FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Label Ranking for Election Outcome Prediction

Miguel Jorge Gonçalves Pereira

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Cláudio Rebelo de Sá

Second Supervisor: Carlos Soares

January 30, 2018

Label Ranking for Election Outcome Prediction

Miguel Jorge Gonçalves Pereira

Mestrado Integrado em Engenharia Informática e Computação

January 30, 2018

Abstract

Predicting the election outcome is a complex task. In a famous recent case, Donald Trump managed to sweep the presidency of the United States with the majority of the pools indicating otherwise. This seems to indicate that currently used approaches, such as classic pools, might not be able to capture the multitude of variables involved in the election results.

Recent approaches, based on Data Mining and Machine Learning have been turning their focus to Social Media sources, which provide a great source of information in this regard.

In this work, we also propose to tackle the problem of Election Outcome Prediction with the use of Data Mining and Machine Learning. However, our focus will be in the study of the preferences of the parties and their socio-economic context. To tackle this, we suggest to transform the prediction of elections into a Label Ranking (LR) problem. LR is a subtask of the Preference Learning field, which uses a set of tools that allow the discovery of patterns in preferences. This has the advantage of allowing the prediction of, not only who is going to win, but also an ordered relation between the political parties or candidates.

In particular, we are going to focus on Pairwise Association Rules (PAR). We will use them for prediction purposes. They come with the advantage that they provide interpretable results, which is useful to analyze the predictions.

The results will be tested both in common LR datasets and in election datasets. We will compare our approach with other LR algorithms.

Part of the election data collected for this study was taken from PORDATA, a vast and up to date Portuguese socio-economical database.

In the end, considering the good results obtained, we believe that this work holds promise both as a contribution to the LR community and the Political Science field.

Resumo

Prever resultados eleitorais é uma tarefa complexa. Num famoso caso recente, Donald Trump conseguiu assegurar a presidência dos Estados Unidos apesar da maioria das sondagens apontar no sentido contrário. Isto parece indicar que as abordagens atualmente usadas, como as sondagens, podem não conseguir capturar a multitude de variáveis envolvidas nos resultados eleitorais.

Abordagens recentes baseadas em Data Mining e Machine Learning têm-se focado em usar Redes Sociais como fonte de dados, o que trouxe uma grande quantidade de informação.

Neste trabalho, também propomos lidar como o problema de Previsões Eleitorais usando Data Mining and Machine Learning. No entanto, o nosso foco vai ser no estudo das preferências dos partidos e dos seus contextos socioeconómicos. Para lidar com esta questão, sugerimos transformar o problema de prever resultados eleitorais num problema de Label Ranking (LR). LR é uma subárea de Preference Learning, que usa um conjunto de ferramentas que permitem descobrir padrões em preferências. Isto tem a vantagem de permitir prever, não só quem vai ganhar, mas também uma relação ordenada entre os candidatos ou partidos políticos.

Em particular, vamo-nos focar em Pairwise Association Rules (PAR) e em usá-las para efeitos de previsão. Elas têm a vantagem de produzir resultados interpretáveis, o que é útil para analisar as previsões.

Os resultados vão ser testados tanto em conjuntos de dados de LR comuns, como em conjuntos de dados relativos a eleições. Fazemos igualmente uma comparação da nossa abordagem com outros algoritmos de LR.

Parte dos dados relativos a eleições provêm da PORDATA, uma vasta base de dados Portuguesa de dados socioeconómicos.

No final, considerando os bons resultados obtidos, nós acreditamos que este trabalho é uma contribuição tanto para a comunidade de LR como para o campo da Ciência Política.

Acknowledgements

Firstly, I would like to thank my mentor Cláudio Sá for the relentless help, guiding and input in shaping this dissertation into what it became. I would also like to thank my mentor Carlos Soares for the always provided expertise.

A work of this dimension is only possible due to the support and love of my family, my Mother, my Father, my sister and brother, my uncles, my cousins and my late but always present grandfather.

A sincere thank you to all my friends whose constant goofing and companionship make life always brighter.

A special word of love to Ayça, for being the brightest light in my life.

Miguel Pereira

"We are such stuff as dreams are made on,
and our little life is rounded with a sleep."

William Shakespeare

Contents

List of Figures

LIST OF FIGURES

List of Tables

Abbreviations

ML	Machine Learning
AR	Association Rules
LR	Label Ranking
LRAR	Label Ranking Association Rules
PAR	Pairwise Association Rules

Chapter 1

Introduction

Election prediction is the science of declaring the outcome of an election, based on the results of a predefined set of methods [Arm01]. Some of those methods [LB90] may also allow us to study the variables that may have an impact on the electorate votes. The advantages of the study of those variables are translated for example in having a more clear map of the population's political tendencies, which therefore allows the parties to improve their political agenda.

The classic methods in this field range from polls to expert opinions to political stock markets [LBT12]. These methods, however, were not developed recently and do not harness the potential of the new data that is available nowadays. For example, in the 2010 US elections, 22 percent of the population used Twitter to voice their opinion on the matter [Smi11]. This presented an opportunity for Data Mining and Machine Learning (ML).

Data Mining and ML-based approaches yield a new way of extracting and using information from the available and aforementioned data. Social Media, for example, where Twitter is included, has been used more as a source to gather information about the opinions of the population [MMGA11]. An example is the POPmine project [SASS15], which uses Sentiment Analysis techniques to access the polarity of politicians and political parties. This is a so called descriptive approach, since it only attempts to describe information that it finds. There are other approaches, however, that focus more on the prediction part. For example, using again Sentiment Analysis but this time to predict Italian elections [CLOP15].

One subfield of ML that we are going to focus on is Preference Learning. This area studies how to discover preferences over a collection of objects [FH11a]. That will allow us to study the preferences between the parties.

These new methods allow for a exploration of existing vectors between socio-economical and demographical markers and election outcomes [BMK+13]. This is the study we are going to focus on in this work.

1.1 Problem

With all the methods available in the election prediction field, we believe there is still room for improvement, mainly because those methods face complex challenges. Classic approaches, on one hand, present diversified inconsistent results [LBT12]. In polls, for example, that problem may be due to the state of mind of the voters when they are inquired in the middle of political campaigns [GK93]. There may also exist a problem in assessing the preferences of the electorate because of the underlying anonymity of the voting process. On the other hand, ML methods that have been applied so far in this study also face complications. Methods that used data from social networks have to tackle problems as social networks sample bias [LB05]. Other difficulties in extracting knowledge from online social platforms include complex textual contexts, dealing for example with sarcasm and irony [CLOP15].

The existing challenges are translated in non optimal predictions to the public. A recent example was the 2016 US presidential elections, where Donald Trump won notwithstanding the number of predictions circulating in the media suggesting a different outcome. Examples like this one lead us to believe that there is still contributions to be made in the field, namely new approaches.

1.2 Motivation and Goals

To explore the Election Prediction problems through a different angle, we suggest the use of Label Ranking. This Preference Learning task [FH11a], defined later on, can be used to study the relations between, in this case, socio-economic markers and rankings of parties. More specifically, we will use Label Ranking Association Rules (LRAR) and Pairwise Association Rules (PAR).

LRAR [dSSJ+11] are an adaptation of Association Rules [AS94a] for Label Ranking problems. As an example, if in an association rule we have associations between economic markers, in LRAR it would be between a marker and a ranking of all the running parties.

PAR [dSAS+18] are also an adaptation of Association Rules for Label Ranking. However, unlike LRAR, they look for subsets of preferences within the rankings (e.g. between only two parties). In elections context, economic markers do not need to be associated with a ranking of all the running parties, it can be a preference between just two. Since PAR were originally proposed for descriptive tasks, part of this work will focus on its adaptation for predictive tasks. We test how they behave in the prediction of rankings on typical LR datasets. We also compare its behavior with state-of-the-art algorithms in the field of LR. Finally, we test them in the context of prediction of elections.

On the specific application of the prediction techniques on elections, we are going to study the 2005 and 2009 German regional elections [BMK+13]. Additionally, we combine publicly available data to create three datasets of Portuguese socio-economic geographical markers for the 2009, 2013 and 2017 regional elections. Interestingly, the 2017 Portuguese regional elections are contemporary with this work. Worth noting that, since we are going to study the voters on a higher level using municipalities, we will avoid the problem of anonymity,

In summary, we aim to:

- Use LRAR for predicting election outcomes;
- Adapt PAR for prediction;
- Test the use of PAR in predicting election outcomes;
- Compare PAR and LRAR, not only in predicting elections but also using traditional label ranking datasets;
- Study the trade-off between the prediction of complete rankings versus incomplete rankings in terms of accuracy;
- Contribute to Election Prediction and well as to Label Ranking existing literature.

1.3 Document structure

This report contains, apart from the introduction, four more chapters.

Chapter 2 is a review of literature related to the topics of this case study.

Chapter 3 describes and formalizes Label Ranking and, in particular, Pairwise Rules.

Chapter 4 details the implementation of the experiments done in the course of this work, as well as the description of the results.

Chapter 5 summaries the entire work, reviews it against the initials goals and considers possible future work.

Chapter 2

Background

This chapter introduces the problem of election prediction (Section 2.1) and association rules mining (Section 2.2).

2.1 Elections

From the times of Plato and Aristotle, almost 2500 years ago, there has been a study of the phenomena surrounding politics [Bar12]. Since then, the pursuit for more empirical information took place, by collecting data, investigating relationships and building theories. The incorporation of political related events in this pursuit let to the study of the one we are focused in this work: elections.

Until the 19th century, the theories developed did not have a strong foundation, being mostly verbal and conformed to established norms, with a lack of analytic methodologies [Mil97]. In the end of the century, that background was finally provided by Economics, originating the contemporary political sciences [Mil97]. The contributions of this merge translated into two different ways of studying political events such as elections: a micro and a macro perspective. This merge was incredibly substantial, allowing to either study political events such as elections from a micro and a macro perspective. On the one hand, a micro perspective allows us to look at the candidates or the voters individually. This enables, for example, studying a candidate's profile and assessing his actions in pre-elections times. A result of that approach was published in 1988 [RS88], where it was concluded that candidates of incumbent governments with a low competency use information that has not reached the public yet to make popular laws. Another example study is analyzing the socio-economic background of the members of the running parties [Rok09] and comparing it to the background of the population of voters. Although correlations were verified for some parties, it is notwithstanding a new piece of information. On the other hand, and much more related to the work developed in this project, we can study the problem from a macro perspective. We can, for example, aim at discovering economic markers that have a major impact on the outcome of election events. In the book Economics and Elections [LB90], Michael Lewis-Beck concluded that some economic performance markers affect party's votes in a linear way. He referenced that,

in a study by Goodhart and Bhansali, based in observations over 21 years, it was pointed out that there were two main factors that influenced political cycles: the annual inflation date and the unemployment level six months prior. The fact that this type of conclusions were obtained sets a positive background for this work. We also strive to find these markers, through a different way, but these kind of markers nonetheless.

