Expert Systems with Applications 38 (2011) 11300–11304

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Neighborhood rough set and SVM based hybrid credit scoring classifier

Yao Ping, Lu Yongheng *

College of Economics & Management, Heilongjiang Institute of Science and Technology, Harbin 150027, China

ARTICLE INFO

Keywords:
Neighborhood
Rough set
SVM
Credit scoring

ABSTRACT

The credit scoring model development has become a very important issue, as the credit industry is highly competitive. Therefore, considerable credit scoring models have been widely studied in the areas of statistics to improve the accuracy of credit scoring during the past few years. This study constructs a hybrid SVM-based credit scoring models to evaluate the applicant's credit score according to the applicant's input features: (1) using neighborhood rough set to select input features; (2) using grid search to optimize RBF kernel parameters; (3) using the hybrid optimal input features and model parameters to solve the credit scoring problem with 10-fold cross validation; (4) comparing the accuracy of the proposed method with other methods. Experiment results demonstrate that the neighborhood rough set and SVM based hybrid classifier has the best credit scoring capability compared with other hybrid classifiers. It also outperforms linear discriminant analysis, logistic regression and neural networks.

1. Introduction

The credit industry has experienced two decades of rapid growth with significant advance in auto-financing, credit card debt, and so on. Meanwhile, credit scoring models have been widely studied in the area of statistics, machine learning, and artificial intelligence. The merits of credit scoring include reducing the cost of credit analysis, enabling faster credit decisions, closely monitoring the existing accounts, and reducing possible risk (Lee, Chiu, Lu, & Chen, 2002; West, 2000). With the growing of the credit industry and large loan portfolios under management today, credit industry is actively developing more accurate credit scoring models.

In the past, many researchers have developed a variety of traditional statistical methods for credit scoring. Linear discriminant analysis (LDA) and logistic regression are the two most commonly used statistical techniques in building credit scoring models. However, the utilization of linear discriminant analysis has often been criticized due to the assumptions of linear relationship between dependent and independent variables, which seldom holds in practice, and the fact that it is sensitive to deviations from the multivariate normality assumption (Karels & Prakash, 1987; Reichert, Cho, & Wangner, 1983). In addition to the LDA approach, logistic regression is another commonly utilized alternative to conduct credit scoring tasks. Basically, both LDA and logistic regression are designed for the case when the underlying relationship between variables is linear, and hence both are reported to be lacking in enough credit scoring accuracy (Thomas, 2000).

Artificial neural networks provide a new alternative to LDA and logistic regression in handing credit scoring, particularly in the case where the dependent and independent variables exhibit the complex non-linear relationship. It is, however, also being criticized for its long training process in obtaining the optimal network's topology, not easy identification of the relative importance of potential input variables, and certain interpretive difficulties, which hence has limited its applicability in handling general classification and credit scoring problems (Craven & Shavlik, 1997; Lee et al., 2002; Piramuhtu, 1999).

Rough set theory, proposed by Pawlak, is a novel mathematic tool handling uncertainty and vagueness, and inconsistent data (Pawlak, 1982a, 1991b). Rough set theory can discover data dependencies and reduce the number of attributes contained in a data set by purely structural methods. So, it is widely used in the area of feature selection and classification.

Recently, researchers have proposed the hybrid data mining approach in the design of the effective credit scoring models, such as the hybrid system based on clustering and neural network techniques, two-stage hybrid modeling procedure with artificial neural networks and multivariate adaptive regression splines, and the back propagation neural network integrated with the traditional discriminant analysis approach (Chen & Huang, 2003; Hsieh, 2005; Lee et al., 2002).

Support vector machines (SVM) is a new technique in the field of data mining, which is a new tool to solve machine-learning by means of optimization methods and is also a machine-learning algorithm based on statistical learning theory developed by Vapnik (1995). The traditional learning methods (e.g. NN) employ empirical

* Corresponding author.
E-mail address: hist_frank@126.com (L. Yongheng).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.02.179

risk minimization (ERM) principle so as to minimize the error of sample, and over-fitting happens inevitably, thus the model generalization is restricted. Whereas the statistic learning theory adopts structural risk minimization (SRM) principle, which minimizes the error of sample and also the upper bound of generalization error of the model, that is, minimizing the model's structural risk to improve the model's generalization. This merit highlights in small sample learning. This theory, which avoids the problems arising from such methods as NN, is taken as the best one in dealing with small sample classification and regression.

When using SVM, two problems are confronted: how to choose the optimal input feature subset for SVM and how to set the best kernel parameters. These two problems are crucial because the feature subset choosing influences the appropriate kernel parameters, and vice verse. In this paper, we proposed hybrid techniques based on three strategies: (1) using neighborhood rough set to select input features; (2) using grid search to optimize RBF kernel parameters; (3) using the hybrid optimal input features and model parameters to solve the credit scoring problem with 10-fold cross validation; (4) comparing the accuracy of the proposed method with other methods.

