2007 IEEE International Conference on Control and Automation
Guangzhou, CHINA - May 30 to June 1, 2007

Satisfactory Feature Selection and Its Application in Enterprise Credit Assessment

Jian Ling
Automation Department
Xiamen University
Xiamen, China
jling@xmu.edu.cn

Chengde Lin
Automation Department
Xiamen University
Xiamen, China
cdlin@xmu.edu.cn

Abstract—**The selection of evaluating index system is one of the key problems in enterprise credit assessment. It is essentially a satisfactory feature selection (SFS) problem. In this paper, several novel satisfactory-rate functions of feature set (SRFFS) are designed, in which the classification performance of the feature subset and its size are considered compromisingly. The accuracy of SVM Cross Validation is employed as evaluation criterion of classification ability, and the SFS algorithm is described in detail. Contrastive experiments are carried on SFS and three other different feature selection methods: S-SFS, Expert+GAFS and GAFS. Results show that SFS, which can pick out the feature subset with low dimension, high classification accuracy and balanced ranking performance, is superior to three other ones.**

Keywords-enterprise credit assessmen; feature selection; satisfactory optimization; support vector machine (SVM)

I. INTRODUCTION

Credit assessment is a key step in loan business of commercial banks. The objective of credit assessment is to decide credit ranks which denote capacity of enterprises to meet their financial commitments. Credit ranks are gained from some evaluating models by assessing the financial statement. Recently, some intelligent classification models have been introduced to credit assessment field [1]-[2]. But there are so many financial indexes in the statement that intelligent models do not performance well. Firstly, complexity of the models and the modeling time are increased by the high dimensional index set. Secondly, high dimensionality leads to low generalization ability of the classification algorithms. So to select a few key financial indexes becomes the chief problem when intelligent models are used in credit assessment.

The selection of financial indexes is a feature selection problem. Feature selection is essentially a satisfactory optimization problem according to its NP hard characteristic and its dependence to the adopted evaluation criterion. So we can only attain suboptimal solution in practical application [3]. The most important part of satisfactory feature selection (SFS) is to define the evaluation criterion of feature set, or the so-called Satisfactory-Rate Function of Feature Set (SRFFS). Several definitions have been brought out in papers [4]-[5] according to their different application backgrounds and research methods.

In this paper, several novel SRFFSs are designed. In these functions, dimension of feature set and its classification performance on each credit rank are considered compromisingly. Support vector machine (SVM) is used as classifier, adapted to the small sample-size and unbalanced data offered by a domestic bank. Contrastive experiments are carried on SFS and three other different feature selection methods: S-SFS, Expert+GAFS [1] and GAFS [6]. Results show that SFS is superior to three others in quality of the selected feature subset, such as the size, the performance on individual rank and so on.

II. SUPPORT VECTOR MASHINE (SVM) AND ITS PARAMETER SELECTION

A. Brief on SVM

In its basic form, SVM learn linear decision rule $f(\boldsymbol{x})=sgn\{\boldsymbol{w}^T\boldsymbol{x}+b\}$ described by a weight vector $\boldsymbol{w}$ and a threshold b. Given training vectors $x_i \in \mathbf{R}^n$, $i = 1, \ldots, L$, in two classes, and a vector of labels $y \in \mathbf{R}^L$ such that $y_i \in \{1,-1\}$,computing this hyperplane is equivalent to solving the following optimization problem(OP1):

$$\begin{aligned} \min_{w,b,\varepsilon} \quad & \frac{1}{2} w^T w + C\sum_{i=1}^{L} \xi_i, \\ s.t \quad & y_i(w^T \cdot x_i + b) \geq 1-\xi_i, \\ & \xi_i \geq 0, i = 1,...,L, \end{aligned} \tag{1}$$

where ξ_i is the slack variable. The factor C is a parameter that allows one to trade off training error vs. model complexity. Instead of solving OP1 directly, it is easier to solve its dual form, (OP2):

$$\begin{aligned} \max \quad & \sum_{i=1}^{L} \alpha_i - \frac{1}{2}\sum_{i=1}^{L}\sum_{j=1}^{L} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j), \\ s.t \quad & \sum_{i=1}^{L} y_i \alpha_i = 0 \\ & 0 \leq \alpha_i \leq C, i = 1,...,L. \end{aligned} \tag{2}$$

The above SVM is an essentially linear method, but it can be easily generalized to non-linear decision rules by replacing the inner-products $(\boldsymbol{x}_i,\boldsymbol{x}_j)$ with a kernel function $k(\boldsymbol{x}_i,\boldsymbol{x}_j)$ in (2) [7].