2.1.1 Election prediction

Although the act of predicting can be considered old if more in the form of an art, it was only in more modern times that it was tackled as a science[LB05]. Predicting an election consists then in declaring a priori and election result, based on the outcome of a predefined set of methods[Arm01]. Those methods allow us to study which variables affect elections the most, and the positive impact resulting of said knowledge is the foundation of this work. It can be explored in depth, for example, the connection some might have with the political cycles(as seen in section 2.1). All this gives the possibility for parties to readjust their political agenda or change their potential candidates. Also, being possible to paint a more accurate map of what drives voters the most, we can have a more transparent view of the democratic reality. Several approaches have been used in the literature for predicting elections, from polls to expert opinions to political stock markets, resorting to economic performance markers and candidate profiles[LBT12]. In an standardization attempt, Lewis-Beck[LB05] labeled the types of models as conditional or unconditional, and after-the-fact or before-the-fact. In conditional models one or more of their variable's weight is unknown, whether in unconditional ones they all have their weight defined. After-the-fact models are applied to situations where the events already occurred and the prediction tries to accurately match what happened. In this case unconstitutionality also applies, since the details are already known. Before-the-fact models, in the other hand, attempt to predict an event that as yet to occur, based on past observations. According to the type of available data, they can be conditional or unconditional. One example of a before-the-fact unconditional approach given by Lewis-Beck is using a statistical model to predict US elections, by selecting a set of political and economic variables that would maximize a vote, as presidential popularity and GDP growth[LB05]. As it is possible to observe, this problem can be described as trying to predict the electoral outcome of a set of socio-economical variables. This is a natural setup of a classic machine learning classification problem, a connection that will be explored in the following section.

2.1.2 Elections, Machine Learning and Social Networks

The modern world has provided us the opportunity of accessibility to large amounts of data throughout the Internet. With the goal of making sense of all this of information and extracting knowledge from it, the efforts naturally turned to the fields of machine learning and data mining. In particular, those efforts depend on text mining, a branch of data mining focused on the techniques involved in extracting meaningful knowledge from unstructured textual sources[TO99]. In the topic of election prediction, the literature available mainly involve social networks and

blogging platforms. Social networks are a popular pathway for election prediction[MMGA11]. They are vastly disseminated in our societies and people use them to voice their opinions on all types of subjects. As long as we are able to filter out the noise, they can be, due to their current volume, a vast source of information. This line of thought came after other attempts to make predictions based on for example Twitter, like for evaluating the market success of movies or consumer goods[MMGA11]. There have been proposed some ways of extracting political information from this data. One example was trying to discover patterns, using a technique named subgroup discovery[BMK+13], in German elections. Other methods include attempting to correlate the search preferences of potential voters with election outcomes[PPK13], or connect sequences of events to said results[TWC16]. There are also approaches that try to forecast the election results from Twitter data using for example sentiment analysis and relative terms frequency [CLOP15]. Sentiment analysis, a commonly used techique explored in the subsequent section, is also used, for example, to explore the impact of crowd wisdom gathered in social forums[WL16]. Those approaches come with their own problems of course, ranging from existing bias to complex textual contexts [CLOP15]. There is also the problem of the different demographies that characterize the voting population and the social networks' users, since not everyone is in one of them. That difference has the danger of being translated into wrong principles and generalizations, that may invalidate any results obtained in association[MMGA11]. Another way to tackle the problem at hands, which is also in the basis of this work, is to look to past election results and try to extract meaningful (disruptive) descriptive associations between socio-economic variables and voters, using for example sub-group discovery[BMK+13]. The prediction techniques of this work stand on the premise that those association rules exist and can be extracted.

2.1.2.1 Sentiment Analysis

Sentiment analysis is a subfield of text mining that aims to extract opinions from different textual sources[SASS15]. Those opinions must be also labeled according to their positive, negative or neutral sentiment (polarity[SRP+13]). One example is the POPmine open-source project, developed with the contribution of several Portuguese institutions, including the University of Porto. This project aims to capture mentions and sentiments alike from a set of web sources related to politicians and political parties in Portugal. The sources of information used include social media (e.g Twitter and blogs) and mainstream media sites. Tweeter presents some complicated challenges, namely due to the nature of each post. A single tweet is small, non-contextualized, and also very likely informal and ambiguous, making it technically challenging to extract meaningful information[SASS15]. POPmine aims therefore at closing the gap between the knowledge that can be extracted from the named sources and interested parties without the skills to perform such analysis. The system is able to output both the frequency as well as the polarity of the political actors' mentions. There are other examples of the use of sentiment analysis techniques, as the case of [CLOP15]. Here it was attempted to predict an Italian election event using Tweeter and machine learning algorithms. Analyzing the sentiment in the tweets, one of the problems encountered was identifying sarcasm and irony. The conclusions suggested that an integration with other

sources of information(e.g. polls) could potentially enhance the method.

A technique also studied[JSZ15] was trying to extract a person psycho-social characteristics from a person's face (automated visual trait analysis)and subsequently use them for predicting the outcomes of social events that involve commitment. The method was applied in a series of elections in the US, with a interesting resulted accuracy.

2.2 Association rules mining

Association rules mining is a data-mining approach that consists in identifying relevant correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories [KK06]. In other words, if we have a set of transactions, where each transaction is a list of items a client bought, the goal is to identify items that are common between transactions, and items that indicate the presence of other items. They were created to discover meaningful correlations in contexts such as market management (using e.g. market baskets) and risk management (using e.g. credit card transactions)[KK06]. More formally, consider an instance space $\mathbb{X}$ where $desc(\mathbb{X})$ is a set of descriptors of the instances in $\mathbb{X}$, consisting of pairs $\langle$attribute, value$\rangle$ (e.g. $\langle$unemployment, high$\rangle$). The data $D = \{\langle x_i \rangle\},\ for\ i = 1,...,n$, where x_i is a vector of the values $x_i^j, j = 1,...,m$ of m independent variables, $\mathcal{A}$, that describe instance i. The set of descriptors of each x_i instance is $desc(x_i)$. An Association Rule (AR) between an antecedent(A) and a consequents(C) can then be defined as the implication:

$$A \rightarrow C$$

where $A \bigcap C = \emptyset$ and $A, C \subseteq desc(\mathbb{X})$.

2.2.1 Interest measures

Typically, there are two main interest measures that need to be specified, minimum support and minimum confidence. Besides these two, other measures can be also found in the literature, such as lift, correlation, conviction, etc [Omi03b]. The specification of the main interest measures [Omi03a] for a rule $A \rightarrow C$ in D is as follows:

Support percentage of instances in D that contain A and C

$$sup(A \rightarrow C) = \frac{\#\{x_i | A \cup C \subseteq desc(x_i), x_i \in D\}}{n}$$

Confidence percentage of instances that contain C from the set of instances that contain A

$$conf(A \rightarrow C) = \frac{sup(A \rightarrow C)}{sup(A)}$$

Lift measures the independence of the consequent, C, relative to the antecedent, A

$$lift(A \rightarrow C) = \frac{sup(A \rightarrow C)}{sup(A) \cdot sup(C)}$$

For example, if A is independent from C, $lift(A \rightarrow C) \sim 1$.

Improvement Improvement is a measure of the *improvement* that a certain rule yields in comparison to it's predecessor [BAG99]. It is the smallest difference between the confidence of a rule and all it's sub-rules that share the same consequent.

$$imp(A \rightarrow C) = min(\forall A' \subset A, conf(A \rightarrow C) - conf(A' \rightarrow C))$$

For example, if the *improvement* = 1%, a generated sub-rule would only be considered if it's confidence was 1% higher than it's predecessor's.

Label ranking algorithms have a tendency to generate a large number of rules [BAG99]. For this reason, pruning methods must be applied to discard rules with little to none added value.

2.2.2 Rules generation

The rules generation algorithms available can be generally be divided in two types: itemset-based and rule-based [dSSJ+11].

Itemset-based algorithms divide their process in two steps: frequent itemset computing and rules generation. A common method of this type is the APRIORI algorithm proposed in 1994 [AS+94b]. APRIORI was the original method for the induction of AR. It identifies all the AR which have a *support* and *confidence* higher than a predefined threshold. However simple and practical, APRIORI is computationally heavy, proportional to the number of possible combinations. It also typically generates a high amount of rules. Alternative approaches include: hashing [PCY95a], dynamic itemset counting [BMUT97], parallel and distributed mining [PCY95b] and mining integrated into relational database systems [TS98].

On the other hand, rule-based algorithms compute the frequent itemsets and generate the rules at the same time. An example is the FP-Growth [HPYM04] approach.

2.2.3 Using Association rules for Prediction

In 1998 it was proposed a solution for using association rules for classification [LHM98]. This method, very developed in recent years[NVHT12], resorts to Classification Association Rules (CAR) [ML98]. These are distinguishable from typical association rules due to their consequent being restricted to a class [LHM98].

Chapter 3

Label Ranking

This chapter starts by formalizing Label Ranking and exploring some of it's methods, then defines some evaluation measures, and finally includes our contribution by using Pairwise Association Rules for prediction.

Label Ranking is a subtask of Preference Learning. A subfield of machine learning, Preference Learning studies the task of learning how to predict an order relation over a collection of objects[FH11a]. Label ranking is a variation of the classic machine learning classification problems where, instead of predicting a class, it predicts the relative order of a predefined set of labels[PRdS16]. In other words, while in classification, the output is a class, in Label Ranking it is a ranking/order of classes.

3.1 Formalization

Let us consider an instance $x \in \mathbb{X}$. In classification the goal is to predict the class λ where x belongs. In Label Ranking the goal is to predict the ranking of a predefined set of classes $\mathcal{L} = \{\lambda_1, ..., \lambda_k\}$, associated with x. We assume the mapping of instances to rankings, $\mathbb{X} \to \Omega$, on the permutation space Ω, to be stochastic rather than deterministic, since they are always associated with a probability.

A ranking, also referred to as a *complete ranking* [FH11a] or a *strict ranking* [VG11] can be represented as a *strict total order* over $\mathcal{L}$ [dSAS+18], defined on Ω. The later can be defined as a binary relation, $\succ$, on a set $\mathcal{L}$, which is Irreflexive, Transitive, Asymmetric and Connected. As an example, the *strict total order* $\lambda_2 \succ \lambda_3 \succ \lambda_1 \succ \lambda_4$ can be represented as $\pi = (3,1,2,4)$. There are situations, however, where there is not enough data to make a complete ranking. Those situations can be either ties, with leads to *partial rankings*, or lack of data, which leads to *incomplete rankings*.

Partial rankings can be represented by *non-strict total orders* over $\mathcal{L}$ [dSAS+18]. The now binary partial order relation, $\succeq$, is Reflexive, Transitive, Antisymmetric and Connected. An an example, the *non-strict total order* $\lambda_1 = \lambda_2 \succ \lambda_3 \succ \lambda_4$ can be represented as $\pi = (1,1,2,3)$.

