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 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
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.

Table 1. A comparison of geographic distributions by regions

<table>
<thead>
<tr>
<th>Affiliation Region</th>
<th>2006 All</th>
<th>2007 All</th>
<th>Top 20</th>
<th>Top 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>North America</td>
<td>18</td>
<td>19</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>East Asia</td>
<td>18</td>
<td>18</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>South Asia</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>West Asia</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>South America</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Europe</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>ANZ</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>45</td>
<td>47</td>
<td>20</td>
<td>10</td>
</tr>
</tbody>
</table>

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/rewighting 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 crosssell 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
Privacy Implications of Organizational Data Mining
www.igi-global.com/chapter/privacy-implications-organizational-data-mining/27908?camid=4v1a

A Novel Multi-Secret Sharing Approach for Secure Data Warehousing and On-Line Analysis Processing in the Cloud
www.igi-global.com/article/a-novel-multi-secret-sharing-approach-for-secure-data-warehousing-and-on-line-analysis-processing-in-the-cloud/125649?camid=4v1a