Classification and Dimension Reduction in Bank Credit Scoring System

Bohan Liu, Bo Yuan, and Wenhuang Liu

Graduate School at Shenzhen, Tsinghua University,
Shenzhen 518055, P.R. China
bohn22@126.com, {yuanb,liuwh}@sz.tsinghua.edu.cn

Abstract. Customer credit is an important concept in the banking industry, which reflects a customer's non-monetary value. Using credit scoring methods, customers can be assigned to different credit levels. Many classification tools, such as Support Vector Machines (SVMs), Decision Trees, Genetic Algorithms can deal with high-dimensional data. However, from the point of view of a customer manager, the classification results from the above tools are often too complex and difficult to comprehend. As a result, it is necessary to perform dimension reduction on the original customer data. In this paper, a SVM model is employed as the classifier and a "Clustering + LDA" method is proposed to perform dimension reduction. Comparison with some widely used techniques is also made, which shows that our method works reasonably well.

Keywords: Dimension Reduction, LDA, SVM, Clustering.

1 Introduction

Customer credit is an important concept in the banking industry, which reflects a customer's non-monetary value. The better a customer's credit, the higher his/her value that commercial banks perceive. Credit scoring refers to the process of customer credit assessment using statistical and related techniques. Generally speaking, banks usually assign customers into good and bad categories based on their credit values. As a result, the problem of credit assessment becomes a typical classification problem in pattern recognition and machine learning.

As far as classification is concerned, some representative features need to be extracted from the customer data, which are to be later used by classifiers. Many classification tools, such as Support Vector Machines (SVMs), Decision Trees, and Genetic Algorithms can deal with high-dimensional data. However, the classification results from the above tools based on the original data are often too complex to be understood by customer managers. As a result, it is necessary to perform dimension reduction on the original data by removing those irrelevant features. Once the dimension of the data is reduced, the results from the classification tools may turn to be simpler and more explicable, which may be easier for bank staff to comprehend. On the other hand, it should be noted that the classification accuracy still needs to remain at an acceptable level after dimension reduction.

F. Sun et al. (Eds.): ISNN 2008, Part I, LNCS 5263, pp. 531–538, 2008.

2 Credit Data and Classification Models

The experimental data set (Australian Credit Approval Data Set) was taken from the UCI repository [2], which has 690 samples, each with 8 symbolic features and 6 numerical features. There are 2 classes (majority rate is about 55.5%) without missing feature values. The data set was randomly divided into training set (490 samples) and test set (200 samples). All numerical features were linearly scaled to be within [0, 1]. In this paper, the SVM model was employed as the classifier, which has been widely used in various classification tasks and credit assessment applications [3, 4, 5].

2.1 Preliminary Results

In order to use the SVM model, all symbolic features need to be transformed into numerical features. A simple and commonly used scheme is shown in Table 1. In this example, a symbolic feature S taking 3 possible values a, b, and c is transformed into 3 binary features (S1, S2, and S3).

Table 1. A simple way to transform symbolic features into numerical features

	S1	S2	S3
S=a	1	0	0
S=b	0	1	0
S=c	0	0	1

In the experimental studies, K-fold cross-validation was adopted [6] where the parameter K was set to 5. In the SVM model, the RBF kernel was used and its parameters were chosen based on a series of trials. The accuracies of the SVM were 86.7347% and 87.5% on the training set and the test set respectively. The implementation of the SVM was based on "libsvm-2.85" [7].

2.2 An Alternative Way to Handle Symbolic Features

There is an alternative way to transform symbolic features, which is based on the idea of probabilities [10]. Let t represent a symbolic feature and its possible values are defined as: $t_1, t_2, \ldots, t_k$. Let ω_i (i=1,2,…,M) denote the i^{th} class label.

For example, the case of $t=t_k$ is represented by:

$$\left(P(\omega_1 \mid t = t_k), P(\omega_2 \mid t = t_k), \ldots, P(\omega_M \mid t = t_k)\right)$$

Since the sum of probabilities should always equal to 1, each symbolic feature can be represented by M-1 numerical features. As a result, for two-class problems, each symbolic feature can be represented by a single numerical feature. Compared to the scheme in Table 1, this new scheme is favorable when the number of classes is small (two classes in this paper) while the cardinality of each symbolic feature is high. With this type of transformation of symbolic features in the credit data, the accuracies of the SVM were 86.939% and 88.0% on the training set and the test set respectively.