Incomplete rankings can be represented by *non-strict partial orders* over $\mathcal{L}$ [dSAS+18]. The binary partial order relation, $\succeq$, is Reflexive, Transitive and Antisymmetric. An an example, the *non-strict partial order* $\lambda_1 \succ \lambda_3 \succ \lambda_4$ can be represented as $\pi = (1,0,2,3)$, with 0 meaning that λ_1, λ_3 and λ_4 are incomparable to λ_2 ($\lambda_1, \lambda_3, \lambda_4 \perp \lambda_2$).

3.2 Methods

The existing Label Ranking methods can be divided into two categories, according to how they approach a ranking problem.

Decomposition methods (also known as Reduction techniques) divide the ranking problem into smaller, simpler problems (e.g. PAR, Ranking by Pairwise Comparison [FH11b] and Rule-Based Label Ranking [GSF+12a]). This simplification allows the problem to be treated as a classification problem, which has the advantage of having more efficient implementations[PRdS16]. However, there is the downside of loss of information when decomposing the ranking.

Direct methods deal with the ranking without any decomposition (e.g LRAR, Decision Trees [CHH09, TBD02] and k-Nearest Neighbors [CHH09, BSdC03]). This path can be divided into two approaches. Statistical-based methods (e.g. [CH09]) use the probabilistic distribution of the rankings (e.g Mallow's[LL03]), which has the advantage of providing a reliability score. Similarity-based methods (e.g. LRAR) use measures of correlation (e.g. Kendall's [Ken48]) to calculate the distance between rankings. Tree-based algorithms, for example, recursively partition the data into small subsets, by deciding when to split using a ranking correlation coefficient to measure the similarity between nodes, like Kendall's or Spearman's [Spe04].

Some of the methods in Label Ranking are descriptive methods. This means that they only find patterns in the data, but have not been used for prediction. Examples of that are PAR, Exceptional Preferences Mining [dSDSK16] and Mining rank data [HH14].

There are two main approaches for the discretization of the features in the data: Multi-interval Discretization of Continuous Attributes for Label Ranking [dSSK+13], and Entropy-based Discretization for Ranking (EDiRa) [dSSK16]. Since EDiRa was the newest approach with higher benchmarks, we chose it. EDiRa is based on finding cutting points in the data by computing correlations using the Kendall Tau.

3.2.1 Label Ranking Association Rules

Label Ranking Association Rules (LRAR) [dSSJ+11] can be defined as a direct adaptation of class association rules (CAR) [ML98] for LR:

$$A \rightarrow \pi$$

where $A \subseteq \mathbb{X}$ and $\pi \in \Omega$. The ranking π can be either a complete or a partial ranking.

Direct adaptation of CAR would have two problems First, the number of classes can be very large, which requires a lot of data to learn a new mapping Second, the nature of class prediction is different when using rankings. In classification, an example either belongs to a class or you does not. In LRAR, however, strictly different rankings can share enough similarity that would increase the probability of seeing a certain observation in a similar yet different ranking.

To tackle these problems, interest measures were adapted [dSSJ+11] to be based on similarity rather than frequency.

Support:

$$sup_{lr}(A \to \pi) = \frac{\sum_{i:A \subseteq desc(x_i)} s(\pi_i, \pi)}{n}$$

Where $s(\pi_i, \pi)$ is a similarity function between rankings.

Confidence:

$$conf_{lr}(A \to \pi) = \frac{sup_{lr}(A \to \pi)}{sup(A)}$$

Improvement:

$$imp_{lr}(A \to \pi) = min(conf_{lr}(A \to \pi) - conf_{lr}(A' \to \pi'))$$

With $\forall A' \subset A$ and $\forall(\pi, \pi')$.

Lift:

$$lift_{lr}(A \to \pi) = \frac{sup_{lr}(A \to \pi)}{sup(A) \cdot sup_{lr}(\pi)}$$

3.2.2 Pairwise Association Rules

Pairwise Association Rules(PAR) are an adaptation of Association Rules, where the consequent is a set of pairwise rules, rather than a complete or partial ranking, as is the case of LRAR. PAR can be defined as follows [dSAS+18]:

$$A \to \{\lambda_a \succeq \lambda_b \oplus \lambda_b \succeq \lambda_a \oplus \lambda_a = \lambda_b \oplus \lambda_a \perp \lambda_b | \lambda_a, \lambda_b \in \mathcal{L}\}$$

Contrary to LRAR, the interest measures of *support*, *confidence* and others have the same definition as in AR. This is, while in LRAR the interest measures were adapted to use similarity, in PAR they are based on frequency. See Section 2.2.1.

One of the reasons PAR were proposed was to tackle the problem of not having enough information for a complete or partial ranking as is the case of LRAR [PRdS16]. Let's see for example the following LRAR:

$$\mathcal{A}_1 = high \wedge \mathcal{A}_2 = low \to party.C \succ party.A \succ party.B \succ party.D$$

If there were not enough interesting rules for the model to find, there is a possibility this ranking could not be obtained. If we used PAR however, there is a possibility we could extract information

in the form of pairwise comparisons, as:

$$\mathcal{A}_3 = high \wedge \mathcal{A}_4 = low \rightarrow party.C \succ party.A$$

There are situations, however, where there is sufficient information for a complete ranking but PAR can be used to attempt to achieve an increase in accuracy nonetheless. This leads us to the trade-off in accuracy versus completeness that is made when using one approach or the other.

LRAR generates rankings which gives us a complete information but might not have enough confidence to give us a result, and so it might abstain. PAR give a set of pairwise preferences that, although possibly not as complete as a ranking, still can provide more and very useful information that would have been lost otherwise It is also worth noting that in the limit, if a PAR has enough pairwise comparisons to cover all the labels, it can be translated into a complete ranking.

One last trade-off between both approaches is confidence. Since PAR predict less, it is expected that what is predicted is done with more confidence (because they do not need to find instances to generate an entire ranking). In other words, it is expected that the rules extracted with a PAR model have, in comparison to LRAR, less information but higher confidence. The goal is for the model to fail as little as possible, instead of finding complete rankings.

Until now, PAR were only used as a descriptive tool. However, in this work they will be used for predictive purposes.

There are situations where the model might fail the prediction of a given instance, in which cases a *default ranking* is used. A common choice, is to use the average ranking of the training data.

3.3 Evaluation

The accuracy of a given label ranking prediction $\hat{\pi}$ of a ranking π can be measured with several loss functions on Ω[PRdS16]. Consider $\pi(a)$ the position, or *rank*, of λ_a in π.

Number of discordant label pairs:

$$D(\pi, \hat{\pi}) = \#\{(a,b) | \pi(a) > \pi(b) \wedge \hat{\pi}(a) < \hat{\pi}(b)\}$$

Number of concordant label pairs:

$$C(\pi, \hat{\pi}) = \#\{(a,b) | \pi(a) > \pi(b) \wedge \hat{\pi}(a) > \hat{\pi}(b)\}$$

Kendall's Tau coefficient [Ken48]: normalized difference between the number of concordant C and discordant D pairs:

$$\tau(\pi, \hat{\pi}) = \frac{C - D}{\frac{1}{2}k(k-1)}$$

Gamma coefficient [GK54]: measure of correlation between two incomplete rankings or between incomplete and complete or partial rankings

$$\gamma(\pi, \hat{\pi}) = \frac{C - D}{C + D}$$

The *gamma* coefficient is equivalent to Kendall Tau in the present of strict total orders, since $C + D = \frac{1}{2}k(k-1)$.

3.4 Using PAR for Prediction

The adaptation of PAR for prediction comes in a natural way but still presented some challenges, such as selecting the rules for prediction, evaluating the predictions against complete rankings, and choosing how many rules to predict.

3.4.1 Rule selection

Let us consider the set of all the PAR generated as $\mathcal{R}$. We need to get all the rules from $\mathcal{R}$ that apply to a given an instance x, $applicableRules(x)$. That means all the rules whose antecedent is contained in the descriptors of x.

While there are many ways for defining and selecting the best rules [ML98], we are going to analyze two simple ways. One way is using the average ranking of the $applicableRules(x)$. Another way is using the ranking of the top applicable rule. The top rule is the first one in the ordered list of $applicableRules(x)$. This list is sorted using as first criteria the *confidence* (the higher the better), and as second criteria the *support* (the higher the better). The analysis and choice of our approach will be done during the implementation phase.

For the practical application of PAR, there are two additional measures that need to be defined: the maximum number of pairwise rules in the consequent, and a measure of completeness of the predicted rules.

3.4.2 Maximum Pairwise

The generation of PAR, as a decomposition approach (see Section 3.2), decomposes the rankings into pairwise comparisons. The decomposition of rankings into all possible pairwise comparisons in this context can lead to the generation of many frequent rules.

For a dataset with k labels, the maximum possible number of pairwise comparisons, $max_{pairwise}(k)$, in a rule can be defined as follows:

$$max_{pairwise}(k) = \frac{k!}{2!(k-2)!}$$

This presents a complexity problem for datasets with a large number of labels. Because the maximum possible number of pairwise is a combination without repetition of the number k of

labels the generation of rules becomes computationally heavy. This means that, for a large dataset, for example one with sixteen labels, if we make a loose model tuning, the algorithm could be running for a large amount of time.

To avoid these problems, one of the parameters that needs to be defined in the decomposition method is the maximum number of pairwise, *maxpairs*. This parameter defines how many pairwise comparisons can be present, at most, in the consequent of the generated rules, with $maxpairs \leq max_{pairwise}(k)$.

For example, with $maxpairs = 1$, we could find:

$$\mathcal{A}_1 = high \rightarrow party.B \succ party.A$$

but with $maxpairs = 2$, we could also find:

$$\mathcal{A}_1 = high \rightarrow party.B \succ party.A \wedge party.A \succ party.C$$

3.4.3 Completeness

We propose the measure *completeness* to measure the proportion of pairwise comparisons that were predicted, in comparison to the total number of pairwise comparisons possible. In other words, it reflects how much of the ranking was in fact predicted.

Assuming *pairs* as the number of pairwise comparisons in a rule, *completeness* is defined as:

$$completeness = \frac{pairs}{max_{pairwise}(k)}$$

Chapter 4

Empirical Evaluation

This chapter details the implementation of our experiments regarding the use of Label Ranking, more specifically Pairwise Association Rules, for predicting elections. We are going to first study the behavior of PAR when making predictions in typical LR datasets, then apply the same methods for election prediction. We begin by exploring the data used for our experiments in Section 4.1. In Section 4.2 we detail our experimental setup. Next, we describe the application of PAR for prediction in Section 4.3. Finally, in Section 4.4 we show and analyze the obtained results.

4.1 Data

The data used in this work to test our PAR-based model is a set of eight LR datasets (Section 4.1.1). These datasets, although not related to elections, will serve to validate our model's results before applying it in our case study. The rules' mining and prediction approaches used on these datasets will then be applied in the context of elections.

