



Credit rating by hybrid machine learning techniques

Chih-Fong Tsai^{a,*}, Ming-Lun Chen^b

^a Department of Information Management, National Central University, 300 Jhongda Rd., Jhongli 32001, Taiwan

^b Taichung Commercial Bank, Taiwan

ARTICLE INFO

Article history:

Received 30 May 2008

Received in revised form 15 April 2009

Accepted 2 August 2009

Available online 8 August 2009

Keywords:

Credit rating

Consumer loans

Machine learning

Hybrid models

Maximum profits

ABSTRACT

It is very important for financial institutions to develop credit rating systems to help them to decide whether to grant credit to consumers before issuing loans. In literature, statistical and machine learning techniques for credit rating have been extensively studied. Recent studies focusing on hybrid models by combining different machine learning techniques have shown promising results. However, there are various types of combination methods to develop hybrid models. It is unknown that which hybrid machine learning model can perform the best in credit rating. In this paper, four different types of hybrid models are compared by 'Classification + Classification', 'Classification + Clustering', 'Clustering + Classification', and 'Clustering + Clustering' techniques, respectively. A real world dataset from a bank in Taiwan is considered for the experiment. The experimental results show that the 'Classification + Classification' hybrid model based on the combination of logistic regression and neural networks can provide the highest prediction accuracy and maximize the profit.

© 2009 Elsevier B.V. All rights reserved.

1. Introduction

As the financial market competes sharply nowadays, the traditional banking profit is largely reduced. This causes banks to focus on consumer banking in order to make higher interest profits, i.e. consumer loans. However, the quality of issuing consumer loans is bank dependent and the audit process is forced to be as simple as possible. Consequently, the potential risk gradually arises.

With the rapid growth in credit industry and the management of large loan portfolios, credit rating (or credit scoring) models have been extensively used for the credit admission evaluation. The credit rating models are developed to classify loan customers as either a good credit group (accepted) or a bad credit group (rejected) with their related characteristics such as age, income and marital status or based on the data of the previous accepted and rejected applicants [1]. The benefits of considering credit scoring include reducing the cost of credit analysis, enabling faster decisions, insuring credit collections, and diminishing possible risks [24]. Even if a slight improvement in credit scoring accuracy might reduce large credit risks and translate into significant future savings.

The traditional approach to predict the consumers' credit risk is based on some statistical methods, such as logistic regression.

However, related studies have shown that machine learning techniques or data mining techniques, such as neural networks, decision trees, etc., are superior to traditional (statistical) methods [2,9,22]. That is, using machine learning techniques can provide higher prediction accuracy.

In machine learning, the hybridization approach has been an active research area to improve the classification/prediction performance over single learning approaches [8,11,13,17,19]. In general, it is based on combining two different machine learning techniques. For example, a hybrid classification model can be composed of one unsupervised learner (or cluster) to pre-process the training data and one supervised learner (or classifier) to learn the clustering result or vice versa [18].

Therefore, to develop a hybrid learning credit model, there are four different ways to combine the two machine learning techniques. They are: (1) combining two classification techniques, (2) combining two clustering techniques, (3) one clustering technique combined with one classification technique, and (4) one classification technique combined with one clustering technique.

In literature, related work developing credit rating models based on hybrid machine learning techniques only compare with some chosen single learning based models as the baselines (see Section 2.4). That is, none of the existing studies compares different hybrid models to identify which hybrid approach can perform the best for credit rating in terms of high prediction accuracy and low error rates.

Therefore, the aim of this paper is to examine the prediction performance of these four types of hybrid learning models for

* Corresponding author. Tel.: +886 3 422 7151; fax: +886 3 425 4604.
E-mail address: cftsai@mgt.ncu.edu.tw (C.-F. Tsai).

credit rating in addition to the single classification and clustering techniques as the baseline models. Moreover, the profit made by these models is also compared based on a chosen bank as the case. Four well-known classification techniques, which are decision trees, Bayes classification, logistic regression, and neural networks, and two clustering techniques, which are K -means and expectation maximization are used to develop the hybrid models. Therefore, the contribution of this paper is to find out which combination method and techniques for the hybrid learning model can perform the best as well as provide maximum profits for credit rating.

The organization of this paper is as follows. Section 2 briefly describes related machine learning techniques used in this paper. In addition, related work is compared and their limitations are discussed. Section 3 presents the research methodology including the data used, the development of the credit rating models, evaluation strategies considered, etc. Section 4 shows the experimental results and the conclusion is provided in Section 5.

