Pattern mining for label ranking

Pinho Rebelo de Sá, C.F.

Chapter 7

Conclusions

In this thesis, we addressed label ranking problems with popular data mining techniques. In most cases, typical data mining approaches had to be adapted to better explore the complexity of the object of study, the rankings. Ranking are objects with multiple dimensions. Hence, one challenge is to define the border between similar and distinct rankings (Chapter 2). On the other hand, this multi-dimensionality allowed us to explore different facets of the rankings (Chapter 5 and Chapter 6).

We proposed methods that are either direct or reduction techniques. Considering the results obtained, we believe that direct and reduction approaches complement each other by providing different perspectives of the label ranking problem (Chapter 5).

Whenever applicable, we compared our findings with the state-of-the-art label ranking approaches. The good results obtained demonstrate that the proposed approaches are meaningful and competitive. In particular, the adaptation of one popular approach for classification and regression tasks, Random Forests, led to a highly competitive label ranking method (Chapter 4).

Label Ranking Association Rules were proposed as a predictive approach for label ranking [36]. In Chapter 2, we consolidated the work on Label Ranking Association Rules and presented an extensive empirical analysis of its behavior. The performance was analyzed from different perspectives, such as *accuracy*, *number of rules* and *average confidence*. The results show that, for label ranking datasets in general, similarity-based interest measures contribute positively to the accuracy of the model. Results also seem to indicate that the higher the entropy of the rankings on a dataset, the more the accuracy can be affected by the similarity threshold. This can be used

as an indicator for setting the threshold according to the characteristics of the data.

In Chapter 3, we proposed two *supervised* discretization approaches for label ranking. The two methods, based on a well-known supervised discretization approach for classification, are referred to as *Minimum Description Length Partition for Ranking* (MDLP-R) and Entropy-based Discretization for Ranking (EDiRa). Both use different heuristic measures of entropy, based on the *Shannon entropy* [54], to discretize numeric variables.

An analysis of MDLP-R was performed in terms of the similarity threshold parameter. It was clear that, in simple scenarios, MDLP-R deals with noisy ranking data appropriately and that the threshold plays a major role in its behavior. When there are only a few distinct rankings in the data, the method can be less sensitive to the ranking similarities. We also observed that, in more complex situations, MDLP-R tends to overfit the data.

For comparison, a supervised discretization method for classification was also used, recurring to a Ranking As Class transformation [39]. Hence, the original MDLP discretization method for classification was also used in label ranking problems. However, as expected the latter failed to distinguish very similar, but not equal, rankings [39]. This RAC transformation also comes with the problem that, the number of classes can be extremely large, up to a maximum of $k!$, where k is the number of labels in $\mathcal{L}$.

Concerning the second method proposed, EDiRa, the experiments indicate that this is a more stable and efficient method when compared to MDLP-R. An analysis of EDiRa shows that it clearly outperforms MDLP-R and does not have the problem of overfitting, in the presence of noisy ranking data, as its predecessor. The proposed supervised discretization approaches can motivate the creation of new methods that, otherwise, could not deal with continuous data.

The measure of entropy used in EDiRa is more simple and showed better sensibility to ranking than the previous one. We also believe that, despite its heuristic nature, makes sense and may be more generally useful in label ranking. Furthermore, it can be also applied to other fields (e.g. regression) since it is based on a distance measure such as Kendall τ.

In Chapter 4, this measure of entropy was implemented in the splitting process of a decision tree, giving rise to a novel ranking tree approach, Entropy Ranking Trees. We also implemented and analyzed an improved version of Ranking Trees [115].

An analysis of the behavior concluded that both are valid and competitive approaches. In general, Entropy Ranking Trees generated trees with much smaller depth than Ranking Trees. On the other hand, Ranking Trees had better accuracy. Statistical tests showed that none of the methods is significantly different from the state-of-the-art approach, Label Ranking Trees [26].

