria decision making for network improvement. *Journal of Advanced Transportation*, 31(1):49--74.

Voorhees, E. M. (2003). Overview of the TREC 2003 robust retrieval track. *In Proceedings of the 12th Text Retrieval Conference (TREC-12)*.

Voorhees, E. M. (2004). Overview of the TREC 2004 robust retrieval track. *In Proceedings of the 13th Text Retrieval Conference (TREC-13)*.

Voorhees, E. M. (2005). Overview of the TREC 2005 robust retrieval track. *In Proceedings of the 14th Text Retrieval Conference (TREC-14)*.

Wang, L., Bennett, P. N., and Collins-Thompson, K. (2012). Robust ranking models via risk-sensitive optimization. *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 761--770.

Wang, S., Wu, Y., Gao, B. J., Wang, K., Lauw, H. W., and Ma, J. (2015). A Cooperative Coevolution Framework for Parallel Learning to Rank. *Journal of IEEE Transactions on Knowledge and Data Engineering*, 27(12):3152--3165.

Wismans, L., Brands, T., Erik, B., and Bliemer, M. (2011). Pruning and ranking the Pareto optimal set, application for the dynamic multi-objective network design problem. *Journal of Advanced Transportation*, 48(6):512-- 525.

Zhang, P., Hao, L., Song, D., Wang, J., Hou, Y., and Hu, B. (2014). Generalized Bias-Variance Evaluation of TREC Participated Systems. *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management - CIKM*, pages 3--6.

Zitzler, E., Laumanns, M., and Thiele, L. (2001). SPEA2: Improving the strength pareto evolutionary algorithm. *Proceedings of Evolutionary Methods for Design, Optimization and Control with Applications to Industrial Problems - EUROGEN*, pages 12--19.