Expert Systems with Applications 37 (2010) 7838–7843

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Vertical bagging decision trees model for credit scoring

Defu Zhang [a,*], Xiyue Zhou [a], Stephen C.H. Leung [b], Jiemin Zheng [a]

[a] Department of Computer Science, Xiamen University, Xiamen 361005, China
[b] Department of Management Sciences, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong, China

ARTICLE INFO

Keywords:
Credit scoring
Decision trees
Bagging
Classification

ABSTRACT

In recent years, more and more people, especially young people, begin to use credit card with the changing of consumption concept in China so that the business on credit cards is growing fast. Therefore, it is significative that some effective tools such as credit-scoring models are created to help those decision makers engaged in credit cards. A novel credit-scoring model, called vertical bagging decision trees model (abbreviated to VBDTM), is proposed for the purpose in this paper. The model is a new bagging method that is different from the traditional bagging. The VBDTM model gets an aggregation of classifiers by means of the combination of predictive attributes. In the VBDTM model, all train samples and just parts of attributes take part in learning of every classifier. By contrast, classifiers are trained with the sample subsets in the traditional bagging method and every classifier has the same attributes. The VBDTM has been tested by two credit databases from the UCI Machine Learning Repository, and the analysis results show that the performance of the method proposed by us is outstanding on the prediction accuracy.

1. Introduction

As an immense global industry, the credit card always maintains to develop rapidly in the last decades due to its huge profits. According to IMF statistics, the annual transactions of the five famous credit card organizations including Visa, Master, and American Express, etc., have increased to $ 4.74 trillion from $ 1.08 trillion between 1993 and 2003. In recent years, the credit card is also very popular with the changing of consumption concept in China; more and more people popularly use all kinds of credit cards. China's credit card market has reported remarkable growth over the past few years, albeit growing from a small base (RNCOS, 2009). As of the end of 2007, the circulation of credit cards had quadrupled and the outstanding level of credit card debt totaled RMB 75 billions, total line of credit is more than RMB 630 billion. China issued more than 50 million credit cards during 2008, taking the total number of credit cards in circulation to over 150 million. These numbers are projected to continue growing in the next few years, despite the current economic crisis (RNCOS, 2009). The volume of business on the credit card grows fast and is more enormous. Therefore, the decision makers need some help to decide whether to grant credit or not for a credit card applicant from some efficient and feasible tools.

Credit-scoring model is a good and effective tool for global financial institutions. Last few years, numerous credit-scoring models have been proposed in literatures to evaluate the consumer loans and improve the credit-scoring accuracy (Crook, Edelman, & Thomas, 2007). These models may be grouped into parametric statistical and non-parametric statistical or data-mining models. Generally, the linear discriminant analysis (LDA) and the logistic regression (LR) are the two most popular parametric statistical models to construct credit-scoring model. LDA is one of the first parametric statistical methods suggested for credit scoring (Reichert, Cho, & Wagner, 1983). This method is criticized because of the categorical property of the data and the fact that the covariance matrices of the good credit data are considerably distinct from that of bad credit data. The LR model used for credit-scoring application is proposed by Henley (1995), which allows overcoming these deficiencies and does not require the multivariate normality assumption. However, both LDA and LR are of low prediction accuracy in the credit scoring, as the relationship among variables is linear. To improve the less accuracy of parametric statistical methods, many models based on data-mining methods are built. These methods include the decision trees (Davis, Edelman, & Gammerman, 1992; Frydman, Altman, & Kao, 1985; Zhou, Zhang, & Jiang, 2008), artificial neural networks (Jensen, 1992; West, 2000; West, Dellana, & Qian, 2005), *k*-nearest neighbor (Henley & Hand, 1996), genetic programming (Abdou, 2009; Huang, Tzeng, & Ong, 2006; Ong, Huang, & Tzeng, 2005), genetic algorithm (Desai, Conway, Crook, & Overstreet, 1997; Walker, Haasdijk, & Gerrets, 1995; Zhang, Huang, Chen, & Jiang, 2007), case-based reasoning (Chuang & Lin, 2009; Jo, Han, & Lee, 1997; Park & Han, 2002), Artificial Immune System Algorithm (Leung, Cheong, & Cheong, 2007), rule extraction based on NN (Setiono, Baesens, & Mues,

* Corresponding author.
E-mail addresses: dfzhang@xmu.edu.cn (D. Zhang), elice.zhou@gmail.com (X. Zhou), mssleung@mail.cityu.edu.hk (S.C.H. Leung).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.04.054

2008), classification based the association rules (Li, Han, & Pei, 2001; Liu, Hsu, & Ma, 1998; Yin & Han, 2003) and support vector machines (Baesens et al., 2003; Gestel, Baesens, Garcia, & Dijcke, 2003; Huang, Chen, & Wang, 2007), etc. Among these data-mining methods, the decision tree, artificial neural networks and support vector machine are generally regarded as the most efficient single scoring models. Recently, some two-stage scoring models (Chuang & Lin, 2009; Huang et al., 2006; Lin, 2009) and hybrid scoring models (Hsieh, 2005; Lee & Chen, 2005; Zhang, Hifi, Chen, & Ye, 2008) are presented to overcome the shortcoming of the single scoring model. These models perform well and have shown promising results on the prediction accuracy.

