Chapter 17

Bagging Probit Models for Unbalanced Classification

Hualin Wang
AllianceData, USA

Xiaogang Su
University of Central Florida, USA

ABSTRACT

This chapter presents an award-winning algorithm for the data mining competition of PAKDD 2007, in which the goal is to help a financial company to predict the likelihood of taking up a home loan for their credit card based customers. The involved data are very limited and characterized by very low buying rate. To tackle such an unbalanced classification problem, the authors apply a bagging algorithm based on probit model ensembles. One integral element of the algorithm is a special way of conducting the resampling in forming bootstrap samples. A brief justification is provided. This method offers a feasible and robust way to solve this difficult yet very common business problem.

INTRODUCTION

The 11th Pacific-Asia Knowledge Discovery and Data Mining Conference (PAKDD 2007) hosted a data mining competition, co-organized by the Singapore Institute of Statistics. The data set is from a consumer finance company with the aim of finding solutions for a cross-selling business problem. The company currently has two databases, one for credit card holders and the other for home loan (mortgage) customers and they would like to make use of this opportunity to cross-sell home loans to its credit card holders. Thus, it is of their keen interest to have an effective scoring model for predicting potential cross-sell take-ups.

The training dataset contains information on 40,700 customers with 40 input variables, most of which are related to the point of application for the company’s credit card, plus a binary target variable indicating the home loan take-up status. This is a sample of customers who opened a new credit card with the company within a specific 2-year period and did not have an existing home loan with the company. The binary target variable has a value of 1 if the customer then opened a home loan with the company within 12 months after opening the credit...
Bagging Probit Models for Unbalanced Classification

card (700 customers), and will have a value of 0 if otherwise (40,000 customers). Another test dataset containing 8,000 sampled cases is also available with same input variables but withholding the target variable. The data mining task is to produce a score for each customer in the test dataset, indicating his/her propensity to take up a home loan with the company within 12 months after opening the credit card. More detailed information on the competition can be found at http://lamda.nju.edu.cn/conf/pakdd07/dmc07/.

Clearly, it is a classification problem. However, the main challenge of the analysis stems from the very unbalanced distribution of the target variable. Namely, the proportion of 1’s is only 700/40,000 = 1.72%. This is a very common problem seen in many application areas such as fraud detection, rare disease studies, marketing strategic modeling, network intrusion analysis, and others. The problem, generally termed as unbalanced classification in data mining practices, is characterized by the fact that one class of the response variable is very much underrepresented in the data. When working directly with severely unbalanced data, most classifiers will encounter numerical problems and yield poor performance.

BAGGING WITH WEIGHTED RESAMPLING

The common approach to unbalanced classification is to modify the weights, borrowing the idea from retrospective designs (see, e.g., Agresti, 1990). This amounts to either decreasing the weight for the majority class by under-sampling or increasing the weight for the minority class by over-sampling. However, how to adjust the weights is quite an art. In the following, we shall present our procedure with justification and compare it with some alternative approaches.

To proceed, we first introduce some notations to set up the problem. Let

\[ L_0 = \{(y_i, \underline{x}_i) : i = 1, \ldots, n_0\} \]

denote the training sample, where \( y_i \) is the \( i \)-th binary 0-1 outcome with Class 1 severely under-represented and \( \underline{x}_i \) is the associated input vector. Let \( L_1 = \{\underline{x}_p : p = 1, \ldots, n_1\} \) denote the test sample that contains the input information only.

Let \( D \) denote the distribution underlying the data. What is under modeling is the conditional probability that \( y \) is equal to 1 conditioning on \( \underline{x} \), i.e.,

\[ \pi = \Pr\{y = 1 \mid \underline{x}\}. \tag{1} \]

There are various modeling tools or classifiers available for modeling this probability, e.g., logistic regression, decision trees, neural networks, support vector machine, to name a few. One is referred to Hastie, Tibshirani, and Friedman (2001) for a full account of different modeling or learning processes.

Our procedure is outlined in Algorithm 1. First, we generate \( B \) bootstrap samples, denoted by \( L^{(b)} \) for \( b = 1, \ldots, B \), from the training sample \( L_0 \) by oversampling the minority class. More specifically, let \( \lambda > 1 \) denote the ratio of sampling probabilities for the two classes. In other words, we randomly select an observation from Class 1 with a higher sampling probability \( \lambda / (1 + \lambda) \), while selecting an observation from Class 0 with a lower probability \( 1 / (1 + \lambda) \). A bootstrap sample \( L^{(b)} \) has the same sample size as \( L_0 \) does. For more background information about bootstrap, see Efron and Tibshirani (1998). With the oversampling strategy, the proportions of 1’s and 0’s will become more balanced in the resultant bootstrap sample \( L^{(b)} \). To prepare for the inference below, we introduce a binary indicator variable \( s \) to indicate whether or not an observation is selected by this resampling scheme.
Related Content

DMA: Matrix Based Dynamic Itemset Mining Algorithm  
[www.igi-global.com/article/dma/105120?camid=4v1a](www.igi-global.com/article/dma/105120?camid=4v1a)

A Hyper-Heuristic for Descriptive Rule Induction  
[www.igi-global.com/article/hyper-heuristic-descriptive-rule-induction/1778?camid=4v1a](www.igi-global.com/article/hyper-heuristic-descriptive-rule-induction/1778?camid=4v1a)

Approximate Matching Between XML Documents and Schemas with Applications in XML Classification and Clustering  
[www.igi-global.com/chapter/approximate-matching-between-xml-documents/60906?camid=4v1a](www.igi-global.com/chapter/approximate-matching-between-xml-documents/60906?camid=4v1a)

A Study of Friendship Networks and Blogosphere  
[www.igi-global.com/chapter/study-friendship-networks-blogosphere/21749?camid=4v1a](www.igi-global.com/chapter/study-friendship-networks-blogosphere/21749?camid=4v1a)