B. Application of SVM to Enterprise Credit Assessmnet

SVM is an effective classifier for problems with small sample set, such as credit assessment. As the sample in different credit ranks is unbalanced, the separating hyperplane will lean to the rank with low sample density when the same C

is used to all the ξ_i as in (1). This defect will cause bad classification performance on prediction [2]. Thus, different penalty coefficients should be attached to the ξ_i corresponding to different ranks. Then the format of SVM in (1) is modified to

$$\min \quad \frac{1}{2} w^T w + C(\beta_i \sum_{y_k=i} \xi_k + \beta_j \sum_{y_k=j} \xi_k), \tag{3}$$

with the same constraints [8]. The ratio of β_i to β_j is determined by the sample-scale proportion of i^{th} rank and j^{th} rank.

Enterprise credit assessment is a multi-classification problem, and 1-vs-1 strategy will be used in this paper to combine SVMs.

C. Parameters Selection for SVM

In our experiments, RBF kernel is used in SVM models:

$$K(x_i, x_j) = \exp(-\gamma \left\| x_i - x_j \right\|^2) \tag{4}$$

Thus, there are two unknown parameters, C and γ, in our model. They will be determined by a validation process as the following:

- Consider a grid space of (C,γ) with $\log_{10} C \in \{-3,-2,...,3\}$ and $\log_2 \gamma \in \{-14,-12,...,2\}$.
- For each pair (C, γ) in the search space, conduct 5-fold cross validation (CV) on the training set.
- Choose the pair (C, γ) that leads to lowest CV error rate.
- Use the best parameter-pair to create a model as the predictor [9].

III. Satisfactory Feature Selection (SFS) for Enterprise Credit Assessment

A. Satisfacory-rate Function of Feature Set (SRFFS)

Let $U=[t_1,t_2,\ldots, t_z]$ be the original feature set with size z, $T=[t_1,t_2,\ldots, t_d]$ a subset of U with size d, $1 \le d \le z$, and M be the total number of credit ranks. In the following, for $1 \le i \le M$, t_i is the total number of training samples in i^{th} rank; a_i and b_i denote the numbers of training samples in i^{th} rank correctly classified by SVM 5-fold-CV on T and on U, respectively.

Definition 1. The Classification-Ability-SRFFS (C-SRFFS) s_c on the feature subset T is defined as:

$$\begin{aligned} s_c &= g(s_{c1}, s_{c2}, \cdots, s_{cM}) \\ &= \sum_{i=1}^{M} \omega_{ci} s_{ci} \Big/ \sum_{i=1}^{M} \omega_{ci}, \end{aligned} \tag{5}$$

where s_{ci} is the satisfactory-rate of classification ability on T for i^{th} credit rank, $s_{ci} = (a_i - b_i)/(t_i - b_i)$, and if $t_i = b_i$, $s_{ci} = a_i / t_i - 1$; ω_{ci} is the weight of s_{ci}, $\omega_{ci} = 1 - b_i/t_i$.

When $a_i=t_i$, i.e. the total training samples of i^{th} rank are correctly classified on T, then $s_{ci}=1$; when $a_i=b_i$, i.e. the number of samples (in i^{th} rank) correctly classified on T is equal to that on U, then $s_{ci}=0$; when $a_i<b_i$, i.e. the number of samples correctly classified on T is smaller to that on U, then $s_{ci}<0$, $|s_{ci}|$ represents the unsatisfactory rate on T.

If $t_i=b_i$, the classification ability on T is unsatisfactory unless $a_i=t_i$. So under this especial situation, s_{ci} is defined as $a_i/t_i - 1$.

Intuitively, heave weight should be attached upon the rank with high classification error. So we define ω_{ci} as $1 - b_i/t_i$, which is the error rate of i^{th} rank from 5-fold CV.

Definition 2. The Size-SRFFS (S-SRFFS) s_d on T is defined as

$$s_d = f(d) = (d_{\max} - d)/(d_{\max} - d_{\min}), \tag{6}$$

where $d_{\max}$ is the size of original feature set and $d_{\min}$ is the minimums size of the feature subset expected to be selected.

S-SRFFS is designed to automatically decide the size of feature subset in the searching process. This approach can improve the weakness of most existing feature selection methods which need to fix the size of feature subset before optimization.