The elections-related data of this work consists in five datasets (Section 4.1.2), three Portuguese and two German, with socio-economical and voting information about each region of each country in election years. The German datasets comes from a study[BMK+13] where subgroup discovery techniques (as seen in Section 2.1.2) were used to find patterns that would indicate which factors would favor one party over another. Because of the interesting patterns discovered in [BMK+13] and the quality of the datasets, they served as a reference for our case study.

The Portuguese datasets are a subset of PORDATA, a Portuguese institutional database where it can be found a range of all types of social and economic markers pertaining to each region. Our curated subset is based on the format of the German data, although it can be expanded with extra variables.

4.1.1 Label Ranking datasets

To develop and test our approach we used a set of common Label Ranking datasets, which are an adaptation of datasets made available by the UCI repository and the Statlog project [CHH09]. These semi-synthetic datasets were generated from classification (type A) and regression (type B)

datasets. Because they were used in several papers on label ranking methods [CHWW12, CH, RDSK12, GSF+12b, dSAS+18], they give us a good term of comparison for our benchmarks.

Table 4.1: LR datasets

dataset	type	# instances	# features	# labels
bodyfat	B	252	7	7
cpu-small	B	8192	6	5
glass	A	214	9	6
housing	B	506	6	6
iris	A	150	4	3
stock	B	950	5	5
vehicle	A	846	18	4
wine	A	178	13	3

4.1.2 Election datasets

For the problem at hands, the elections datasets we used (Table 4.2) can be divided in two sets: 1) the German datasets, which have data from the years of 2005 and 2009 of German regional elections; 2) the Portuguese datasets, which have data from the years of 2009, 2013 and 2017 of Portuguese regional elections. In the German datasets, the labels represent:

- CDU (conservative)
- SPD (center-left)
- FDP (liberal)
- GREEN (center-left)
- LEFT (left-wing)

In the Portuguese datasets, the labels represent:

- BE (left-wing)
- CDS-PP (conservative)
- PCP-PEV (left-wing)
- PPD/PSD (center-right)
- PS (center-left)

The features of these datasets include socio-economic variables such as Income, GDP growth, Workforce and Unemployment (the complete list can be seen in Appendix A).

Table 4.2: Election Datasets

dataset	# instances	# features	# labels
germany 2005	412	29	5
germany 2009	412	29	5
portugal 2009	308	20	5
portugal 2013	308	20	5
portugal 2017	308	20	5

While the German datasets are originated from [BMK+13], the Portuguese datasets we created by us using the PORDATA[1] database as a source. This database is a Portuguese centralized source of socio-economic and geographical data.

The transformation process for the Portuguese data was based on the features included in the German datasets. We then proceeded to identify which categories in the PORDATA database would better reflect that same information, and we created a automated script to parse PORDATA and fetch the data for the features we wanted in the years of Portuguese regional elections.

These datasets will not be used, however, in the same way as the common LR datasets. They will not be trained and tested individually, with a cross validation technique. Due to the added interest in predicting from one year to another, they are going to be used two at a time, with the older year being used to train the model, and the most recent one to test, as shown in Table 4.3.

An important detail in the generation of the Portuguese datasets is that the socio-economic variables were not always available for the specific year that the dataset refers. In those situations, we use the most up-to-date variables available, up to the year of the dataset.

Table 4.3: Merged Election datasets

merged dataset	train data	test data
germany 2005-2009	germany 2005	germany 2009
portugal 2009-2013	portugal 2009	portugal 2013
portugal 2013-2017	portugal 2013	portugal 2017

In a field where real world label ranking data is scarce [PRdS16], these Portuguese election datasets contribute to existing Label Ranking and Election Prediction literature with three real world datasets (*portugal 2009*, *2013* and *2017*)[2].

4.1.3 Data preparation

To normalize the German and Portuguese data and transform it to be used in our algorithm, we created a script that parsed our raw data to a standard csv. In the Portuguese data, it merged all the categories fetched from PORDATA, and normalized the resulting table to be in the same format as the German. Some absolute values, e.g. the number of employees per sector, were normalized

[1] https://www.pordata.pt/
[2] http://dx.doi.org/10.17632/cxgft7c85x.1

using the total number of inhabitants per region. In the cases where two or more parties are in a coalition, we do not consider that there is an order relation between them i.e. we consider a tie.

We used the EDiRa algorithm (Section 3.2) to discretize the numeric variables of our datasets. In cases where this method did not manage to find cutting points in an attribute, we used an *equal width* discretization. This discretization method creates intervals of equal width in the data, according to the number of intervals (also referred as bins) desired. In our case, we use 4 bins.

4.2 Experimental setup

Traditionally in LRAR, because it uses measures of similarities between rankings, the support is typically 1%. As it was demonstrated empirically in [dSAS+18], a minimum support generates better predictions. Another important reason to use a small minimum *support* is to minimize the use of the default ranking. This is because, lowering the minimum *support* threshold potentiates the generation of more rules. So the minimum support was set to 1%.

To define the value of minimum *confidence*, we assume that we would like to predict with high confidence. This is because we are going to predict an incomplete ranking, hence the smaller output given must have a high *confidence*. However, the rule selection method for the prediction part might mitigate a lower initial threshold (see Section 4.3). Because of this, we do not need to limit too much the *confidence* and reduce our rules generation, and so after some tests we defined the minimum *confidence* at 70%.

Improvement is used to prune rules that do not increase the confidence of sub-rules that share the same consequent by a specific amount. However, in our case we do not want to limit the generation of rules that could potentially have a higher completeness, even if the confidence is the same. Disabling the *improvement* allows the model to generate sub-rules even if there is no gain in confidence, which will be useful to study the impact of the number of comparisons in the prediction results (see Section 4.4). For this reason, we defined the minimum *improvement* as 0.

The rules generation process is divided in two parts: the mining and the post-processing. For the rules mining part of our experiments, we are going to use the CAREN [AJ10] software[3]. CAREN uses a rule-based (see Section 2.2.2) bitwise depth-first frequent pattern mining algorithm to generate the association rules. Since CAREN yields significant rules [Web06], the Fisher test is used to evaluate when a generalization of a rule is significant or if it should be discarded instead. The pruning of the rules uses the Fisher exact test [AJ10].

PAR that are obtained from CAREN, come with a set of pairwise comparisons in the consequent. Since these are not guaranteed to be sufficient to generate a complete ranking, we might not be able to represent them as a permutation. For this reason we propose to use a pairwise ranking matrix, as in [HFCB08].

[3] http://www4.di.uminho.pt/~pja/class/caren.html

As an example, one of the rules consequent could be

$$\lambda_2 \succ \lambda_1 \wedge \lambda_2 \succ \lambda_4$$

For that, we use a matrix to represented the preferences. Each line represents the preferences between label λ_l and all the labels in the ranking. Therefore, each entry represents a pairwise preference between two labels, λ_l and the column label λ_c. We assume 1 when $\lambda_l > \lambda_c$, -1 when $\lambda_l < \lambda_c$ and 0 when $\lambda_l = \lambda_c$. We also use "NA" when there is no information.

-	λ_1	λ_2	λ_3	λ_4	λ_5
λ_1	0	-1	NA	NA	NA
λ_2	1	0	NA	1	NA
λ_3	NA	NA	0	NA	NA
λ_4	NA	-1	NA	0	NA
λ_5	NA	NA	NA	NA	0

4.3 Prediction with PAR

As discussed in Section 3.4.1, there are different approaches to aggregate the predictions of rules. In this work we tested two simple approaches 1) use the *average ranking* of all the applicable rules or 2) using the *ranking* of the top applicable rule.

Preliminary results showed us that taking option 1 leads us to approximate our results to the baseline. We believe that this can be explained by the fact that using all the possible rules, their average ranking tends to be a complete ranking, similar to the default ranking. Therefore, in our opinion, it does not emphasize the advantage of PAR. This approach is not interesting for us, since what we are trying to achieve is higher accuracy in the predictions, even if that means a lower completeness.

Reaching option 2, our tests showed that using only the rule with the highest *support* and *confidence* yields accurate results, in spite of a lower completeness.

With the list of $applicableRules(x)$, we need to fetch the best rule of them all. Ordering it by *support* and *confidence*, with *confidence* as the primary sorting method, the best rule, $bestRule$, is the first element of the list. That best rule will be used to the prediction of the x instance.

4.3.1 Evaluation

Since our model sacrifices completeness to accuracy, we wanted to use, on top of *gamma*, a measure of accuracy that is more penalizing. To achieve that, we propose the use of an accuracy measure similar to the one used in classification problems.

The accuracy measure, as it is used in classification problems, is 1 if the prediction is completely correct, and 0 if incorrect. The most relevant part to take into account is that in the Label Ranking context, it will be also 0 even if the ranking is partially correct. That way we penalize any errors or flaws. This measure would then give us a strict accuracy, instead of measuring until

what point is our prediction similar to reality, as is the case of *gamma*. In this way, we can assess which percentage of rules correctly predicted the examples.

In Table 4.4 we can see an example of the behavior of this accuracy in comparison to Kendall Tau and gamma.

Table 4.4: Accuracy measures comparison

π	$\hat{\pi}$	Kendall Tau	gamma	accuracy
$\lambda_1 \succ \lambda_2 \succ \lambda_3 \succ \lambda_4 \succ \lambda_5$	$\lambda_1 \succ \lambda_2 \succ \lambda_3 \succ \lambda_4 \succ \lambda_5$	1	1	1
$\lambda_1 \succ \lambda_2 \succ \lambda_3 \succ \lambda_4 \succ \lambda_5$	$\lambda_1 \succ \lambda_2 \succ \lambda_3 \succ \boldsymbol{\lambda_5 \succ \lambda_4}$	0.8	0.8	0
$\lambda_1 \succ \lambda_2 \succ \lambda_3 \succ \lambda_4 \succ \lambda_5$	$\boldsymbol{\lambda_5 \succ \lambda_4 \succ \lambda_3 \succ \lambda_2 \succ \lambda_1}$	-1	-1	0
$\lambda_1 \succ \lambda_2 \succ \lambda_3 \succ \lambda_4 \succ \lambda_5$	$\lambda_1 \succ \lambda_2 \wedge \lambda_2 \succ \lambda_3 \wedge \lambda_3 \succ \lambda_4$	NA	1	1
$\lambda_1 \succ \lambda_2 \succ \lambda_3 \succ \lambda_4 \succ \lambda_5$	$\boldsymbol{\lambda_2 \succ \lambda_1} \wedge \lambda_2 \succ \lambda_3 \wedge \lambda_3 \succ \lambda_4$	NA	0.33(3)	0
$\lambda_1 \succ \lambda_2 \succ \lambda_3 \succ \lambda_4 \succ \lambda_5$	$\boldsymbol{\lambda_2 \succ \lambda_1} \wedge \lambda_2 \succ \lambda_3 \wedge \boldsymbol{\lambda_4 \succ \lambda_3}$	NA	-0.33(3)	0

The implementation of the *gamma* coefficient can be seen in Appendix B. The implementation of the classification *accuracy* can be seen in Appendix C.

4.4 Results and Analysis

We will first analyze the LRAR and PAR prediction results using the common LR datasets to validate our model. We will also compare our results against the methods based on the Plackett-Luce model [CDH10], as well as two other these used for comparison.

