Chapter 16

A Solution to the Cross-Selling Problem of PAKDD-2007: Ensemble Model of TreeNet and Logistic Regression

Mingjun Wei
Zhejiang University and Sherpa Consulting, China

Lei Chai
Sherpa Consulting, China

Renyong Wei
China Mobile Group Zhejiang Co., China

Wang Huo
China Mobile Group Zhejiang Co., China

ABSTRACT

Our team has won the Grand Champion (Tie) of PAKDD-2007 data mining competition. The data mining task is to score credit card customers of a consumer finance company according to the likelihood that customers take up the home loans offered by the company. This report presents our solution for this business problem. TreeNet and logistic regression are the data mining algorithms used in this project. The final score is based on the cross-algorithm ensemble of two within-algorithm ensembles of TreeNet and logistic regression. Finally, some discussions from our solution are presented.

INTRODUCTION

The PAKDD-2007 competition requires development of a cross-selling model for a financial services company offering both credit cards and home loans. The main challenge we face is high class imbalance, and the low incidence rate of the target class in the overall population. In the modeling dataset, the number of target class is extremely small, only 700 records out of 40,700.
The imbalance ratio is as high as 57.143. To deal with such class imbalance, certain data partition strategy and ensemble method are employed to effectively improve the performance of predictive models. TreeNet and logistic regression are the data mining algorithms used in this project. The final score is based on the cross-algorithm ensemble of two within-algorithm ensembles of TreeNet and logistic regression.

**SOFTWARE**

To achieve the analysis objectives, we used the following tools:

- Weka
- CART
- TreeNet

The data mining package Weka (Witten & Frank, 2005) is used to carry out logistic regression modeling task. CART® is the decision tree tool developed by Breiman, Friedman, Olshen and Stone (1984) and exclusively licensed to Salford systems. CART® can automatically sift large, complex databases, searching for significant patterns and relationships. TreeNet® is the tool developed by Friedman (1999) and exclusively licensed to Salford systems. TreeNet® uses Stochastic Gradient Boosting to develop hundreds of small trees, each of which contributes just a tiny adjustment to the overall mode. TreeNet is a breakthrough technology that can offer surprisingly high accuracy while remaining remarkably robust in the presence of common data problems such as missing values and misrecorded training data.

**DATA PREPROCESSING**

There is a need to preprocess the original modeling data into data used for modeling.

### Data Preparation

The supplied dataset went through a standard series of data preparation steps including:

- Study the univariate and bivariate analysis and frequency tables;
- Handle missing values and potential outliers;
- Create new variables from the original ones;
- Compress categorical variables, and so forth.

There were 40 candidate predictors, plus a target variable in the raw forms. Only a couple of predictors had a severe missing problem. B_DEF_UNPD_L12M was excluded from our analysis because it was equal to 0 for all the records. MVIs (Missing Value Indicators) were created for all the predictors with missing values. Using CART, we also made imputations for the missing values. In addition, there were inconsistent values in the categorical variable ANNUAL INCOME RANGE. In the data there was a value “\$0K-<30K”. We deleted the preceding “0” in “\$0K-<30K,” making it consistent with “\$0K-<30K” in the data dictionary.

### Data Partition via 5-Fold Cross-Validation Method

We adopted the 5-fold cross-validation (CV) method that evenly and randomly partitioned the modeling dataset (40,700 records) into 5 nonoverlapping subsets (CV1, CV2, CV3, CV4 and CV5), each of which was approximately as big as the prediction dataset (8,000 records). CV parameter is usually set to 10, but in this case, 10-fold CV would make each fold contain too few positives. Hence, we reduced the CV parameter to 5. Each of 5 subsets was used as a testing dataset, while the rest of 4 subsets were merged and used as a learning dataset. Therefore, every observation...
3 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the product's webpage:
www.igi-global.com/chapter/solution-cross-selling-problem-pakdd/40410?camid=4v1

This title is available in InfoSci-Books, Business-Technology-Solution, InfoSci-Database Technologies, Library Science, Information Studies, and Education, InfoSci-Library Information Science and Technology. Recommend this product to your librarian:
www.igi-global.com/e-resources/library-recommendation/?id=1

Related Content

PaKDD-2007: A Near-Linear Model for the Cross-Selling Problem
www.igi-global.com/article/pakdd-2007-near-linear-model/1806?camid=4v1a

Identifying and Analyzing Popular Phrases Multi-Dimensionally in Social Media Data
Zhongying Zhao, Chao Li, Yong Zhang, Joshua Zhexue Huang, Jun Luo, Shengzhong Feng and Jianping Fan (2015). *International Journal of Data Warehousing and Mining* (pp. 98-112).
www.igi-global.com/article/identifying-and-analyzing-popular-phrases-multi-dimensionally-in-social-media-data/129526?camid=4v1a

Handling Local Patterns in Collaborative Structuring
Ingo Mierswa, Katharina Morik and Michael Wurst (2008). *Successes and New Directions in Data Mining* (pp. 167-186).
www.igi-global.com/chapter/handling-local-patterns-collaborative-structuring/29959?camid=4v1a

Anonymous Spatial Query on Non-Uniform Data
www.igi-global.com/article/anonymous-spatial-query-on-non-uniform-data/105119?camid=4v1a