2009 International Conference on Business Intelligence and Financial Engineering

Hybrid Classifier Using Neighborhood Rough Set and SVM for Credit Scoring

Ping Yao
School of Economics & Management,
Heilongjiang Institute of Science and Technology,
Harbin, 150027, China
12308157@sina.com

Abstract—**Credit scoring model development became a very important issue as the credit industry has many competitions. Therefore, most credit scoring models have been widely studied in the areas of statistics to improve the accuracy of credit scoring models during the past few years. This study constructs a hybrid SVM-based credit scoring models to evaluate the applicant's credit score from 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 in comparing with other hybrid classifiers. It also outperforms linear discriminant analysis, logistic regression and neural networks.**

Keywords-credit socring; neighborhood rough set; SVM; hybrid classifier;

I. INTRODUCTION

The credit industry has experienced two decades of rapid growth with significant increases in auto-financing, credit card debt, and so on. Credit scoring models have been widely studied in the area of statistics, machine learning, and artificial intelligence. The advantages of credit scoring include reducing the cost of credit analysis, enabling faster credit decisions, closer monitoring of existing accounts, and reducing possible risk [1]. With the growth of the credit industry and the 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 were 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, and the fact that it is sensitive to deviations from the multivariate normality assumption [2, 3]. 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 are linear and hence are reported to be lack of enough credit scoring accuracy [4, 5]

Artificial neural networks provide a new alternative to LDA and logistic regression in handing credit scoring tasks, particularly in situations where the dependent and independent variables exhibit complex non-linear relationships. it is, however, also being criticized for its long training process in obtaining the optimal network's topology, not easy to identify the relative importance of potential input variables, and certain interpretive difficulties and hence has limited its applicability in handling general classification and credit scoring problems [6-8].

Rough set theory, proposed by Pawlak, is a novel mathematic tool handling uncertainty and vagueness, and inconsistent data [9, 10]. Rough set theory can discover data dependencies and reduce the number of attribute 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 an effective credit scoring model. 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; integrating the back propagation neural network with traditional discriminant analysis approach [11-13].

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 a machine-learning algorithm based on statistical learning theory developed by Vapnik (1995). The traditional learning methods (e.g. NN) employed empirical risk minimization (ERM) principle so as to minimize the error of sample, and over-fitting appears inevitably, thus the model generalization is restricted. Whereas the statistic learning theory adopted structural risk minimization (SRM) principle, which minimizes the error of sample and also the upper bound of generalization error of the model, that is, minimize the model's structural risk to improve the model's generalization. This advantage 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.

978-0-7695-3705-4/09 $25.00 © 2009 IEEE
DOI 10.1109/BIFE.2009.41

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 choice 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.

II. BASIC IDEAS OF METHODS

A. Neighborhood rough set

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 figure.1.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 grannles 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 figure 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 [14-16].

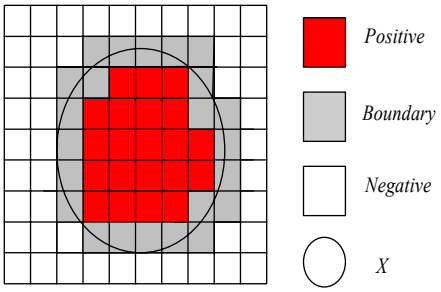

Figure1. Pawlak's rough sets

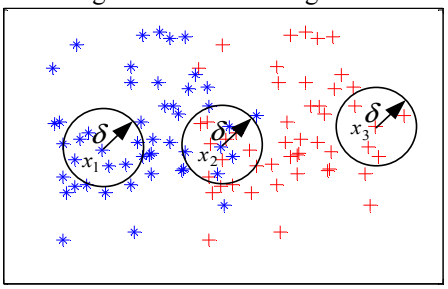

Figure 2. Neighborhood rough sets

B. SVM

The SVM developed by Vapnik implemented 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 figure 3.Those circled points are called 'support vectors' for which $y_i (x_i \cdot w) + b = 1$ holds and which confine the margin. Moving of any of them will change the hyper plane normal vector w.

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

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

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

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

quadratic programming problems can be re-written as:
with the decision function:

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

In this paper, we define the credit scoring as a none-linear problem and use Gaussian kernel to optimize the hyper plane.