Definition 3. Based on Definitions 1 and 2, the Integrated SRFFS (I-SRFFS) S is defined as

$$S = \varphi(s_d, s_c) = (\omega_c s_c + \omega_d s_d)/(\omega_c + \omega_d), \tag{7}$$

where ω_d and ω_c are the weight coefficients of C-SRFFS s_c and S-SRFFS s_d. Here, $S \le 1$; when $S<0$, $|S|$ represents the unsatisfactory rate on T.

B. Feature Selection Algorithm

Feature selection is a typical combinatorial optimization problem. There are 2^d subsets when selection is carried on an original feature set with d dimension (size). Therefore, Genetic Algorithm (GA) is often used in the optimizing process. In our problem, binary coding is used to create individuals. The dimension of each individual is equal to the size of the original feature set. Other words, an individual is represented by a 0-1 string, where '1' denotes that the feature corresponding to this bit is selected and '0' denotes not being selected. We take I-SRFFS (see (7)) as the fitness function of GA. The detailed algorithm is as the following:

1) For the SVM model with i^{th} rank and j^{th} rank in 1-vs-1 strategy, calculate the β_i and β_j (see (3)) according to the sample-scale proportion of these two ranks.

2) Find the best (C, γ) for SVM models as described in section Ⅱ.C.

3) Set $d_{\max}=z$ (size of the original feature set) and $d_{\min}=2$ in Definition 2.

4) Do 5-fold CV on the original feature set, and calculate ω_{ci} (see (5)).

5) Implement GA to find the satisfactory feature subset on the chosen parameters (number of the initial population, crossover and mutation proportion, maximum generation…).

IV. Emulational Experiments

A. Pretreatment of samples

In our experiments, there are 1,143 examples of enterprises taken from a commercial bank's client database (which is in FuJian Province), which include the financial statement and their credit ranks given by the bank. There are five credit ranks totally: AAA, AA, A, B and C, corresponding to the credit

degree from excellent to bad. The indexes in the financial statements show great difference in different cases. So we use financial ratios gained from these indexes to eliminate the effect, such as scales, industries, and so on. According to the suggestion of financial experts, we adopt 24 financial ratios as the original feature set. These ratios and their numbers are shown in TABLE Ⅰ.

In our data, the samples in B and C ranks are too small to establish effectual SVM model. Thus, we are forced to add them into rank A, and re-categorize the total sample into three ranks, noted as AAA, AA and ABC. We randomly select 75% examples in each rank as training sample for feature selection, and the rest 25% as testing sample to verify the predicting performance of the selected feature subset. The details about our data are: 857/286(training/testing) in total sample set, including 437/138 in AAA rank, 347/124 in AA rank, and 73/24 in ABC rank.

B. Experiment and Results Analysis

All the parameters needed in our experiments are calculated according to the description in section Ⅲ.*B*, and shown in TABLE Ⅱ. It is obvious that the choice of weight coefficients ω_d and ω_c in I-SRFFS (see (7)) would affect the result of experiment. According to the experiences and the result of plenty experiments, ω_d =0.1 and ω_c =0.9 are appropriate choices.

For comparison, three other different feature selection methods are introduced. The first one is denoted as S-SFS, which is actually a simplification of SFS (see Ⅲ). In S-SFS, the C-SRFFS (see (5)) s_c is re-defined simply as: $s_c = (a-b)/(t-b)$(if $t-b$, $s_c = a/t-1$), where t is the total number of training samples; a and b denote the numbers of training samples correctly classified by SVM 5-fold-CV on the selected feature subset T and on the original set U, respectively. The selection algorithm of S-SFS does not need to calculate ω_{ci} (see (5)).

TABLE Ⅰ. FINANCIAL RATIOS AND THEIR NUMBERS

No.	Financial Ratios
1	assets-liabilities ratio
2	current ratio
3	current assets turnover times
4	profit to sales ratio
5	rate of return on total assets
6	times interest earned
7	working capital ratio
8	Operational cash ratio
9	rate of return on equity
10	growth rate of net asset
11	rate of Contingent Liabilities
12	rate of sales of finished products
13	ratio of income sale to loan share
14	deposit-loan ratio
15	repayment interest rate
16	repayment rate of maturity credit
17	liquid ratio
18	ratio of fixed assets to total assets
19	ratio of pre-interest taxable income to total liabilities
20	ratio of pre-interest taxable income to working capital
21	ratio of sales revenue to total assets
22	ratio of current liabilities to net assets
23	ratio of inventories to sales revenue
24	ratio of net cash flow to total liabilities.