Following that, we will apply LRAR and PAR to our election datasets.

4.4.1 Experimental results with typical LR datasets

In this experimental part we study the behavior of LRAR as well as PAR when making predictions in the standard LR datasets. In this experimental part, the models were evaluated using a 10-fold cross-validation technique. We trained the model with 10% of the data and tested it in the remaining 90%. This was done 10 times for each 10% block of the data. The presented results are the average of all individual results.

To compare our method against state of the art approaches, we used the results of the methods based on the Plackett-Luce(PL) model in [CDH10]. Those methods are: instance-based approach to LR with PL(IB-PL) and generalized linear approach to LR with PL (Lin-PL). We will also compare with two other methods used in the same paper: instance-based approach using the Mallows model (IB-Mal), and loglinear models for LR (Lin-LL). We note that, since we are simply using the reported results, these are only indicative as they were not tested under the same conditions.

As we have seen in Sections 3.4.2 and 3.4.3, the maximum number of pairwise comparisons in the consequent of the rule directly affects the completeness of the prediction. Intuitively, we expected that the accuracy drops as we increase the completeness (i.e. increase the *maxpairs*).

This leads to the question of how much pairwise comparisons do we want to generate and use. In other words, how much can the *maxpairs* parameter affect our accuracy. Because this not an

Table 4.5: Experimental results using the common LR datasets

	baseline	IB-PL	IB-Mal	Lin-PL	Lin-LL	LRAR	1 Pair	2 Pairs	3 Pairs	4 Pairs
bodyfat	-.04	.23	.23	.27	.27	.02	.33	.19	-.37	-.58
cpu-small	.23	.50	.50	.43	.42	.45	.81	.74	.61	.33
glass	.68	.84	.84	.83	.82	.71	.97	.98	.95	.94
housing	.05	.71	.74	.66	.63	.69	1	.99	.97	.91
iris	.09	.96	.93	.83	.82	.86	.95	.95	.89	-
stock	.06	.92	.93	.71	.70	.82	.99	.99	.99	.95
vehicle	.18	.86	.86	.84	.78	.82	.95	.95	.94	.83
wine	.33	.95	.94	.95	.94	.92	.88	.87	.5	-

answer we can respond a priori since it will possibly depend on the dataset, what we can study is the behavior of the prediction accuracy across a predefined range of *maxpairs*. That range must be wide enough to demonstrate the trade-off between completeness versus accuracy/gamma.

We will test different *maxpairs* values to demonstrate the trade-off in different datasets.

In Table 4.5 we can see the results obtained for each dataset, where "1 Pair" stands for using PAR with *maxpairs* = 1, and so on. However, since we wanted to test the impact of the number of pairwise comparisons of a rule in the prediction accuracy, in "1 Pair" only rules with 1 comparison in the consequent are used, and so on. The pruning was done outside CAREN because *maxpairs* only sets the *maximum* comparisons, not the exact number. For the case of LRAR the value is Kendall's tau coefficient, and for PAR it is the gamma. As we have seen in section 3.3, they are equivalent.

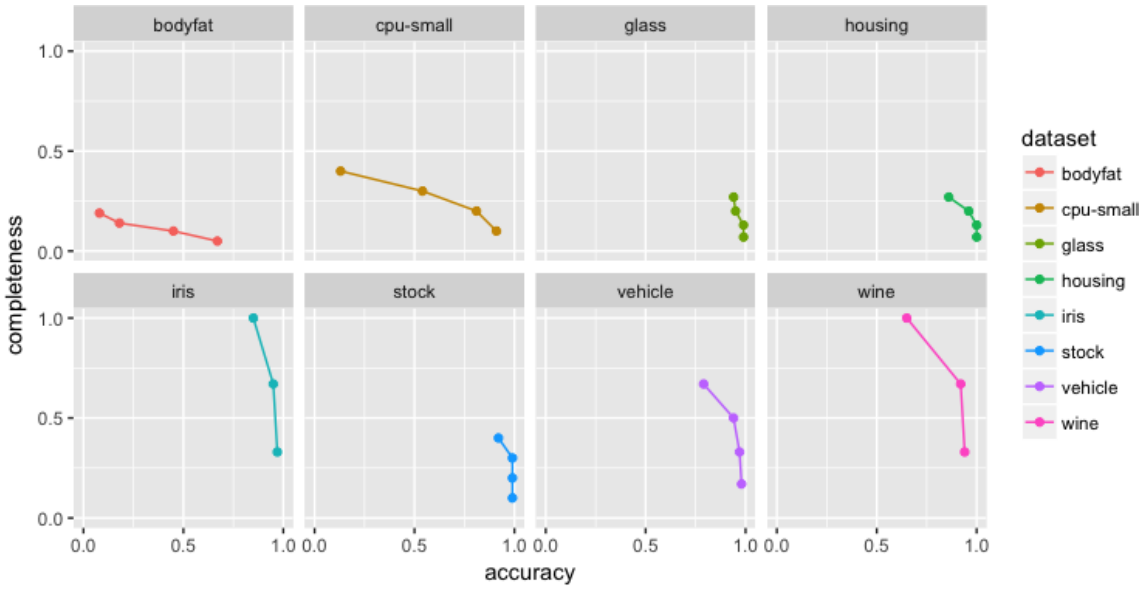
We tested PAR with "Pairs" ranging from 1 to 4. Some datasets were, however, limited by their smaller number of labels. For that reason, some of these datasets do not have more than 3 possible maximum pairwise rules.

Using LRAR as a baseline, it is possible to see that PAR are always better when using "1 Pair". The only case where that does not happen is in *wine*, where LRAR has better results. For example, in *cpu-small* there was a big improvement. However, we note that the model is predicting only one pairwise comparison. This behavior was expected and in fact helps us achieve a goal of this work, which is to suggest that PAR, in spite of predicting less, can be more certain of what they predict in certain situations.

Looking at the PAR results alone, we can also see the tendency of, generally, the gamma decreasing when "Pairs" increases. This happens because the less the rankings have to be complete, the higher chance there is of it finding rules with higher support and confidence. That means that, with PAR, we can tune the model accordingly to what is more important to us regarding the information, quantity(completeness) or quality(gamma/ accuracy).

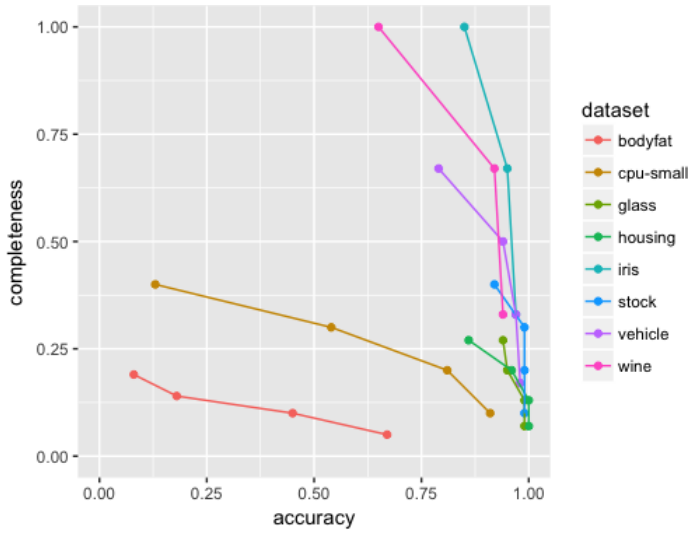
To explore more the impact of "Pairs", lets look at Figure 4.1. However, instead of using gamma here, we opted to use the classical classification accuracy measure, since it better represents what we are trying to illustrate(see section 4.3.1). This set of graphs allows us to see the curve that represents the aforementioned trade-off between the completeness and accuracy. The less we attempt to predict the more accurate it is.

Figure 4.1: Completeness vs Accuracy in the LR datasets



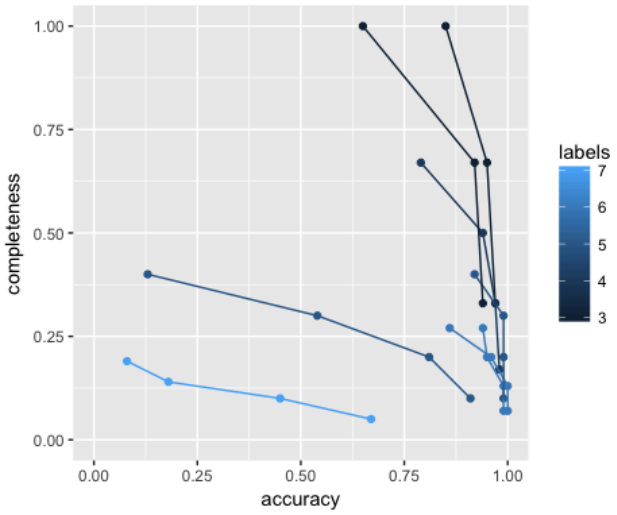
Looking at Figure 4.2 we can see all the completeness versus accuracy curves merged in the same graph. This, apart from making the trade-off tendency more visually clear, allowed us to spot groups of datasets. To analyze this we grouped the datasets by number of labels.

Figure 4.2: Completeness vs Accuracy in the LR datasets - Merged



In Figure 4.3 we can now see the datasets grouped by their number of labels. This graph shows how datasets with the same number of labels tend to have similar accuracy and completeness outputs across the range of "Pairs". This seems to indicate that if the size of the rankings is small, the accuracy is not very affected. One example of this is the behavior of the "iris" dataset observed in Figure 4.1. Still, the number of observations in this study are not enough for a definitive conclusion.

Figure 4.3: Completeness vs Accuracy in the LR datasets grouped by number of labels



4.4.2 Experimental results with LR Elections datasets

With this experiments we would like to know which socio-economic indicators affect the votes of the different parties. For that we use Label Ranking Association Rules, to look for patterns of the type:

$$\mathcal{A}_1 = low \wedge \mathcal{A}_2 = high \rightarrow party.C \succ party.A \succ party.B \succ party.D$$

and Pairwise Association Rules, which will find patterns like:

$$\mathcal{A}_1 = low \wedge \mathcal{A}_2 = high \rightarrow party.C \succ party.A \wedge party.C \succ party.B$$

Each transaction of the data represents a geographical region, which includes information about the number of votes and socio-economic metrics. We study the behavior of PAR for mining and predicting electoral results and compare to LRAR. We will use them to identify meaningful association rules that relate the socio-economic indicators with the preferences of the inhabitants of certain regions.

In this section we present the results obtained with PAR on election data. In terms of model evaluation, in spite of using cross validation for the LR datasets, we used the merged elections datasets of two years. That is because, in this case it is more interesting to study the training of the model on the previous election year, and testing it in the year we are trying to predict.

In Table 4.6 we can see the results of applying the LRAR-based and PAR-based models to the elections datasets. As we can see, all the cases display an accuracy improvement when using PAR. For example, in *portugal 2009-2013* there was a 0.32 gamma improvement, with the respective completeness trade-off.