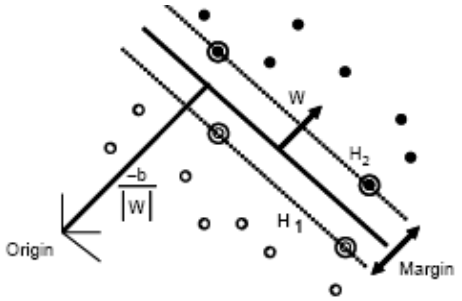

Figure 3. Linear separating hyperplanes for the separable case

In order to obtain robust results, it is important to set related parameters such as selection of kernels. The most often used kernels are linear, polynomial, radial basis function, sigmoid as follows:

Linear kernel: $K(x_i, x_j) = x_i^T x_j$

Polynomial: $K(x_i, x_j) = (\gamma(x_i^T x_j)+c)^q$

Radial basis function: $K(x_i, x_j) = \exp(-||x_i-x_j||^2/|\sigma^2)$

Sigmoid function: $K(x_i, x_j) = \tanh(\gamma(x_i^T x_j)+r)$

III. HYBRID SVM MODEL USING NEIGHBORHOOD ROUGH SET AND GRID SEARCH

A. Feature selection using neighborhood rough set

In this section, we present a fast feature selection algorithm based on neighborhood rough set for credit scoring, for more details, please refer to references [14-16].

Consider a decision table $<U, C \cup D, V, f>$, where A is the set of condition attributes, D is the decision. $B \subset A$, $a \in A - B$, the significance of attribute a relative B and D is defined as: $SIG_1(a, B, D) = \gamma_{B \cup a}(D) - \gamma_B(D)$.

SIG(a, B, D) reflects the increment of dependency, which means the positive region increases if we add attribute a in B, the increment of positive region is the significance of attribute a. It's easy to know $0 \le$ SIG (a, B, D) ≤ 1. If SIG (a, B, D) = 0, we say a is superfluous, which means a is useless for B to approximate D.

Formally, a forward greedy algorithm for mixed feature reduction can be formulated as follows.

Algorithm: forward attribute reduction based on variable precision neighborhood rough set

Input: Hybrid decision table<U, $A^c \cup A^n \cup D$ > and β and d // A^c and A^n are categorical and numerical attributes, respectively.

// β is the threshold for computing variable precision lower approximations; d is the radius of neighborhoods

Output: One reduct red

Step 1: $\forall a \in A^c$: compute equivalence relation R_a;

$\forall a \in A^n$: compute neighborhood relation N_a;

Step 2: $\varnothing \rightarrow$ red; // red is the pool to contain the selected attributes

Step 3: For each $a_i \in A - red$

Compute $\gamma_{red \cup a_i}(D) = \dfrac{|POS_{red \cup a_i}|}{|U|}$

Compute $SIG(a_i, red, D) = \gamma^{\beta}_{red \cup a_i}(D) - \gamma^{\beta}_{red_i}(D)$

// we define $r^{\beta}_{\varnothing} = 0$

end

Step 4: select the attribute ak which satisfies: SIG (a_k, red, D) = max_i (SIG (a_k, red, D))

Step 5: If SIG (a_k, red, D)$>\varepsilon$ //ε is a positive real number use to control the convergence

$red \cup a_k \rightarrow red$

go to step 3

else

return red

Step 6: end

B. Process of the Hybrid SVM model for credit scoring

Credit scoring is, in fact, classification, and the score (good, bad) is treated as classification label and the applicant's personnel condition as classification attributes. Credit scoring procedure based on SVM is, data collecting and preprocessing, selection of input features, selection of a kernel, training, optimal SVM classifier i.e. SVM with optimal kernel and corresponding parameters, tests, credit scoring for new samples.

In this paper, we use neighborhood rough set to select the input features; grid search algorithm is adopted to obtain the optimal parameters, for each pair of parameters, 10-fold cross validation is conducted on the training set. The algorithm of our credit scoring model is as follows:

Step 1: Use neighborhood rough set to obtain the variable importance.

Step 2: Drop redundant features; rebuild data set and randomly partition into a training set and testing set.

Step 3: Choose a kernel function (RBF).

