Chapter 23
Ranking Potential Customers Based on Group-Ensemble

Zhi-Zhuo Zhang
South China University of Technology, China

Qiong Chen
South China University of Technology, China

Shang-Fu Ke
South China University of Technology, China

Yi-Jun Wu
South China University of Technology, China

Fei Qi
South China University of Technology, China

Ying-Peng Zhang
South China University of Technology, China

ABSTRACT

Ranking potential customers has become an effective tool for company decision makers to design marketing strategies. The task of PAKDD competition 2007 is a cross-selling problem between credit card and home loan, which can also be treated as a ranking potential customers problem. This article proposes a 3-level ranking model, namely Group-Ensemble, to handle such kinds of problems. In our model, Bagging, RankBoost and Expending Regression Tree are applied to solve crucial data mining problems like data imbalance, missing value and time-variant distribution. The article verifies the model with data provided by PAKDD Competition 2007 and shows that Group-Ensemble can make selling strategy much more efficient.

INTRODUCTION

Data mining plays an increasingly important role in business application and practice. To maximize commercial profits, how to discover potential customers is one of the hottest topics in both data mining and e-business. Although huge amounts of commercial data provide good opportunities to approach the task more thoroughly and precisely, some difficult problems pop up. Among them, imbalance distribution of data, and existence of missing value and dynamic sample distribution are well-known ones, which also occur in the PAKDD Competition 2007. The detail about competition task can be found at the official Web site (LeVis Group, 2007).

The modeling dataset consists of 40,700 customers, only 700 of whom bought home loan as well as a credit card, that is, only 700 of them have a target flag of 1 with others having 0. Besides, among the 40 modeling variables, there exist many missing values. Nearly 90% of the sample more or less suffers this problem. The reason of overlap being small remained unknown, which excluded the possibility of additional assumptions. The provided dataset just gave us the information about whether customers have opened a home loan with the company within 12 months after opening the credit card. It would just so happen that the distribution of potential customers is different from distribution of customers who open a home loan account in the first year. What's more, it is better to treat this problem as a ranking problem, but not a classification problem described in Lecun, Chopra, Hadsell, Huang, and Ranzato (2006), because it would be more convenient for the company decision makers to put the limited resources to the most potential ones.

In this article, we proposed a 3-level learning model named Group-Ensemble to handle the potential ranking associating with data imbalance, missing value and time variant distribution. Different from other learning models, this model is designed for ranking which applies RankBoost as its subalgorithm. Moreover, we slightly modify the traditional bagging by reserving all minority class in each bag, which greatly improve the learners' performance in a serious imbalance case.

Group-Ensemble

After analyzing the problem, we think the task has the following difficulties which have to be tackled.

- **Distribution is time variant:** the target flag in the modeling dataset is based on the record within 1 year; however, the task is to predict the propensity of a customer not limited to 1 year.
- **Serious missing value problem:** the modeling dataset comes from the real world, so that nearly 90% of variables in the dataset encounter a serious missing value problem.
- **Serious data imbalance problem:** in the modeling dataset, the ratio of the positive class is only about 1.71%, which means the negative class dominates the whole dataset and causes poor performance in many traditional classifiers.

To handle the difficulties above, we introduce the 3-level Regression model. In the bottom level, ERTree (Expending Regression Tree) is applied to expand the probability distribution from 1 year to overall. Then in middle level, a metalearning method, RankBoost, presented in Freund, Iyer, Robert, Schapire, and Singer (1998), is used for optimizing the AUC value of the model to achieve the best ranking result (Ataman, Streets, & Zhang, 2006; Corinna & Mehryar, 2003). It is also helpful in the imbalance case. Finally, in the top level, a modified bagging method is used in dealing with the imbalance problem, which reduces the time complexity of the model.