Chapter 21
Selecting Salient Features and Samples Simultaneously to Enhance Cross-Selling Model Performance

Dehong Qiu
Huazhong University of Science and Technology, China

Ye Wang
Huazhong University of Science and Technology, China

Qifeng Zhang
Huazhong University of Science and Technology, China

ABSTRACT
The task of the 2007 PAKDD competition was to help a finance company to build a cross-selling model to score the propensity of a credit card customer to take up a home loan. The present work tries to increase the prediction accuracy and enhance the model comprehensibility through efficiently selecting features and samples simultaneously. A new framework that coordinates feature selection and sample selection together is built. The criteria of optimal feature selection and the method of sample selection are designed. Experiments show that the new algorithm not only raises the value of the area under ROC curves, but also reveals more valuable business insights.

INTRODUCTION
The rapid growth in information science and technology has lead to generation of huge amount of valuable data in many areas. In finance for example, over the past five years, many banks have experienced exceptional growth in service and have built up bank’s Group Data Warehouse. In order to realize faster, more effective decisions and provide more excellent customer services, new technologies to handle or extract fully the latent knowledge within the data are urgently required. The finance company that donated the data for 2007 PAKDD competition would like to build a cross-selling model to predict the potential take-ups of cross-selling home loans to its credit card customers (Qiu, Wang & Bi, 2008).
One critical issue in the building of cross-selling model is the processing of a huge amount of data having a high dimensionality. In the modeling dataset of 2007 PAKDD competition each sample is with 40 modeling feature variables and a categorical variable. Usually excessive features contain irrelevant or redundant features, which reduce the performance of data mining and increase computing costs. Feature selection, i.e., selecting an optimal subset of the features available for describing the data is an effective way to improve the performance of data mining (Hall, 2000; Peng, 2005). It has been proven in both theory and practice that there are many advantages of feature selection, such as the dimension reduction to reduce the computational cost, the improvement of predicative accuracy, and more interpretable business insights from the scoring cross-selling model.

Another problem in data mining is that the continued expansion of huge dataset contributes to the difficulty of obtaining a training sample set of good quality in appropriate size. The instances in huge dataset may come from a large variety of sources, be collected carefully or carelessly, be updated normally or not, satisfy the same distribution or not. In the modeling dataset of 2007 PAKDD competition, 40700 samples are provided, which come from a customer base of credit card customers as well as a customer base of home loan customers. The overlap between these two customer bases is very small. Obviously the training dataset is extremely class imbalanced, coming from different sources and may not satisfy the same distribution. In sample-based data mining, how well a predicting model eventually turns out depends heavily on the quality of training samples it receives. To use all training samples uniformly seems suboptimal. It is desirable to pick out training samples that are informative and representative for the final decision function.

Although there exists previous work addressing feature selection and sample selection respectively (Zadrozny, 2004; Liu, 2005; Huang, 2006; Zhang, 2008), there is relatively fewer work on combining them together into processing. In this paper, we first build up the criteria of feature selection and negative sample selection respectively, and then a uniform framework is build to select optimal features and informative samples efficiently and simultaneously. The remainder of this paper is organized as follows. Section 2 describes the problem of feature selection and sample selection respectively. In section 3, we set up the criteria for feature selection. Negative sample selection method is explained in section 4. In section 5, we propose a new framework that realizes feature selection and sample selection simultaneously and efficiently. Section 6 contains an empirical study of the new algorithm on the 2007 PAKDD competition dataset. Section 7 concludes this work.

PROBLEM DESCRIPTION

Feature Selection

Consider a given training set, in which an instance is typically described as \( Z = [(x_1, x_2, \ldots, x_N), c] \), where \((x_1, x_2, \ldots, x_N)\) is one possible value of the vector \((X_1, X_2, \ldots, X_N)\) with \(N\) features and \(c\) is one possible class of the categorical label \(C\). Each sample in the dataset of 2007 PAKDD competition has 40 features. Feature selection is a process that selects a subset from the original features and ignoring others not in this set before applying data mining. The optimality of a feature subset is measured by an evaluation criterion. It is reasonable to believe that for a given problem under the same evaluation criterion, there may be several optimal feature subsets, as shown by figure 1. The entire feature set can be conceptually divided into several optimal feature subsets, an irrelevant feature set and a redundant feature set.

As one of the most important techniques in data preprocessing for data mining, feature selection has received lots of attentions as data mining is...