Mining Association Rules for Label Ranking

No Author Given

No Institute Given

Abstract. Label ranking (LR) is an increasingly popular topic in machine learning [8, 4, 20]. It studies the problem of learning a mapping from instances to rankings over a finite number of predefined labels. It can be considered as a variant of the conventional classification problem, where only a single label is requested instead of a ranking of all labels [4]. One such example is the predictions of rankings of financial analysts, in terms of the value of their recommendations. The predicted rankings can be used by investors to decide which analysts to follow.

In parallel, Association Rules (AR) Mining have been one of the most successful data mining algorithms. It has been used in market basket analysis and recommendation systems applications. Despite being originally created for descriptive tasks, it has been adapted for predictive ones.

The goal of this project is to adapt AR for LR. To achieve this, several adaptations to the algorithm must be made. Additionally, pre-processing methods which are often used with AR, such as discretization of continuous attributes, must also be adapted. This project builds on our previous work [16] where a preliminary adaptation was made.

1 State of the art

There are two main approaches to the problem of label ranking that we may refer to as decomposition and direct methods. Decomposition methods decompose the problem into several simpler problems. An example of the former is the ranking by pairwise comparisons, where the problem of predicting a ranking of n labels is split into up to n! problems of predicting the order of one pair of labels [10]. Each pairwise comparison problem is a binary classification task, which can be solved using one of the many methods which have been developed and thoroughly investigated for this purpose. However, if the number of labels, n, is very large, the computational cost of this approach can be very high. Additionally, to predict a ranking, it is necessary to combine the individual predictions of all the pairwise models, which may require solving some conflicts (e.g., when the predictions concerning the pairwise ranks of labels A, B and C are A ¿ B, B ¿ C and C ¿ A). Direct methods adapt existing algorithms or develop new ones to treat the rankings as target variables without any transformation. Examples of algorithms that were adapted to deal with rankings as the target objects include decision trees [19, 4], k -Nearest Neighbor [3, 4] and the linear utility transformation [6]. In this approach, one or more parts of the algorithm are adapted (e.g., the splitting criterion in decision trees), while retaining the general characteristics of the learning method (e.g., top-down induction of decision trees). Direct Methods can be divided into two approaches. The first one contains methods (e.g., [4]) that are based on statistical distributions of rankings, such as Mallows [5]. The other group of methods are based on measures of similarity or correlation between rankings (e.g., [19, 2, 15]). The most commonly used correlation measures are the Kendall's tau [12] and the Spearman's Rho [18]. Association Rules mining is a very important and successful task in data mining. It is used to discover interesting relationships between attributes in generally large databases [1]. An AR has the form A ¿ B, meaning that when the set of values A is observed in the data, there is a high probability of observing B. APRIORI [1] is one of the most used and studied algorithms in this field of study. APRIORI identifies the set of all rules that have a support (i.e., the proportion of examples containing the set of values A and B) and confidence (i.e., the proportion of examples that contain B out of those that contain A) higher than the corresponding minimum thresholds that are given as parameters. Although AR were originally developed for descriptive tasks, their success has quickly lead to their adaptation for prediction problems. In contrast to association rule induction, classification does not intend to explore the data to discover interesting behaviors, but rather to decide how new cases should be classified. The motivation for adapting AR for classification is that a classification rule model built from such an unrestrained set of rules can potentially be more accurate than the ones using a greedy search approach [14]. So, AR have been proposed for the first time as complete and competitive classification models by Liu et al in 1998 [14] with the CBA algorithm, which is based on APRIORI. In CBA, the output is a set of rules of the form - A -¿ C - where A is a set of items and C the label/class predicted by the rule. After that, alternative adaptations have been proposed, such as CMAR [13]. One very simple adaptation of the APRIORI for LR (APRIORI-LR [16]) was developed in the MSc Thesis of the candidate, which originated the publications [15, 17]. The main adaptations were in terms of re-defining the support and confidence measures, in order to take into account the nature of label rankings. Based on the categorization given earlier, this is a similarity-based, direct LR method. The results obtained show that this is a promising approach. Furthermore, we observed that there are many opportunities for further improvement. The goal of this project is to work on those opportunities.

2 Objectives

The goal of this project is the investigation of different variants of Association Rules-based methods for Label Ranking. As a starting point, it will be considered the preliminary work developed in [17] and [16]. We plan to:

– improve the method of generation of itemsets that was developed previously based on the APRIORI algorithm [1]. Other possibilities will be considered such as FP-Tree.
– implement alternative similarity functions for the similarity-based support and confidence, extending our previous work.