2. Basic methods

2.1. VNRS

Pawlak's rough set model is built on equivalence relations and equivalence classes. Equivalence relations can be directly induced from categorical attributes based on attributes values. The samples are said to be equivalent or indiscernible if their attribute values are identical to each other. However, some attributes in data are numerical in real-world applicants, such as credit scoring problem. Let's consider a two class problem, as Fig. 1. In the left plot, the sample space is divided into a set of information granules induced with some categorical attributes, where each box denotes an information granule of objects with the same feature values. The granules in the boundary are inconsistent because some of objects in these granules belong to X and the others do not belong to it.

A similar case can also be found in numerical feature spaces, as plot 2. We associate a neighborhood to each object in the sample, as x_1, x_2 and x_3. It is easy to find that the neighborhood of x_1 are completely contained in class 1, marked with "$*$", and the neighborhood of x_3 are completely contained in class 2, marked with "+", we say that x_1 and x_3 are the objects in lower approximations of 1 and 2, respectively. In the same time, the objects in the neighborhood of x_2 come from class 1 and 2. Then we define that the samples as x_2 are the boundary objects of the classification. Generally speaking, we hope to find a feature subspace where the boundary region is as little as possible because the samples in boundary region are inconsistent and are easily misclassified. Here we can find that numerical and categorical features can be unified into a framework. In this framework, categorical features generate equivalence information granules of the samples, and numerical features forms neighborhood information granules, and then they are both used to approximate the decision class in the framework of rough sets (Hu, Liu, & Yu, 2008a; Hu, Xie, & Yu, 2007b, 2008c).

2.2. SVM

The SVM developed by Vapnik implements the principal of structural risk minimization by constructing an optimal separating hyper plane: $w \cdot x + b = 0$.

To find the optimal hyper plane: $\{x \in S(w, x) + b = 0\}$, the norm of the vector needs to be minimized, on the other hand, the margin $1/||w||$ should be maximized between two classes. $min|(w, x) + b| = 1$.The solution for the typical two classes in linear problems has the form as shown in Fig. 2. Those circled points are called 'support vectors' for which $y_i(x_i\, w) + b = 1$ holds and which confine the margin. Moving of any of them will change the hyper plane normal vector w.

In the nonlinear case, we first mapped the data to some other Euclidean space H, using a mapping: $\Phi : R^d \to H$.

Then instead of the form of dot products, 'kernel function' K is issued such that $K(x_i, y_i) = \Phi(x_i) \cdot \Phi(y_i)$. There are several Kernel functions. Using a dual problem, the quadratic programming problems can be re-written as:

$$Q(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

$$\text{Subject to} : 0 \le \alpha_i \le C \sum_{i=1}^{l} \alpha_i y_i = 0$$

with the decision function:

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{l} y_i \alpha_i k(x, x_i) + b \right)$$

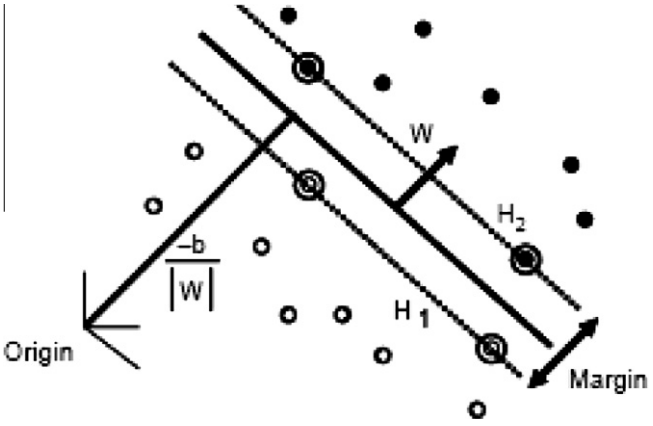
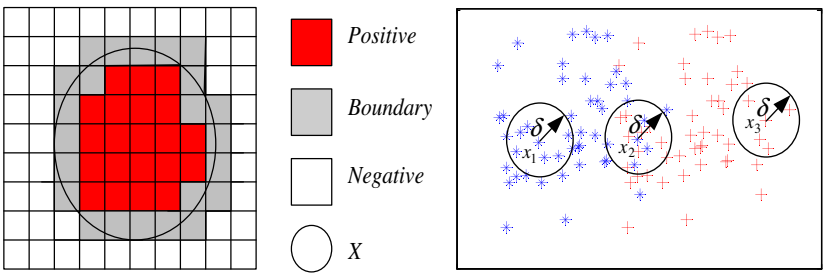


Fig. 2. Linear separating hyper planes for the separable case.



(1) Rough sets (2) Neighborhood rough sets

Fig. 1. Pawlak's rough sets and neighborhood rough sets.