Information Sciences 000 (2015) 1–16

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Entropy-based discretization methods for ranking data

Cláudio Rebelo de Sá[a,b,*], Carlos Soares[b,c], Arno Knobbe[a]

[a] LIACS, Universiteit Leiden, Netherlands
[b] INESC TEC, Porto, Portugal
[c] Faculdade de Engenharia, Universidade do Porto, Portugal

ARTICLE INFO

Article history:
Received 1 April 2014
Revised 9 April 2015
Accepted 11 April 2015
Available online xxx

Keywords:
Label ranking
Discretization
Association Rule Mining
Minimum description length

ABSTRACT

Label Ranking (LR) problems are becoming increasingly important in Machine Learning. While there has been a significant amount of work on the development of learning algorithms for LR in recent years, there are not many pre-processing methods for LR. Some methods, like Naive Bayes for LR and APRIORI-LR, cannot handle real-valued data directly. Conventional discretization methods used in classification are not suitable for LR problems, due to the different target variable. In this work, we make an extensive analysis of the existing methods using simple approaches. We also propose a new method called *EDiRa* (Entropy-based Discretization for Ranking) for the discretization of ranking data. We illustrate the advantages of the method using synthetic data and also on several benchmark datasets. The results clearly indicate that the discretization is performing as expected and also improves the results and efficiency of the learning algorithms.

1. Introduction

Research in Label Ranking (LR) has been increasing over the last few years [30,10,7,8,32,33]. LR studies the problem of learning a mapping from instances to rankings over a finite number of predefined labels. An example of an LR problem is the ranking of a set of restaurants according to the preferences of a given person. It can be considered as a variant of the conventional classification problem [6]. However, in contrast to a classification setting, where the objective is to assign examples to a specific class, in LR we are interested in assigning a complete preference order of the labels to every example. An additional difference is that the true (possibly partial) ranking of the labels is available for the training examples.

As in any machine learning task, data preparation is essential for the development of accurate LR models. For instance, some algorithms are unable to deal with numeric variables, such as the basic versions of Naive Bayes and Association Rules [28,1], in which case numeric variables should be discretized beforehand. Discretization, from a general point of view, is the process of partitioning a given interval into a set of discrete sub-intervals. It is normally used to split continuous intervals into two or more sub-intervals which can then be treated as nominal values. In theory, a good discretization should have a good balance between the loss of information and the number of partitions [23]. While there has been a significant amount of work on the development of learning algorithms for LR in recent years, there are not many pre-processing methods specifically for this task.

Discretization methods are typically organized in two groups, depending on whether or not they involve target variable information. These are usually referred to as *supervised* and *unsupervised* discretization, respectively. Previous research found

* Corresponding author at: LIACS, Universiteit Leiden, Netherlands.
E-mail addresses: c.f.de.sa@liacs.leidenuniv.nl (C.R. de Sá), csoares@fe.up.pt (C. Soares), a.j.knobbe@liacs.leidenuniv.nl (A. Knobbe).

http://dx.doi.org/10.1016/j.ins.2015.04.022
0020-0255/

that the supervised methods produce more useful discretizations than unsupervised methods [12]. The difference in nature between the target variable in classification and in LR problems implies that supervised discretization methods developed for the former are not suitable for the latter. In fact, in classification, two target values (i.e., classes) are either equal or different, while in LR, the difference between two rankings is closer to a continuous function, similar to the error in a regression setting. In this work, we make an extensive empirical analysis of the existing methods. We also propose a new method based on Minimum Description Length Principle (MDLP) [14] for the discretization of ranking data. The new method of supervised discretization for ranking data, which we refer to as EDiRa (Entropy-based Discretization for Ranking), follows the line of work in [11]. Based on MDLP for classification, it adapts the concept of entropy to LR based on the distance between rankings.

We also make an extensive study of the *Minimum Description Length Principle for Ranking data (MDLP-R)* method proposed in [11], which is also based on MDLP [14]. This analysis includes varying its parameter to assess how it affects the performance of the learner.