Step 4: Consider a grid space of (C, γ) with $\log_2$C {-5,-3,-1,…,13} and $\log_2\gamma$ {-13,-11,-9,…,5}, for each pair of parameters, conduct 10-fold cross-validation on the training set, in the neighborhood of the parameters (C, γ) that leads to the lowest CV error classification rate, choose a fine grid, and repeat this step.

Step 5: Choose parameters (C, γ) that lead to the lowest CV error classification rate to create a model as the classifier.

IV. EMPIRICAL ANALYSIS

A. Real world credit data set

The two real world data sets illustrated in table I, the Australian and German credit data sets, are available from the UCI Repository of Machine Learning Databases and are adopted herein to evaluate the predictive accuracy. The Australian credit data consists of 307 instances of creditworthy applicants and 383 instances where credit is not creditworthy. Each instance contains 6 nominal, 8 numeric

attributes, and 1 class attribute (accepted or rejected). The German credit scoring data more unbalanced and it consists of 700 instances of creditworthy applicants and 300 instances where credit should not be extended. For each applicant, 24 input variables describe the credit history, account balances, loan purpose, loan amount, employment status, personal information, age, housing, and job title. This data set only consists of numeric attributes.

TABLE I. DATASETS FROM THE UCI REPOSITORY

Names	# classes	# instances	Nominal features	Numeric features
German	2	1000	0	24
Australian	2	690	6	8

B. Comparing the feature selection algorithm

Feature selection is very important step to improve the prediction accuracy of hybrid model, this paper comparing the neighborhood rough set with 6 other feature selection methods, they are t-test, correlations matrix, stepwise regression, Pawlak's Rough Set, Classification and regression tree (CART) and Multivariate adaptive regression splines (MARS). Results of features selected with different methods for Australian and German data set are shown in table II and table III respectively.

TABLE II. FEATURES SELECTED WITH DIFFERENT METHODS FOR AUSTRALIAN DATA SET

method	Features selected
t-test	2,3,4,5,6,7,8,9,10,12,13,14
correlations	2,3,4, 7,8,9,10,14
stepwise	7,8,9,12,14
CART	5,7,8,9,10,14
MARS	3,5,7,8,10,14
Pawlak's Rough Set	1,2,3,4,6,7,8,9,11,14
Neighborhood rough set	8,9,14

TABLE III. FEATURES SELECTED WITH DIFFERENT METHODS FOR GERMAN DATA SET

method	Features selected
t-test	1,2,3,4,5,6,7,8,9,10,12,13,14,15,20
correlations	1,2,3,5,6,7,12,13,14,15,20
stepwise	1,2,3,7,8,15,20
CART	1,2,3,4,5,9,10,18
MARS	1,2,3,4,5,6,9,15,16,17,20
Pawlak's Rough Set	-
Neighborhood rough set	1,2,3,4,6,7,8,9,11,12,13,14

C. Experiment results

After feature selection step, we employ CART, RBF-SVM and KNN (K=5) classifier as validation function to compare the four feature selection algorithms above. All of the results are obtained with 10-fold cross validation. Table 4-5 shows the comparisons of accuracies with the different feature selection methods for Australian data set and German data set respectively. From table II-V, we can find all of the feature selection algorithms can remove parts of the candidate features while keep or improve classification accuracies in most of the cases. Comparing the accuracies in table IV-V, we can find that the selected features based on neighborhood rough set outperform those feature selection methods. Especially, although neighborhood method deletes most of the candidate features, average classification accuracy also improve. It shows that neighborhood rough set is able to find the most informative features for classification.

TABLE IV. ACCURACIES WITH THE DIFFERENT FEATURE SELECTION METHODS FOR AUSTRALIAN DATA SET

method	CART	RBF-SVM	5NN
t-test	0.8276	0.8392	0.8538
correlations	0.8306	0.8639	0.8379
stepwise	0.8203	0.8537	0.8232
CART	0.8145	0.8465	0.8363
MARS	0.8261	0.8509	0.8407
Pawlak's Rough Set	0.8131	0.8530	0.8407
Neighborhood rough set	0.8409	0.8752	0.8481

TABLE V. ACCURACIES WITH THE DIFFERENT FEATURE SELECTION METHODS FOR GERMAN DATA SET

method	CART	RBF-SVM	5NN
t-test	0.6890	0.7070	0.7220
correlations	0.7020	0.7210	0.7200
stepwise	0.6910	0.7280	0.7230
CART	0.7020	0.7110	0.7170
MARS	0.6970	0.7230	0.7300
Pawlak's Rough Set	-	-	-
Neighborhood rough set	0.6920	0.7660	0.7370

D. Comparing with other classifiers

In order to evaluate the effectiveness of the proposed hybrid credit scoring classifier, the classification results are also compared with those using linear discriminant analysis, logistic regression and neural networks. It can be concluded, from table VI, that the neighborhood rough set and SVM based hybrid credit scoring classifier has the best credit scoring capability in terms of the overall classification rate.