- adapt methods for pruning rules for LR, including methods that are based in confidence, lift, etc
- adapt AR-based methods to generate predictive models, addressing LR-specific issues such as solving conflicts and generation of predictions for unobserved target values.
- adapt measures for the evaluation of AR, such as lift, conviction, etc.
- adapt the method of dyadic learning [9] for association rules in the context of rankings of financial analysts [2]

We will also investigate preprocessing methods, focusing on operations that are essential for AR mining, including:

- discretization of numerical values, taking inspiration from methods used in classification, such as Fayyad and Irani's entropy-based feature selection.
- feature selection methods that measure the information about the target contained in the attributes.

To assess the approach proposed, we will carry out empirical studies. Namely, we will:

- test the methods on benchmark problems, such as the ones in the KEBI repository http://www.uni-marburg.de/fb12/kebi/research/repository/.

We will also test them on our applications, including metalearning [3], ranking of financial analysts [2] and music preferences in a portuguese music portal, PalcoPrincipal, a company with which the research team has been collaborating.

- compare our methods with selected state-of-the-art approaches, including representatives from the different types described earlier.

3 Detailed description

Label ranking is an increasingly popular topic in the machine learning literature [8, 4, 20]. Label ranking studies the problem of learning a mapping from instances to rankings over a finite number of predefined labels. It can be considered as a variant of the conventional classification problem [4]. In contrast to a classification setting, where the objective is to assign examples to a specific class, in label ranking we are interested in assigning a complete preference order of the labels to every example. A detailed description of the LR Problem can be found in [20].

Association Rule Mining is a commonly used and studied technique to discover interesting relationships between attributes in generally large databases [1]. Although originally developed for descriptive tasks, AR were adapted for the first time as complete and competitive classification models by Liu et al in 1998 [14] with the CBA algorithm.

A preliminary adaptation of AR for LR has been proposed as part of the M.Sc. thesis of the candidate. The main change was in the support and confidence measures, which were adapted to take into account the nature of rankings. In classification, two target values are either the same or not. In LR, we can measure the similarity between two rankings. As it is demonstrated in [17], which was developed by the candidate and a team including one of the supervisors, this approach revealed promising results.

The goal of this project is to propose and evaluate algorithms based in Association Rules for Label Ranking. For that, we will develop a theoretical framework by adapting concepts from Association Rule Mining and Label Ranking. The starting point is the adaptation that was previously proposed [15].

The adaptation of AR for LR implies changes to the basic algorithm, namely:

- generation of itemsets,
- basic measures of support and confidence, extending our previous work, and
- pruning the rule set.

In our previous work, we proposed a simple adaptation of the the APRIORI algorithm and the basic measures of support and confidence [16]. We will extend on this work namely by proposing further improvements and by considering alternative algorithms, such as FP-Tree. AR algorithms are known to generate a very large number of rules, which makes pruning a very important step of the model generation process. However, for effective pruning, the selection criteria must depend on the task. We will adapt existing pruning methods (e.g., based on confidence and lift) for LR.

To develop AR-based methods for LR we will search for inspiration in other adaptations of AR for predictive tasks, in particular, for classification. However, due to the differences in nature between LR and other prediction tasks, further changes must be made. For instance, it is possible for several rules with different rankings to fire for a single example. Therefore, we must design suitable strategies to solve the conflicts, which, are usually called consensus ranking methods [11]. We will start with methods that were previously proposed for this purpose in machine learning [3] and develop new ones, if necessary.

Concerning the evaluation of the AR, a number of measures have been developed, including lift and conviction. Given the differences between LR and other prediction tasks, specific measures were also developed for this task, typically based in rank correlation measures such as Spearman rho and Kendall tau [20]. Adapting AR for LR requires the creation of new measures, possibly combining existing measures from both fields. Two particular aspects of LR should be taking into account. Usually it is more important to predict the items in the top ranks than the ones ranked lower. For instance, when predicting the ranking of financial analysts to choose which ones to follow, it is more important to predict the best ones correctly than the worst ones. Additionally, labels are frequently associated with cost and benefit values, which determine the real value of the ranking. For instance, to follow a given analyst, I may have to buy its recommendations. On the other hand, different analysts make recommendations which yield different gains or losses in the market. The empirical evaluation of ranking methods will only be useful in practice if these issues are taken into account.

Concerning the methodologies for experimental evaluation, the main difference relative to other supervised learning problems is the evaluation measures mentioned above. Thus, it is possible to reuse many of the methodologies used in those problems, such as n-fold cross validation.

It is necessary to prepare the data for label ranking as in any machine learning or data mining task. Failure to prepare the data adequately may compromise the results obtained with the learning algorithms. Some of these methods depend on the target variable (e.g., Fayyad and Irani's entropy-based discretization of numerical variables). However, to the best of our knowledge, there has been no work in this area. Thus, we will also investigate preprocessing methods, focusing on operations that are essential for AR mining, including:

- discretization of numerical values, taking inspiration from methods used in classification, such as Fayyad and Irani's entropy-based feature selection [7].
- feature selection methods that measure the information about the target contained in the attributes.