For good classifiers, superior accuracy may be one of the most important performance measures. Some researches have shown that aggregating approach can easily achieve the improved accuracies by an aggregation of individual classifiers for credit scoring as well as the classification application. Hoffmann (2002) showed that the boosted genetic fuzzy classifier performed better than both the neurofuzzy classifier and the well-known C4.5 algorithm. In addition, the model by an ensemble of neural networks obtained the higher accuracy than the single neural networks in credit scoring and bankruptcy prediction application (West et al., 2005). In this paper, we propose a powerful credit-scoring model: called vertical bagging decision trees model (abbreviated to VBDTM), which is different from the model based on traditional bagging. The VBDTM model gets a set of classifiers by means of the combination of attributes. In this paper, we use the result of attributes reduction for the combination of attributes (Zhou et al., 2008). In the VBDTM model, all training samples and just parts of attributes take part in learning of every classifier, which is a vertical method. By contrast, classifiers are only trained with the sample subsets in the traditional bagging method and every classifier has the same attributes, which is a horizontal method. The VBDTM has been tested using two credit databases from the UCI Machine Learning Repository (Asuncion & Newman, 2007), and the analysis results show that the performance of our proposed model is outstanding on the prediction accuracy.

The rest of this paper is organized as follows. We will introduce vertical bagging decision tree model in the next section. Following this, the description of the credit data, the accuracy of our model and the comparison of the prediction accuracy of other models will be showed; and the performance of bagging will also be given. Section 4 addresses the conclusion and discusses the possible future research work.

2. Vertical bagging decision trees

2.1. Decision trees

A decision tree is a mapping from observations about an item to conclusion about its target value as a predictive model in data mining and machine learning. Generally, for such tree models, other descriptive names are classification tree (discrete target) or regression tree (continuous target). In these tree structures, the leaf nodes represent classifications, the inner nodes represent the current predictive attributes and branches represent conjunctions of attributions that lead to the final classifications. The popular decision trees algorithms include ID3 (Quinlan, 1986), C4.5 (Quinlan, 1993) which is an extension of ID3 algorithm and CART. The C4.5 will be simply introduced as follows.

C4.5 builds decision trees from a set of training data with every sample of that classified, using the concept of information entropy (Zhou et al., 2008). The training data is a set $S = (s_1, s_2, ..., s_n)$ and each sample $s_i = (x_1, x_2, ..., x_m, c_i)$ is a vector, where $x_1, x_2, ..., x_m$ represents predictive attributes or features of the sample and c_i represents the class of the sample s_i as a target attribute. At each inner node of the tree, C4.5 chooses one attribute that most effectively splits its set of samples into subsets. Its criterion is the normalized information gain (Zhou et al., 2008) that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is chosen to make the decision. Then the C4.5 algorithm is recursively executed for the smaller subsets. The C4.5 algorithm then recurses on the smaller subsets.

When the tree is built, it would have some base cases as follows (Quinlan, 1993).

1. All the samples in the subsets belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.
2. None of the attributes provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class.
3. Instance of previously unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value.

The Fig. 1 is an illustration of the structure of decision tree built by some credit database with the C4.5, where *x*, *y*, *z*, *u* in inner nodes of the tree are predictive attributes and "good" and "bad" are the classifications of target attribute in the credit database.

2.2. Rough sets

A rough set, first described by Pawlak (1991), is a formal approximation of a crisp set in terms of a pair of sets which give the lower and the upper approximation of the original set. One of important applications of rough sets theory is attribute reduction (Zhou et al., 2008) or feature selection. In rough sets theory, the sample data set is called information system, denoted $I = (U, A)$, where U is a non-empty finite set of observers called universe and A is a non-empty finite set of attributes (Zhou et al., 2008). After attribute reductions, we can get a series of subsets $A_1, A_2, ..., A_m \subseteq A$. In this paper, we applied the $A_1, A_2, ..., A_m$ as the different combinations of predictive attributes when building VBDTM.

2.3. Bagging

Bagging (Bootstrap aggregating) (Breiman, 1994) is a meta-algorithm to improve classification or regression models in terms of stability and classification accuracy in machine learning. Also it reduces the variance and helps to avoid overfitting. It can be used for any type of model such as NN (West et al., 2005), although it is most applied to decision tree models (Dietterich, 2000).

Given a training set S of size n, bagging gets m new sample subsets $S_1, S_2, ..., S_m \subset S$, by sampling observers from S uniformly and

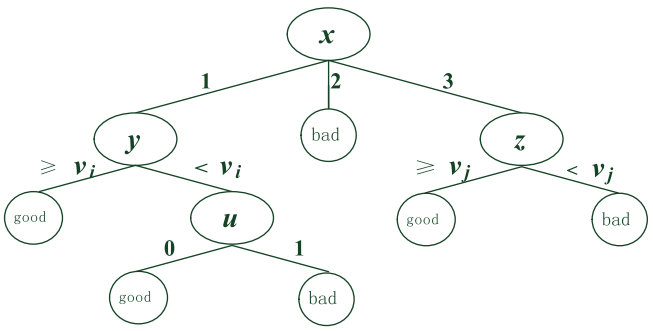

Fig. 1. A structure of decision tree.