2. Machine learning techniques

2.1. Classification techniques

Classification (or supervised learning) techniques are based on learning by examples that map input vectors into one of several desired output classes. That is, a pattern classifier can be created through the training or learning process. The learning process of creating a classifier is to calculate the approximate distance between input–output examples and make correct output labels of the training set. This process is called the model generation phase. When the model is generated, it can classify an unknown instance into one of the learned classes in the training set [20].

2.1.1. Decision trees

A decision tree is a classification approach which is based on the tree structure to analyze data. One major advantage is that some decision rules can be produced that are easy to understand by humans. A decision tree classifies an instance by sorting it through the tree to the appropriate leaf node, i.e. each leaf node represents a classification. Each node represents some attribute of the instance, and each branch corresponds to one of the possible values for this attribute. C4.5 is the mostly used decision tree approach, and it is a later version of the ID3 algorithm [23].

2.1.2. Artificial neural networks

An artificial neural network, also called neural network, is composed of a group of neural nodes that link with the weighted nodes. Every node can simulate a neuron of creatures, and the connection among these nodes is equal to the synaptic that connects among the neurons. The most common type of neural networks consists of three layers of units: input layers, hidden layers, and output layers. It is called multilayer perceptron (MLP). A layer of “input” units is connected to a layer of “hidden” units, which is connected to a layer of “output” units [6].

2.1.3. Naïve Bayes classification

Naïve Bayesian classification [4] is based on Bayes theorem, which uses all kinds of beforehand probabilities and probabilities that are observed in the population to predict afterward probabilities. It is an effective tool to predict the relation of class members in the unknown situation.

The naïve Bayes classifier requires all assumptions be explicitly built into models which are then used to derive ‘optimal’ decision/classification rules. It can be used to represent the dependence between random variables (features) and to give a concise and tractable specification of the joint probability distribution for a domain. It is constructed by using the training data to estimate the

probability of each class given the feature vectors of a new instance.

2.1.4. Logistic regression

Logistic regression is a simply parametric statistical approach. It is similar to traditional regression analysis. Therefore, the use of logistic regression should also conform to some hypothesis of traditional regression analysis, such as to avoid the autocorrelation in residuals, to avoid multi-collinearity in independent variables, and the collected data must conform to a normal distribution [7].

Logistic regression uses a series of numerical computations to establish a model by known classification parameters to find out which parameters have the more discriminated ability for each group and the classification rules for each group. Logistic regression is the same as discriminant analysis, which is used to deal with the relationship between independent variables and dependent variables when the dependent variable is a list of categories to which objects can be classified. Therefore, the difference between logistic regression and discriminant analysis is that discriminant analysis must satisfy the assumption of normal distribution and the equal covariance matrixes to find out the optimal value. However, logistic regression does not need these assumptions, even if these assumptions are satisfied, logistic regression can still provide relatively high prediction accuracy.

2.2. Clustering techniques

Clustering (or unsupervised learning) techniques can be regarded as the process of grouping similar objects into a cluster. In particular, labeled examples are not available. The purpose of clustering techniques is to improve the similarity of the members in a group and make the data in each cluster have the highest similarity, but the highest dissimilarity between different clusters [20].

Clustering algorithms can be classified into two categories, which are hierarchical and partitional clustering algorithms [12]. Hierarchical clustering creates a hierarchy of clusters by using the agglomeration algorithm. Then, a distinct singleton cluster will be combined one by one until satisfying some rules. The result will produce a series of arborescence partitions. On the other hand, partitional clustering is much more popular, which have been extensively used in many business problems [21]. Two well-known partitional clustering algorithms are K -means and expectation maximization (EM) algorithms described below.

2.2.1. K -means

The K -means clustering algorithm is a simple and efficient clustering method. K -means clustering is performed based on the following steps [5]:

- Given a group of feature vectors (or data points) as the dataset to be clustered.
- Randomly select the amount of seed by k to be the cluster center.
- Assign the nearest data points to the clusters.
- Average the position of every data point in the clusters in order to find out the new cluster, and then every data point will be assigned to their nearest cluster center.
- Repeat the two previous steps until some convergence criterion is met or the assignment cannot be changed.

2.2.2. Expectation maximization

The EM algorithm performs as the following steps [3]:

- *Step 1: Estimation.* First, assume the average of cluster parameter μ^i ; standard deviation σ^i . Then, we can figure out the probability of p^i that every points to cluster q^i , and $q^i = 1$ where i represents

متن کامل مقاله

دریافت فوری ←

ISIArticles

مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات