Pattern mining for label ranking

Pinho Rebelo de Sá, C.F.

English Summary

Preferences have always been present in many tasks in our daily lives. Buying the right car, choosing a suitable house or even deciding on the food to eat, are trivial examples of decisions that reveal information, explicitly or implicitly, about our preferences. The recent trend of collecting increasing amounts of data is also true for preference data.

Extracting and modeling preferences can provide us with invaluable information about the choices of groups or individuals. In areas like e-commerce, which typically deal with decisions from thousands of users, the acquisition of preferences can be a difficult task. For these reasons, artificial intelligence (in particular, machine learning) methods have been increasingly important to the discovery and automatic learning of models about preferences.

The subfield of machine learning which focuses on the study and modeling of preferences is *Preference Learning* (PL). We focus on one subtask of PL, Label Ranking (LR). In simple terms, a LR dataset consists of a set of observations described by attributes (independent variables) and a ranking of a (finite) set of labels (target or dependent variable). In LR, we are interested in predicting the ranking of the labels for a new observation based on the values of the independent variables.

In this Ph.D. project, several approaches were analyzed and proposed to deal with the LR problem. We investigated Label Ranking Association Rules (LRAR), which are the equivalent of Class Association Rules for the LR task. A LRAR is an association rule, where the items are based on the values of the independent values and the right-hand side is a ranking of the labels. Furthermore, we proposed Pairwise Association Rules (PAR), which are defined as association rules with a set of pairwise preferences in the consequent. Like LRAR, PAR can be used both as descriptive and predictive models. However, our analysis of PAR has focused on its descriptive properties, while LRAR have been essentially studied as predictive models.

Preprocessing methods are well known to be an essential part of machine learning processes. This is true for LR as for any other machine learning task. For example, LRARs, like association rules, cannot handle numeric data directly, which needs to be discretized beforehand. However, no LR-specific methods existed. Hence, we proposed two discretization approaches that are specific for LR problems. Both approaches are based on new measures of *ranking entropy*.

Most of this project has focused on pattern mining methods. However, given the popularity of decision tree methods and how these can clearly express information about the problem, we proposed Entropy Ranking Trees. Although previous approaches existed to adapting decision tree (DT) algorithms for LR, having proposed a measure of ranking entropy, we found it natural to investigate its integration on DT algorithms. Another very popular modeling approach is ensemble learning. In particular, the Random Forests (RF) algorithm has been very successful but was not adapted for LR. RF are an ensemble learning method that combines different trees obtained using different randomization techniques. Hence, we proposed an ensemble of decision trees for Label Ranking, based on Random Forests, which we refer to as Label Ranking Forests (LRF).

We continued our journey on the field of preference learning by combining it with local pattern mining. The task is named Exceptional Preferences Mining (EPM) and can be seen as a local pattern mining task that finds subsets of observations where the preference relations between subsets of the labels significantly deviate from the norm. In other words, it is a variant of Subgroup Discovery, with rankings as the target. We employed three quality measures that highlight subgroups featuring exceptional preferences, where the focus of what constitutes 'exceptional' varies with the measure. The results also illustrate how the visualization of the preferences in a Preference Matrix can aid in interpreting exceptional preference subgroups.

Finally, we proposed an approach to test the relation between the rankings and independent variables in LR datasets. As in other supervised learning tasks, target swap randomization methods have been used to test it. So, we proposed two target swap randomization approaches for LR and apply them on LR datasets.

Experimental results show the potential of the approaches mentioned.