Overview of PAKDD Competition 2007

Junping Zhang, Fudan University, China
Guo-Zheng Li, Shanghai University, China

ABSTRACT

The PAKDD Competition 2007 involved the problem of predicting customers' propensity to take up a home loan when a collection of data from credit card users are provided. It is rather difficult to address the problem because 1) the data set is extremely imbalanced; 2) the features are mixture types; and 3) there are many missing values. This article gives an overview on the competition, mainly consisting of three parts: 1) The background of the database and some statistical results of participants are introduced; 2) An analysis from the viewpoint of data preparation, resampling/reweighting and ensemble learning employed by different participants is given; and 3) Finally, some business insights are highlighted.

Keywords: ensemble learning; feature selection; imbalanced data; PAKDD competition; resampling; reweighting

INTRODUCTION

The PAKDD Competition 2007 is about finding better solutions for a cross-selling business problem donated by a consumer finance company. In one of their markets, the company currently has a credit card customer base and a housing loan (mortgage) customer base with few customers overlapping between the two. The company would like to make use of this opportunity to cross-sell home loans to its credit card customers.

In the problem setting of the competition, a modeling database of 40,700 customers with 40 modeling features plus a binary target feature is provided to the participants. This is a sample of customers who opened a new credit card with the company within a specific 2-year period, while they did not have an existing home loan with the company. In the database, 700 cases that the customer opened a home loan with the company within 12 months after opening the credit card are assumed to be positive examples, and the other 40,000 cases are assumed to be negative. Participants were tasked to create cross selling response scores to predict the propensity of 8,000 customers to take up a mortgage. Data are made up of large imbalanced distribution and small absolute number of positive cases, as well as a mixture of feature types. Furthermore, there are a large number of missing data, invalid data and special values in the database. These problems result in a formidable challenge to
participants. A more detailed introduction of the data set can be found in the Web site (PAKDD Competition, 2007).

Moreover, the AUC (area under the curve) value, which is the area under the ROC curve (Receiver Operating Characteristic)\(^1\), was used as the evaluation criterion for the competition. The AUC calculated from a team’s submission for the prediction set was regarded as a major factor for determining the winners; the higher, the better.

This overview is organized as follows. In Section 2, participation and competition results are reported. A survey on data preparation and resampling (or reweighting) techniques, as well as ensemble approaches, are given in Section 3. Business insights into the problem provided by the participants are summarized in Section 4. Finally, Section 5 concludes.

**PARTICIPATION AND RESULTS**

In this competition, 47 entries from 12 countries submitted their write-ups and predictive results for the 8,000 observations in the unlabelled prediction set. There are 7,650 negatives and 350 positives in the prediction set. A set of comparisons on geographic distributions, participant types and modeling techniques between participants in 2006 and participants in 2007 are tabulated in Table 1 to Table 4.

Compared with the PAKDD 2006 Data Mining Competition (Noriel & Tan, 2007), the number of entries slightly increased. Secondly, the entries from industry in 2007 is nearly twice as those in 2006, indicating that the competition has received more attention from industry this year. The percentage of entries from industry are also high in the Top 10 and the Top 20. It shows that entries from industry benefit from their professional experiences and knowledge on how to address data mining problems.

For determining who are the champion and first runner up, the AUC values using trapezoidal rule are computed for all the submitted predictive results. The evaluation results are summarized, as in Figure 1. From the figure, it can be found that the highest AUC is 70.01\%, and the lowest AUC is 47.73\%. The top four entries in terms of AUC values, together with modeling techniques, are reported in Table 5.

**DISCUSSION**

The participants make use of different ways for processing the data set provided by the competition. These steps can be mainly categorized into three levels: data preparation, resampling/reweighting and ensemble learning techniques. Here, we would like to give an in-depth analysis on the approaches used for each of these three levels.

**Data Preparation**

In this competition, mixture types of features, such as ordinal, binary, nominal and interval features, are concurrently present in the cross-sell data set. Specifically, the database is made up of 24 numerical and 16 categorical features. Due to some reasons, missing values also exist in some features. To address the two issues, missing values should be either removed or imputed before performing modeling. Meanwhile, some important features as well as correlation among features should be extracted from the training data.

**Feature Selection and Extraction**

Feature selection and extraction procedures were widely used for capturing important

---

\(^1\) The area under the ROC curve (Receiver Operating Characteristic) is a non-parametric measure of a test's ability to discriminate between groups. It is calculated as the area under the curve of the true positive rate (sensitivity) plotted against the false positive rate (1-specificity) at various threshold settings.
Related Content

Classification Of 3G Mobile Phone Customers
[www.igi-global.com/article/classification-mobile-phone-customers/1782?camid=4v1a](www.igi-global.com/article/classification-mobile-phone-customers/1782?camid=4v1a)

Data Mining Techniques for Communities’ Detection in Dynamic Social Networks
[www.igi-global.com/chapter/data-mining-techniques-communities-detection/73467?camid=4v1a](www.igi-global.com/chapter/data-mining-techniques-communities-detection/73467?camid=4v1a)
Clinical Data Mining in the Age of Evidence-Based Practice: Recent Exemplars and Future Challenges
www.igi-global.com/chapter/clinical-data-mining-age-evidence/44295?camid=4v1a

Weighted Fuzzy-Possibilistic C-Means Over Large Data Sets
Renxia Wan, Yuelin Gao and Caixia Li (2012). International Journal of Data Warehousing and Mining (pp. 82-107).
www.igi-global.com/article/weighted-fuzzy-possibilistic-means-over/74756?camid=4v1a