To add a terms of comparison to our results, we included the baseline score as well. This baseline was achieved by using the average of the election results of the train data and applying it

on all the instances of the test data.

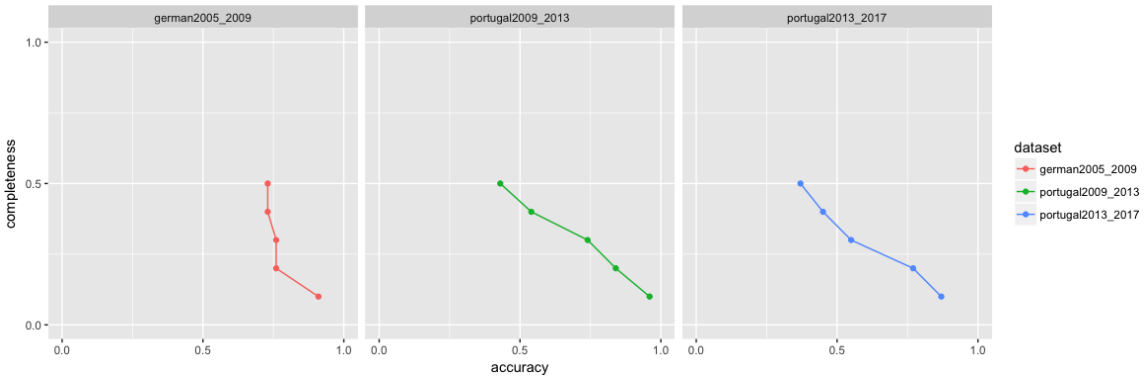
The reason the baseline is relatively high, being higher or on pair with the LRAR results, is because of the nature of elections. There are parties that have a high representation in the voting population, and those parties almost always stay in front of other small parties. For example, PS, a Portuguese center-left party, almost always is in front of BE, a left-wing party. That fact does not change from year to year, which leads to high baseline results. However, is is interesting to note that PAR still offer an interesting accuracy increase, comparing to both the LRAR and the baseline results. That happens because both the baseline and LRAR are based on complete rankings, while PAR decomposes those rankings into easier to accurately predict pairwise comparisons.

Table 4.6: Experimental results using the elections datasets

	baseline	LRAR	1 Pair	2 Pairs	3 Pairs	4 Pairs	5 Pairs
germany 2005-2009	.72	.64	.82	.75	.84	.86	.89
portugal 2009-2013	.65	.6	.92	.82	.79	.66	.67
portugal 2013-2017	.44	.47	.75	.75	.63	.63	.62

After analyzing the accuracy versus completeness evolution in the LR datasets results, we expected that the same behavior would occur on the elections datasets. Inspecting Figure 4.4 we identify the expected curve of increased accuracy on decreasing the completeness.

Figure 4.4: Completeness vs Accuracy in the Elections datasets



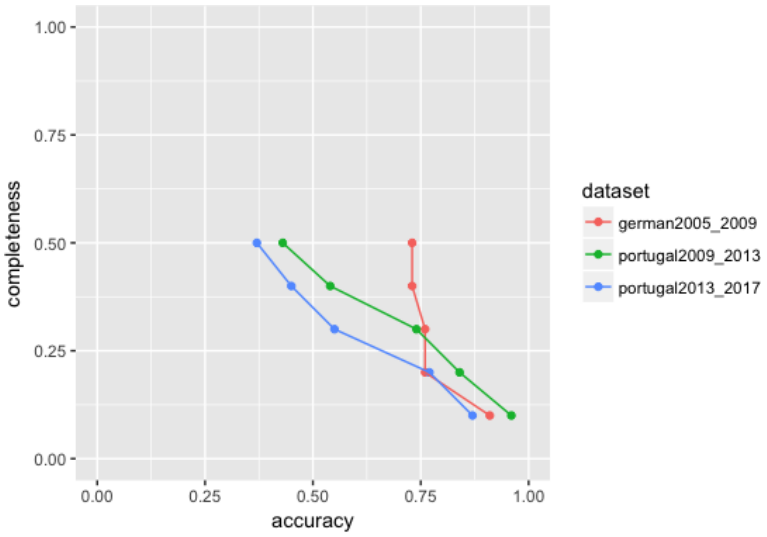
At the same time, if we merge the datasets in one graph as it is done in Figure 4.5, that seems to support what we found on Figure 4.3: datasets seem to have similar behavior according to their number of labels.

With these prediction results suggesting the increased accuracy in using PAR for elections, in the next section we will explore the descriptiveness of the pairwise rules used.

4.4.3 Rule Analysis

One of the advantages of using a descriptive method like PAR for prediction is that in the end we can analyze the rules and understand why the predictions were made. The simple descriptiveness of the rules makes it easy for a person of any background understand the output of our model.

Figure 4.5: Completeness vs Accuracy in the Elections datasets - Merged



In Table 4.7 we have a subset of the socio-economic variables used by our model. To exemplify the clarity of reading a rule, let's look at a real example. For a representing the discretized interval between 12% and 17% in $\mathcal{A}_1$:

$$\mathcal{A}_1 = a \rightarrow PS \succ BE(conf = 90\%)$$

This is easily read as: if the youth unemployment is between 12% and 17% the party PS will be preferred to the party BE in 90% of the cases.

Table 4.7: Subset of Socio-economic variables of the election datasets

id	description
$\mathcal{A}_1$	Unemployment below 25 years
$\mathcal{A}_2$	Residents between 15 and 64 years
$\mathcal{A}_3$	Elderly Population
$\mathcal{A}_4$	Average Monthly Income
$\mathcal{A}_5$	Employees in Construction
$\mathcal{A}_6$	Employees in Administration and Services
$\mathcal{A}_7$	Employees in Agriculture
$\mathcal{A}_8$	Employees in Transformation Industries
$\mathcal{A}_9$	Employees in Gross and Retails Markets
$\mathcal{A}_{10}$	Employees in Banks
$\mathcal{A}_{11}$	Employees in other Economic Sectors
$\mathcal{A}_{12}$	Population without Education
$\mathcal{A}_{13}$	Registered Web Domains
$\mathcal{A}_{14}$	Population with no school degree

To explore the rules behind the PAR model, Table 4.8 displays the top five rules found for each elections dataset. For this case study we used "2 Pairs". For readability purposes, the antecedent was reverse discretized.

Table 4.8: Top 5 Rules analysis

	antecedent	consequent	*%sup*	*%conf*
germany 2005	$\mathcal{A}_{13} > 74 \wedge \mathcal{A}_{14} < 10\%$	SPD $\succ$ FDP $\wedge$ SPD $\succ$ LEFT	81.31	100
	$\mathcal{A}_{13} > 74 \wedge \mathcal{A}_{14} < 10\%$	CDU $\succ$ FDP $\wedge$ SPD $\succ$ LEFT	81.31	100
	$\mathcal{A}_{13} > 74 \wedge \mathcal{A}_{14} < 10\%$	SPD $\succ$ GREEN $\wedge$ SPD $\succ$ LEFT	81.31	100
	$\mathcal{A}_{13} > 74 \wedge \mathcal{A}_{14} < 10\%$	CDU $\succ$ GREEN $\wedge$ SPD $\succ$ LEFT	81.31	100
	$\mathcal{A}_3 < 30\%$	SPD $\succ$ FDP $\wedge$ SPD $\succ$ LEFT	80.58	100
portugal 2009	$\mathcal{A}_5 \in [14\%, 27\%] \wedge \mathcal{A}_{11} < 1\%$	PS $\succ$ BE $\wedge$ PS $\succ$ CDS.PP	42.53	100
	$\mathcal{A}_{12} < 9\% \wedge \mathcal{A}_6 < 7\% \wedge \mathcal{A}_{11} < 1\%$	PS $\succ$ BE $\wedge$ PS $\succ$ PCP.PEV	16.88	100
	$\mathcal{A}_{12} \in [15\%, 20\%] \wedge \mathcal{A}_7 < 13\%$	PS $\succ$ BE $\wedge$ PS $\succ$ PCP.PEV	15.58	100
	$\mathcal{A}_2 > 66\% \wedge \mathcal{A}_6 < 7\% \wedge \mathcal{A}_{11} < 1\%$	PS $\succ$ BE $\wedge$ PS $\succ$ PCP.PEV	15.58	100
	$\mathcal{A}_{10} \in [1\%, 3\%] \wedge \mathcal{A}_5 \in [14\%, 27\%]$	PS $\succ$ CDS.PP $\wedge$ PS $\succ$ PCP.PEV	10.06	100
portugal 2013	$\mathcal{A}_5 < 14\% \wedge \mathcal{A}_6 < 7\% \wedge \mathcal{A}_{10} < 1\%$	PCP.PEV $\succ$ BE $\wedge$ PS $\succ$ BE	24.03	100
	$\mathcal{A}_1 \in [12\%, 17\%] \wedge \mathcal{A}_8 < 16\%$ $\wedge \mathcal{A}_9 \in [17\%, 27\%]$	PCP.PEV $\succ$ BE $\wedge$ PS $\succ$ BE	18.18	100
	$\mathcal{A}_4 \in [780, 1000] \wedge \mathcal{A}_8 < 16\%$ $\wedge \mathcal{A}_9 \in [17\%, 27\%]$	PCP.PEV $\succ$ BE $\wedge$ PS $\succ$ BE	17.53	100
	$\mathcal{A}_{12} < 9\% \wedge \mathcal{A}_6 < 7\%$	PSD $\succ$ BE $\wedge$ PS $\succ$ BE	17.21	100
	$\mathcal{A}_8 \in [16\%, 31\%] \wedge \mathcal{A}_4 \in [1000, 1730]$	PSD $\succ$ BE $\wedge$ PS $\succ$ BE	16.88	100

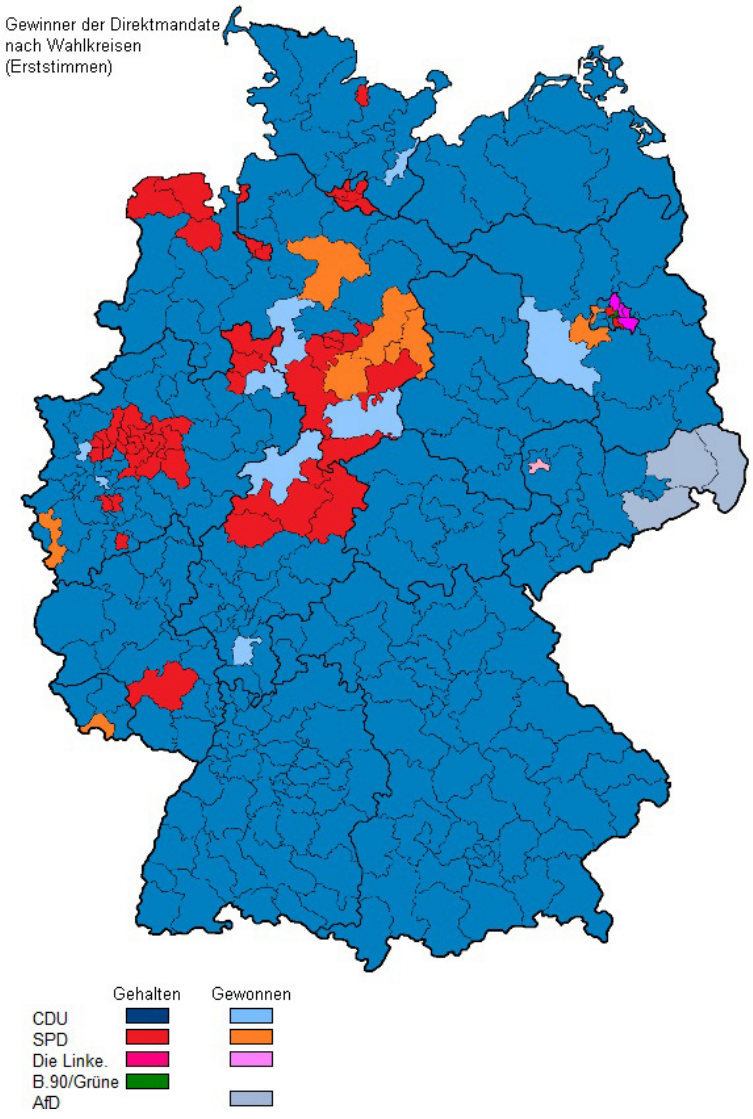
Looking at the table, it is possible to see that some of these rules could be aggregated by their antecedent. An obvious example are the first four rules generated from the *germany 2005* dataset. This antecedent aggregation was not explored in this work but lay an interesting ground for future developments.

Another detail that can be observed is the unusually high *support*, compared to typical LR problems, of the German rules. However, looking at Figure 4.6, we can see that the German regional elections reality is very homogeneous. This map show the evolution from 2009 to 2017 of the elected parties in each region, with the regions held ("Gehalten") and the regions won ("Gewonnen"). We can see that there are a multitude of regions that can support rules that say CDU or SPD are preferred to other small parties.

A different situation is the mining of relatively obvious rules. As it was referred in the previous section when discussing the high baseline for election datasets, the PS $\succ$ BE is a rule that is going to be discovered several times due to PS being a much bigger party than BE. One way of minimizing this situation is using techniques like Subgroup discovery, as it was done in [BMK+13], to find deviating patterns. However, it is worth noting that there are situations where PS $\succ$ BE are not such obvious conclusions. Examples of those situations are rankings where they are in close situations, like PS is first and BE is second, or PS is forth and BE is last.

[4] https://www.wahlen-in-deutschland.de/bubtwvgl091317.htm

Figure 4.6: German regional elections map 2009 to 2017[4]



Chapter 5

Conclusions

In this dissertation, we proposed to study the problem of Election Prediction with Label Ranking. In particular, we tackled it with a new approach: Pairwise Association Rules (PAR). We present an empirical study of PAR, not only on election datasets, but also on benchmark Label Ranking datasets. We wanted to understand if we could, in some way, force the model to keep only the preferences from which it was more certain. The results were compared with state of the art algorithms in the field.

The analysis of the experimental results shows that in all datasets the completeness versus accuracy trade-off is clear. Figures 4.1 and 4.4 illustrate the impact of the number of pairwise comparisons. It is possible to see the tendency for the accuracy of the model to decrease when we increase the completeness. In general the highest accuracy is obtained when predicting the order between two labels ("2 Pair").

One interesting observation that can be drawn from Figure 4.3 is that the aforementioned behavior seem to have a strong relation with the size of the rankings. That is, if the size of the rankings is small, the accuracy is not very affected. However, the number of datasets is not high enough for a definitive conclusion.

Having tested our PAR prediction approach in the LR datasets, we tested in election datasets the same experimental setup. These datasets contained socio-economical variables of the different countries' regions, as well as the parties results. The only difference was that the model was trained with the data from the previous elections, the ones immediately before the year to be predicted.

We verified that the election datasets experienced a similar behavior, in comparison to the typical LR datasets. However, further analysis of the results reveals limitations in this approach.

5.1 Limitations

When analyzing in depth the predictions of the 2017 Portuguese regional elections using "2 Pairs"(with two pairwise comparisons in the consequent) we observe a gamma of 0.75. In de-

tail, the model made 78.57% accurate predictions in a universe of 308 regions. In spite of the seemingly high numbers, the reality is that most predictions were, in the current context, considerably obvious. For example, is was predicted that in 85.71% percent of the cases *Partido Socialista (PS)* (a center left party) would have more votes than *Bloco de Esquerda (BE)* (a left party). Knowing the Portuguese political reality, this is an obvious prediction to make and does not represent in most cases much added value. On the other hand, precisely because it found obvious information, these results validate our approach.

This suggests that this PAR approach, while having it's viability validated by the results in the LR datasets, might not be best suited for this specific scenario of elections. Parties that have a disproportionate representation in the voting population lead to unbalanced ranking data. A model that focus more on preference shifts might be more successful in discovering non-obvious patterns.

Another situation that limits this approach is the coalitions between parties. These situations are very problem-specific and, in spite of us handling them as draws, it is arguable if it is a good approach..

A different type of limitations is related to the processes of discretization, and mining of LRAR and PAR. These processes are computationally heavy, scaling with the number of features and size of rankings, which limited us in our choice of LR datasets to perform our tests.

5.2 Accomplished goals and Contributions

Despite the limitations of our approach in election outcome prediction, we successfully show that PAR are a solid alternative to LRAR in solving complex LR problems.

PAR showed that, in general, the less information it is required, the better it is at predicting without mistakes. This strongly suggests that in situations where it is more important to have an accurate rather than more complete prediction, PAR can provide a meaningful contribution.

We also made a contribution to the existing Label Ranking community by contributing with three real-world datasets. The Portuguese socio-economic datasets that were created are hopefully a welcome addition to a field where semi-synthetic datasets are common but real-world datasets are not.

5.3 Future Work

As a future work, we believe that it would be interesting to study the use PAR-based prediction models in other real world examples that could better suit their application. This is because the PAR had very interesting academic results and can very well have other applications rather than election prediction where it would bring more added value. There could also be a study of the situations, in terms of dataset characteristics, where PAR are better suited than LRAR.

In terms of interest measures and pruning techniques, since PAR have to deal with multiple items in the consequent, future research might aid the generation and filtering of such rules.

Another situation that could be interesting in terms of dataset study would be exploring the accuracy and completeness versus number of labels (Figure 4.3) in PAR. Studying this phenomena with more datasets could tell us if indeed datasets with the same number of labels exhibit similar accuracy behavior.

Something else also worth exploring is optimizing the PAR-based model by aggregating rules by their consequent. This would be even more valuable in a case where we do not want to use just be best rule for prediction, as we do, but rather a set of top rules.

One different research path could be exploring the ranking entropy of the mined PAR, while varying the maximum number of pairwise comparisons, against the accuracy drop, similar to what was done for LRAR in [dSAS+18].

Lastly, for all the mentioned possible future studies, it would be pivotal to improve the computational time of the discretization and rules mining processes, possibly with more efficient algorithms. This would allow us to perform tests across a variety of more complex LR datasets, which would possibly yield more comprehensive results.

Conclusions

References

[AJ10] Paulo J Azevedo and Alípio Mário Jorge. Ensembles of jittered association rule classifiers. *Data Mining and Knowledge Discovery*, 21(1):91–129, 2010.

[Arm01] J S Armstrong. *Principles of Forecasting: A Handbook for Researchers and Practitioners*. International Series in Operations Research & Management Science. Springer, 2001.

[AS94a] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc.

[AS+94b] Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.

[BAG99] Roberto J Bayardo, Rakesh Agrawal, and Dimitrios Gunopulos. Constraint-based rule mining in large, dense databases. In *Data Engineering, 1999. Proceedings., 15th International Conference on*, pages 188–197. IEEE, 1999.

[Bar12] E. Barker. *The Political Thought of Plato and Aristotle*. Dover Publications, 2012.

[BMK+13] Mario Boley, Michael Mampaey, Bo Kang, Pavel Tokmakov, and Stefan Wrobel. One click mining. *Proceedings of the ACM SIGKDD Workshop on Interactive Data Exploration and Analytics - IDEA '13*, pages 27–35, 2013.

[BMUT97] Sergey Brin, Rajeev Motwani, Jeffrey D Ullman, and Shalom Tsur. Dynamic itemset counting and implication rules for market basket data. In *ACM SIGMOD Record*, volume 26, pages 255–264. ACM, 1997.

[BSdC03] Pavel B. Brazdil, Carlos Soares, and Joaquim Pinto da Costa. Ranking learning algorithms: Using ibl and meta-learning on accuracy and time results. *Machine Learning*, 50(3):251–277, Mar 2003.

[CDH10] Weiwei Cheng, Krzysztof Dembczyński, and Eyke Hüllermeier. Label ranking methods based on the plackett-luce model. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 215–222, USA, 2010. Omnipress.

[CH] W Cheng and E Hüllermeier. Label ranking with abstention: Predicting partial orders by thresholding probability distributions (extended abstract). computing research repository, corr abs/1112.0508 (2011). *URL http://arxiv. org/abs/1112.0508*.

REFERENCES

[CH09] Weiwei Cheng and Eyke Hüllermeier. A new instance-based label ranking approach using the mallows model. In *International Symposium on Neural Networks*, pages 707–716. Springer, 2009.

[CHH09] Weiwei Cheng, Jens Hühn, and Eyke Hüllermeier. Decision tree and instance-based learning for label ranking. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 161–168, New York, NY, USA, 2009. ACM.

[CHWW12] Weiwei Cheng, Eyke Hüllermeier, Willem Waegeman, and Volkmar Welker. Label ranking with partial abstention based on thresholded probabilistic models. In *Advances in Neural Information Processing Systems*, pages 2501–2509, 2012.

[CLOP15] Mauro Coletto, Claudio Lucchese, Salvatore Orlando, and Raffaele Perego. Electoral Predictions with Twitter: A Machine-Learning approach. In *IIR*, 2015.

[dSAS+18] Cláudio Rebelo de Sá, Paulo Azevedo, Carlos Soares, Alípio Mário Jorge, and Arno Knobbe. Preference rules for label ranking: Mining patterns in multi-target relations. *Information Fusion*, 40:112 – 125, 2018.

[dSDSK16] Cláudio Rebelo de Sá, Wouter Duivesteijn, Carlos Soares, and Arno Knobbe. Exceptional preferences mining. In *International Conference on Discovery Science*, pages 3–18. Springer, 2016.