Finally we present a comparison between the newly proposed approach EDiRa and MDLP-R, along with the original MDLP (i.e. for classification). The results observed show that EDiRa behaves better in many scenarios and is also more robust.

The paper is organized as follows: Section 2 introduces the LR problem and the learning algorithms used in this paper. Section 3 introduces discretization and Section 4 describes the method proposed here. Section 5 presents the experimental setup and discusses the results. Finally, Section 6 concludes this paper.

2. Label Ranking

The LR task is similar to classification. In classification, given an instance x from the instance space $\mathbb{X}$, the goal is to predict the label (or class) λ to which x belongs, from a pre-defined set $\mathcal{L} = \{\lambda_1, \ldots, \lambda_k\}$. In LR, the goal is to predict the ranking of the labels in $\mathcal{L}$ that are associated with x [19]. A ranking can be represented as a total order over $\mathcal{L}$ defined on the permutation space Ω. In other words, a total order can be seen as a permutation π of the set $\{1, \ldots, k\}$, such that $\pi(a)$ is the position of λ_a in π.

As in classification, we do not assume the existence of a deterministic $\mathbb{X} \to \Omega$ mapping. Instead, every instance is associated with a *probability distribution* over Ω [6]. This means that, for each $x \in \mathbb{X}$, there exists a probability distribution $\mathcal{P}(\cdot\,|x)$ such that, for every $\pi \in \Omega$, $\mathcal{P}(\pi|x)$ is the probability that π is the ranking associated with x. The goal in LR is to learn the mapping $\mathbb{X} \to \Omega$. The training data contains a set of instances $D = \{\langle x_i, \pi_i\rangle\}, i = 1, \ldots, n$, where x_i is a vector containing the values $x_i^j, j = 1, \ldots, m$ of m independent variables describing instance i and π_i is the corresponding target ranking.

Given an instance x_i with label ranking π_i, and the ranking $\hat{\pi}_i$ predicted by an LR model, we evaluate the accuracy of the prediction with a loss function on Ω. One such function is the number of discordant label pairs,

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

If normalized to the interval $[-1, 1]$, this function is equivalent to Kendall's τ coefficient [21], which is a correlation measure where $\mathcal{D}(\pi, \pi) = 1$ and $\mathcal{D}(\pi, \pi^{-1}) = -1$ (π^{-1} denotes the inverse order of π).

The accuracy of a model can be estimated by averaging this function over a set of examples. This measure has been used for evaluation in recent LR studies [6,11] and, thus, we will use it here as well. However, other correlation measures, like Spearman's rank correlation coefficient [31], can also be used.

Given the similarities between LR and classification, one could consider workarounds that treat the label ranking problem essentially as a classification problem. One such workaround is *Ranking As Class (RAC)* [11], which replaces the rankings with classes:

$$\forall \pi_i \in \Omega, \quad \pi_i \to \lambda_i.$$

This approach allows the use of all pre-processing and prediction methods for classification in LR problems.

2.1. Association Rules for Label Ranking

Label Ranking Association Rules (LRAR) [10] are a straightforward adaptation of Class Association Rules (CAR):

$$A \to \pi$$

where $A \subseteq desc(\mathbb{X})$ and $\pi \in \Omega$. Where $desc(\mathbb{X})$ is the set of descriptors of instances in $\mathbb{X}$, typically pairs $\langle attribute, value\rangle$. Similar to how predictions are made with CARs in CBA (Classification Based on Associations) [26], when an example matches the antecedent of the rule, $A \to \pi$, the predicted ranking is π.

If the RAC approach is used, the number of classes can be extremely large, up to a maximum of $k!$, where k is the number of labels in $\mathcal{L}$. This means that the amount of data required to learn a reasonable mapping $\mathbb{X} \to \Omega$ can be very large.

Alternatively, mining of LRAR uses similarity-based support and confidence measures [10].

2.1.1. Similarity-based support and confidence

Given a measure of similarity $s(\pi_a, \pi_b)$, the *support* of the rule $A \to \pi$ is defined as follows:

$$sup_{lr}(A \to \pi) = \frac{\sum_{i:A \subseteq desc(x_i)} s(\pi_i, \pi)}{n} \tag{1}$$