As a natural extension of this work, and considering the success of Random Forests for classification and regression tasks [13], we proposed Label Ranking Forests in Chapter 4. Two versions were proposed. One approach used Ranking Trees as base model and the other used Entropy Ranking Trees. We observed a clear improvement of the accuracy in comparison to the corresponding base methods. The results confirmed that Label Ranking Forests are highly competitive label ranking methods.

In Chapter 5 we introduced Exceptional Preferences Mining for mining label ranking data. It consists of a supervised local pattern mining task where the target concept is a ranking of a fixed set of labels. The result of this task is a set of subgroups, described as a conjunction of conditions, where the label preferences are exceptional in some sense. Three quality measures were developed to measure different kinds of exceptionality in preferences, *Pairwise*, *Labelwise* and *Norm*. A discussion of the relative merits, drawbacks, and foci of the quality measures was provided, including guidelines regarding when to use which measure.

One of the main benefits of a local pattern mining method such as Exceptional Preferences Mining is that it delivers interpretable results. That means that the resulting subgroups are ideally suited to instigate real-world policies and action. In particular, the experiments on the Algae and Sushi datasets provided a valuable exploration of the data with interpretable results. In terms of visualization of rankings, the Preference Matrix visualization was able to reveal information that was not easy to obtain with the usual representations of rankings.

In Chapter 2, we proposed Pairwise Association Rules (PAR) as a decomposition method for mining label ranking datasets. Pairwise Association Rules successfully found interesting subranking patterns in both the *Algae* and *Sushi* datasets. The results clearly show the potential of this relaxed approach that finds subsets of data for which, some parts of rankings are frequently observed. This approach is more relaxed than Label Ranking Association Rules, in the sense that it does not force to find complete rankings. In future research, Pairwise Association Rules could also be used for predictive tasks.

In Chapter 6, we investigated the usefulness of the type B datasets from the KEBI repository, and proposed two *swap randomization* methods specifically for label ranking datasets. As in [62], we used statistical tests to validate the significance of the findings.

We conclude that, even though KEBI datasets have a semi-synthetic nature, they carry relevant preference information that can be learned by contemporary label rankers. In particular, there were no obvious differences between the type A and type B datasets.

As a side note, one minor contribution was the adoption of the Algae dataset (Chapter 5). This dataset, originally for multi-regression problems (referred as COIL 1999 Competition Data [96]), was approached in this thesis from a label ranking perspective. Here, the set of frequencies of the algae were interpreted as rankings. This led to a different approach of the problem, where we want to understand in which conditions some algae prevail and others not.

We proposed *Preference Rules*, as a generic term of association rules for mining ranking data. Label Ranking Association Rules and Pairwise Association Rules can be regarded as specialization of general association rules that handle ranking data. We strongly believe that such a distinction is important to emphasize the complexity of the rankings, in comparison to other type of targets [57].

As future work, we believe that PAR have potential to be used as predictive models. However, a straightforward implementation might not give satisfactory results since pairwise conflicts can appear (e.g. $A \rightarrow \lambda_1 \succ \lambda_2 \wedge \lambda_2 \succ \lambda_1$). For this, proper aggregation techniques must be used. Also, giving the unusual structure of PAR, with multiple items in the consequent, appropriate interest measures can be developed to handle this type of rules [8].

In our opinion, Exceptional Preferences Mining, can be useful in other fields, other than the ones explored in this work (Chapter 5). For example, in the discovery of profiles with same voting trends. Also, as we broaden the scope of Exceptional Preferences Mining, more quality measures can be developed to better suit the problems at hand.

List of Acronyms

DM Data Mining

AR Association Rules

LR Label Ranking

LRAR Label Ranking Association Rules

PAR Pairwise Association Rules

MDLP Minimum Description Length Principle

MDLP-R Minimum Description Length Principle for Ranking data

EDiRa Entropy-based Discretization for Ranking data

RAC Ranking As Class