[dSSJ+11] Cláudio Rebelo de Sá, Carlos Soares, Alípio Mário Jorge, Paulo Azevedo, and Joaquim Costa. Mining association rules for label ranking. In Joshua Zhexue Huang, Longbing Cao, and Jaideep Srivastava, editors, *Advances in Knowledge Discovery and Data Mining*, pages 432–443, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[dSSK+13] Cláudio Rebelo de Sá, Carlos Soares, Arno J. Knobbe, Paulo J. Azevedo, and Alípio Mário Jorge. Multi-interval discretization of continuous attributes for label ranking. In *Discovery Science - 16th International Conference, DS 2013, Singapore, October 6-9, 2013. Proceedings*, pages 155–169, 2013.

[dSSK16] Cláudio Rebelo de Sá, Carlos Soares, and Arno J. Knobbe. Entropy-based discretization methods for ranking data. *Inf. Sci.*, 329:921–936, 2016.

[FH11a] Johannes Fürnkranz and Eyke Hüllermeier. *Preference Learning: An Introduction*, pages 1–17. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[FH11b] Johannes Fürnkranz and Eyke Hüllermeier. *Preference Learning and Ranking by Pairwise Comparison*, pages 65–82. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[GK54] Leo A Goodman and William H Kruskal. Measures of association for cross classifications. *Journal of the American Statistical Association*, 49(268):732–764, 1954.

[GK93] Andrew Gelman and Gary King. Why are american presidential election campaign polls so variable when votes are so predictable? *British Journal of Political Science*, 23(4):409–451, 1993.

REFERENCES

[GSF+12a] Massimo Gurrieri, Xavier Siebert, Philippe Fortemps, Salvatore Greco, and Roman Słowiński. Label ranking: A new rule-based label ranking method. In Salvatore Greco, Bernadette Bouchon-Meunier, Giulianella Coletti, Mario Fedrizzi, Benedetto Matarazzo, and Ronald R. Yager, editors, *Advances on Computational Intelligence*, pages 613–623, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[GSF+12b] Massimo Gurrieri, Xavier Siebert, Philippe Fortemps, Salvatore Greco, and Roman Słowiński. Label ranking: A new rule-based label ranking method. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 613–623. Springer, 2012.

[HFCB08] Eyke Hüllermeier, Johannes Fürnkranz, Weiwei Cheng, and Klaus Brinker. Label ranking by learning pairwise preferences. *Artificial Intelligence*, 172(16):1897 – 1916, 2008.

[HH14] Sascha Henzgen and Eyke Hüllermeier. Mining rank data. In *International Conference on Discovery Science*, pages 123–134. Springer, 2014.

[HPYM04] Jiawei Han, Jian Pei, Yiwen Yin, and Runying Mao. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data mining and knowledge discovery*, 8(1):53–87, 2004.

[JSZ15] J Joo, F F Steen, and S C Zhu. Automated Facial Trait Judgment and Election Outcome Prediction: Social Dimensions of Face. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3712–3720, 2015.

[Ken48] Maurice George Kendall. Rank correlation methods. 1948.

[KK06] Sotiris Kotsiantis and Dimitris Kanellopoulos. Association rules mining: A recent overview. *GESTS International Transactions on Computer Science and Engineering*, 32(1):71–82, 2006.

[LB90] M S Lewis-Beck. *Economics and Elections: The Major Western Democracies*. University of Michigan Press, 1990.

[LB05] Michael S Lewis-Beck. Election Forecasting: Principles and Practice. *The British Journal of Politics & International Relations*, 7(2):145–164, 2005.

[LBT12] Michael S Lewis-Beck and Charles Tien. Election Forecasting for Turbulent Times. *PS: Political Science & Politics*, 45(4):625–629, 2012.

[LHM98] Bing Liu, Wynne Hsu, and Yiming Ma. Integrating Classification and Association Rule Mining. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, KDD'98, pages 80–86. AAAI Press, 1998.

[LL03] Guy Lebanon and John D Lafferty. Conditional models on the ranking poset. In *Advances in Neural Information Processing Systems*, pages 431–438, 2003.

[Mil97] Gary J. Miller. The Impact of Economics on Comtemporary Political Science. *Journal of Economic Literature*, 35(3):1173–1204, 1997.

[ML98] Bing Liu Wynne Hsu Yiming Ma and Bing Liu. Integrating classification and association rule mining. In *Proceedings of the fourth international conference on knowledge discovery and data mining*, 1998.

REFERENCES

[MMGA11] Panagiotis T Metaxas, Eni Mustafaraj, and Dani Gayo-Avello. How (not) to predict elections. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 165–171. IEEE, 2011.

[NVHT12] Loan T T Nguyen, Bay Vo, Tzung-Pei Hong, and Hoang Chi Thanh. Classification based on association rules: A lattice-based approach. *Expert Systems with Applications*, 39(13):11357–11366, 2012.

[Omi03a] E. R. Omiecinski. Alternative interest measures for mining associations in databases. *IEEE Transactions on Knowledge and Data Engineering*, 15(1):57–69, Jan 2003.

[Omi03b] Edward R Omiecinski. Alternative interest measures for mining associations in databases. *IEEE Transactions on Knowledge and Data Engineering*, 15(1):57–69, 2003.

[PCY95a] Jong Soo Park, Ming-Syan Chen, and Philip S Yu. *An effective hash-based algorithm for mining association rules*, volume 24. ACM, 1995.

[PCY95b] Jong Soo Park, Ming-Syan Chen, and Philip S Yu. Efficient parallel data mining for association rules. In *Proceedings of the fourth international conference on Information and knowledge management*, pages 31–36. ACM, 1995.

[PPK13] S E Polykalas, G N Prezerakos, and A Konidaris. An algorithm based on Google Trends' data for future prediction. Case study: German elections. In *IEEE International Symposium on Signal Processing and Information Technology*, pages 69–73, 2013.

[PRdS16] C.F. Pinho Rebelo de Sá. *Pattern Mining for Label Ranking*. PhD thesis, Universiteit Leiden, 2016.

[RDSK12] Geraldina Ribeiro, Wouter Duivesteijn, Carlos Soares, and Arno Knobbe. Multilayer perceptron for label ranking. *Artificial Neural Networks and Machine Learning–ICANN 2012*, pages 25–32, 2012.

[Rok09] S Rokkan. *Citizens, Elections, Parties: Approaches to the Comparative Study of the Processes of Development*. ECPR Press Classics. ECPR Press, 2009.

[RS88] Kenneth Rogoff and Anne Sibert. Elections and Macroeconomic Policy Cycles. *The Review of Economic Studies*, 55(1):1, 1988.

[SASS15] Pedro Saleiro, S??lvio Amir, M??rio Silva, and Carlos Soares. POPmine: Tracking political opinion on the web. *Proceedings - 15th IEEE International Conference on Computer and Information Technology, CIT 2015, 14th IEEE International Conference on Ubiquitous Computing and Communications, IUCC 2015, 13th IEEE International Conference on Dependable, Autonomic and Se*, pages 1521–1526, 2015.

[Smi11] Aaron Smith. Twitter and social networking in the 2010 midterm elections. *Pew Research*, 2011.

[Spe04] Charles Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101, 1904.

REFERENCES

[SRP+13] Pedro Saleiro, Lus Rei, Arian Pasquali, Carlos Soares, Jorge Teixeira, Fabio Pinto, Mohammad Nozari, Catarina Felix, and Pedro Strecht. POPSTAR at RepLab 2013: Name ambiguity resolution on Twitter. *CEUR Workshop Proceedings*, 1179, 2013.

[TBD02] Ljupco Todorovski, Hendrik Blockeel, and Saso Dzeroski. Ranking with predictive clustering trees. In Tapio Elomaa, Heikki Mannila, and Hannu Toivonen, editors, *Machine Learning: ECML 2002*, pages 444–455, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg.

[TO99] Ah-Hwee Tan and Others. Text mining: The state of the art and the challenges. In *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, volume 8, pages 65–70. sn, 1999.

[TS98] Shiby Thomas and Sunita Sarawagi. Mining generalized association rules and sequential patterns using sql queries. In *KDD*, pages 344–348, 1998.

[TWC16] Kuan-Chieh Tung, En Tzu Wang, and Arbee L P Chen. *Mining Event Sequences from Social Media for Election Prediction*, pages 266–281. Springer International Publishing, Cham, 2016.

[VG11] Shankar Vembu and Thomas Gärtner. *Label Ranking Algorithms: A Survey*, pages 45–64. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[Web06] Geoffrey I Webb. Discovering significant rules. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 434–443. ACM, 2006.

[WL16] M H Wang and C L Lei. Boosting election prediction accuracy by crowd wisdom on social forums. In *2016 13th IEEE Annual Consumer Communications Networking Conference (CCNC)*, pages 348–353, jan 2016.

REFERENCES

Appendix A

Portuguese datasets variables

Table A.1: Variables of the Portuguese elections datasets

	features
$\mathcal{A}_1$	Regions
$\mathcal{A}_2$	Votes Total
$\mathcal{A}_3$	Votes BE
$\mathcal{A}_4$	Votes CDS-PP
$\mathcal{A}_5$	Votes PCP-PEV
$\mathcal{A}_6$	Votes PPD/PSD
$\mathcal{A}_7$	Votes PS
$\mathcal{A}_8$	Residents Total
$\mathcal{A}_9$	% Residents 0-14
$\mathcal{A}_{10}$	% Residents 15-64
$\mathcal{A}_{11}$	% Residents 65+
$\mathcal{A}_{12}$	% Population without Education
$\mathcal{A}_{13}$	% Population with High School
$\mathcal{A}_{14}$	Companies Created
$\mathcal{A}_{15}$	Average Monthly Income
$\mathcal{A}_{16}$	Employees per sector: Total
$\mathcal{A}_{17}$	% Employees per sector: Agriculture
$\mathcal{A}_{18}$	% Employees per sector: Extraction Industries
$\mathcal{A}_{19}$	% Employees per sector: Transformation Industries
$\mathcal{A}_{20}$	% Employees per sector: Construction
$\mathcal{A}_{21}$	% Employees per sector: Gross and Retails Markets
$\mathcal{A}_{22}$	% Employees per sector: Administration and Services
$\mathcal{A}_{23}$	% Employees per economic sector: Banks
$\mathcal{A}_{24}$	% Employees per economic sector: Others
$\mathcal{A}_{25}$	Companies Dissolved

Portuguese datasets variables

$\mathcal{A}_{26}$	% Unemployment Total
$\mathcal{A}_{27}$	% Unemployment <25
	labels
λ_1	BE
λ_2	CDS-PP
λ_3	PCP-PEV
λ_4	PPD/PSD
λ_5	PS

Appendix B

Gamma coefficient implementation in R

```
gamma <- function(realRankingMatrix, predictedRankingMatrix)
{
  comp <- realRankingMatrix == predictedRankingMatrix
  diag(comp) <- NA
  return((sum(comp, na.rm = TRUE)-sum(!comp, na.rm = TRUE))/sum((comp)>=0, na.rm =
      T))
}
```

Appendix C

Classification Accuracy implementation in R

```r
accuracy <- function(realRankingMatrix, predictedRankingMatrix)
{
  comp <- realRankingMatrix == predictedRankingMatrix

  diag(comp) <- NA
  return(sum(!comp, na.rm = TRUE) == 0)
}
```