We note that it is expected that the discretization and preprocessing methods developed may be useful also for other LR methods.

To assess the approach proposed, we will carry out empirical studies. Namely, we will compare the methods developed in this project with baseline methods, such as the default ranking [3] as well as state-of-the-art approaches. Concerning the datasets, we will use:

- benchmark problems, such as the ones in the KEBI repository [5],
- metalearning problems [3],
- ranking of financial analysts [2]
- music preferences in a portuguese music portal, PalcoPrincipal. This is a company with which the research team has been collaborating. Preliminary analysis indicates that they have interesting label ranking problems.

We will identify the most interesting state-of-the-art methods and implement them. Our selection will take into account not only the quality of the results obtained by those methods, but we will ensure that there will be at least one method from each of the different categories that we identified earlier (decomposition and direct methods; distribution and similarity-based methods).

This PhD project will be integrated in the project PTDC/EIA/81178/2006, Rank!: Development of a methodology to predict rankings of items, which has been approved by Fundação para Ciência e Tecnologia (http://www.fct.mct.pt), initiated on the 1st of June of 2007 where the main researcher is the Prof. Dr. Carlos Soares, which is the supervisor of this PhD proposal. Besides, some label ranking applications might be tested and implemented in Palco AdI project Palco3.0 financed by QREN and Fundo Europeu de Desenvolvimento Regional (FEDER) where two supervisors of this PhD proposal participate. The work will be carried out at INESC TEC - Porto.

References

1. Agrawal, R., Imielinski, T., Swami, A.N.: Mining association rules between sets of items in large databases pp. 207–216 (1993)
2. Aiguzhinov, A., Soares, C., Serra, A.P.: A similarity-based adaptation of naive bayes for label ranking: Application to the metalearning problem of algorithm recommendation. In: Discovery Science. pp. 16–26 (2010)
3. Brazdil, P., Soares, C., da Costa, J.P.: Ranking learning algorithms: Using ibl and meta-learning on accuracy and time results. Machine Learning 50(3), 251–277 (2003)
4. Cheng, W., Huhn, J.C., Hüllermeier, E.: Decision tree and instance-based learning for label ranking. In: ICML. p. 21 (2009)
5. Cheng, W., Hüllermeier, E.: Instance-based label ranking using the mallows model pp. 143–157 (2008)
6. Dekel, O., Manning, C.D., Singer, Y.: Log-linear models for label ranking. In: NIPS (2003)
7. Fayyad, Irani: Multi-interval discretization of continuous-valued attributes for classification learning. In: International Conference on Machine Learning. pp. 1022–1027 (1993)
8. Fürnkranz, J., Hüllermeier, E.: Preference learning. KI 19(1), 60– (2005)
9. Hofmann, T., Puzicha, J., Jordan, M.I.: Learning from dyadic data. In: NIPS. pp. 466–472 (1998)
10. Hüllermeier, E., Fürnkranz, J., Cheng, W., Brinker, K.: Label ranking by learning pairwise preferences. Artif. Intell. 172(16-17), 1897–1916 (2008)
11. Kemeny, J., Snell, J.: Mathematical Models in the Social Sciences. MIT Press (1972)
12. Kendall, M., Gibbons, J.: Rank correlation methods (1970)
13. Li, W., Han, J., Pei, J.: Cmar: Accurate and efficient classification based on multiple class-association rules pp. 369–376 (2001)
14. Liu, B., Hsu, W., Ma, Y.: Integrating classification and association rule mining. Knowledge Discovery and Data Mining pp. 80–86 (1998)
15. Sá, C., Soares, C., Jorge, A., Azevedo, P., Costa, J.: Mining Association Rules for Label Ranking, PL-10, ECML-PKDD (2010)
16. Sá, C.: APRIORI Algorithm for Label Ranking. Master's thesis, FCUP (2010)
17. de Sá, C.R., Soares, C., Jorge, A.M., Azevedo, P.J., da Costa, J.P.: Mining association rules for label ranking. In: PAKDD (2). pp. 432–443 (2011)
18. Spearman, C.: The proof and measurement of association between two things. American Journal of Psychology 15, 72–101 (1904)
19. Todorovski, L., Blockeel, H., Dzeroski, S.: Ranking with predictive clustering trees. In: ECML. pp. 444–455 (2002)
20. Vembu, S., Gärtner, T.: Label Ranking Algorithms: A Survey. In: Johannes Fürnkranz, E.H. (ed.) Preference Learning. Springer–Verlag (2010)