TABLE VI. ACCURACIES WITH THE DIFFERENT METHODS FOR AUSTRALIAN AND GERMAN DATA SETS

method	Australian	German
linear discriminant analysis	0.8520	0.6600
logistic regression	0.8570	0.7240
neural networks	0.8683	0.7520
The proposed Hybrid classifier	0.8752	0.7660

V. CONCLUSIONS

Credit scoring is a widely used technique that helps banks decides whether to grant credit to consumers who submit an applicant. Constructing the credit scoring model from a credit database can be taken as a task of data mining. The statistical classification models perform favorably only when the essential assumptions are satisfied. In contrast to traditional statistical techniques, the artificial intelligence

techniques do not require the knowledge of the underlying relationships between input and output variables. SVM is a modern data mining technique and suitable for classification.

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. 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. Experiment results demonstrate that the neighborhood rough set and SVM based hybrid classifier has the best credit scoring capability in comparing with other hybrid classifiers. It also outperforms linear discriminant analysis, logistic regression and neural networks.

REFERENCE

[1] D. West, "Neural network credit scoring models", Computers and Operations Research, vol.27, pp. 1131-115, 2000.

[2] G. Karels and A. Prakash, "Multivariate normality and forecasting of business bankruptcy". Journal of Business Finance Accounting, vol. 14, pp. 573-593, 1987.

[3] A. K. Reichert, C. C. Cho and G. M. Wangner, "An examination of the conceptual issues involved in developing credit-scoring models". Journal of Business and Economics Statistics, vol.1, pp. 101-114, 1983.

[4] L. C.Thoms, "A survey of credit and behavioral scoring: Forecasting financial risks of lending to customers". International Journal of Forecasting, vol. 16, pp. 149-172, 2000.

[5] D. West, "Neural network credit scoring models". Computers and Operational Research, vol. 27, pp. 1131-1152, 2000.

[6] M. W. Craven, & J. W. Shavlik, "Using neural networks for data mining".Future Generation Computer Systems, vol. 13, pp. 221-229, 1997.

[7] T. S. Lee, C.C. Chiu, C.J. Lu, & I. F. Chen, "Credit scoring using the hybrid neural discriminant technology". Expert Systems with Applications, vol. 23, pp. 245-254, 2003.

[8] S. Piramuhtu, "Financial credit-risk evaluation with neural and neurofuzzy systems". European Journal of Operational Research, vol. 112, pp. 310-321, 1999.

[9] Z. Pawlak, "Rough sets". International Journal of Computer and Information Sciences , vol.11, pp. 341–356, 1982.

[10] Z. Pawlak, Rough sets: theoretical aspects of reasoning about data. Kluwer Academic Publishing, 1991.

[11] H.C. Hsieh, "Hybrid mining approach in the design of credit scoring models". Expert System with Applications, vol. 28, pp. 655-665, 2005.

[12] T. S. Lee, C. C. Chiu, C. J. Lu, & I. F. Chen, "Credit scoring using the hybrid neural discriminant technique". Expert System with Applications, vol. 23, pp. 245-254, 2002.

[13] M. C. Chen, & S. H. Huang, "Credit scoring and rejected instances reassigning through evolutionary computation techniques". Expert System with applications, vol. 24, pp 433-441, 2003.

[14] Q.H. Hu, Z.X. Xie, D.R. Yu, "Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation", Pattern Recognition, vol. 40, pp. 3509-3521, 2007.

[15] Q.H. Hu, J.F. Liu, D.R. Yu, "Mixed feature selection based on granulation and approximation", Knowledge-Based Systems, vol. 21, pp. 294-304, 2008.

[16] Q.H. Hu, D.R. Yu, Z.X. Xie, "Neighborhood classifiers", Expert Systems with Application, vol. 34, pp. 866-876. 2008.