TABLE Ⅱ. PARAMETERS IN THE EXPERIMENT

For SVM	(C, γ)	(β_1, β_2)	(β_1, β_3)	(β_2, β_3)
—	$(10^2, 2^{-2})$	(1,1.2594)	(1,5.9863)	(1,4.7534)
For SRFFS	ω_{c1}	ω_{c2}	ω_{c3}	—
—	0. 2362	0. 3558	0. 3145	—

The second one (denoted as Expert+GAFS in this paper) is available in [1], which considers the financial experts' suggestion and the classification accuracy compromisingly in the feature selection process. The last one is a total GA method (denoted as GAFS) available in [9][1].

We set the same GA parameters in these four methods: the size of initial population is 50, the proportion of crossover is 0.95, the proportion of mutation is 0.08, and the maximum generation is 100. All the methods are implemented 20 times and the average results are shown in TABLE Ⅲ. As a comparison, the performance of the original feature set is also listed in the last line.

From TABLE Ⅲ, we can see that the size of the feature set is largely reduced after feature selection, no matter which method is used. Synchronously, the testing accuracies on feature subsets selected by the four methods are both higher than that of the original set. It is obvious that feature selection can eliminate interferential and redundant features and thus improve the classification performance. Furthermore, SFS and S-SFS get fewer features than the other two, which owes to the S-SRFFS (see (6)) they used. S-SRFFS can restrict the size of selected feature subset in the selection process, and will accordingly lead to many advantages for the subsequent modeling, such as low computational complexity, good generalization ability and so on .

For the testing accuracy shown in TABLE Ⅲ, We can see that the original set does bad in AA rank, which is 11.72% less than AAA rank and 4.84% less than ABC rank. After feature selection, this phenomenon does not ameliorate by the methods besides SFS. SFS, which attaches heavy weight on ranks with bad performance (see C-SRFFS in (5)), can select features on its own initiative to improve the performance on these ranks. Just because SFS considers the capability of feature subset on each credit rank, it shows the most balanced performance over three credit ranks and synchronously remains high total testing accuracy. The accuracy variance of SFS is only $4.0438*10^{-4}$. It is much smaller than that of the three others, which only require high total accuracy in the selection algorithm. Therefore the feature subset selected by SFS is more practical than that selected by the three other ones.

The reason why SFS is inferior to GAFS in total testing accuracy may due to the characteristic of SFS, which trades off

[1] To make an impartiality comparison, we also use SVM 5-fold CV accuracy as the evaluation criterion of classification ability on feature subset in Expert+GAFS and GAFS.

TABLE III. QUALITY COMPARISON BETWEEN ORIFINAL FEATURE SET AND THE FEATURE SUBSETS SELECTED BY FOUR METHODS

Methods	Average Size of The Selected Feature Sets	Average Testing Accuracy (%)				Variance of Average Testing Accuracy in Three Ranks (10^{-4})
		Total	*AAA*	*AA*	*ABC*	
SFS	12.55	77.63	80.02	75.14	77.00	4.0438
S-SFS	12.80	77.59	83.48	71.22	76.56	24.5556
EXPERT+GAFS	14.15	76.64	81.19	71.80	75.68	14.8430
GAFS	15.50	78.16	85.02	70.43	78.43	35.5885
NONE(ORIGINAL)	24	76.22	81.88	70.16	75.00	23.1243

the size of feature subset and its classification performance in selection process. SFS may reduce the dimension of feature subset at the expense of tiny decrease in total accuracy.

V. CONCLUSION

The selection of the financial ratios, which is one of the key problems for enterprise credit assessment, is essentially a satisfactory feature selection (SFS) problem. In this paper, some novel satisfactory-rate functions of feature set (SRFFS) are defined, in which the classification performance of the feature set and its size are considered compromisingly, and SVM technique is used to build intelligent models. Many skills are used in the definition, such as SVM cross validation and separating SRFFSs designed on each individual credit rank. Three other feature selection methods: S-SFS, Expert+GAFS [1] and GAFS [6] are introduced to do contrastive experiment. Performance on the original feature set and the selected subsets (including four methods) is compared after experiment. The results show that SFS method, which can select feature subset with low dimension, high classification accuracy and balanced ranking performance, is superior to three other ones.

REFERENCES

[1] Wei Yang. "Credit Rating Assessment of Enterprises Based on Claddifiers Combination," XMU Knowledge Resource Portal, Xiamen University,2005.

[2] Min Liu, Chengde Lin."A Model Based on Support Vector Machine for Credit Risk Assessment in Commercial Banks ," Journal of Xiamen University (Natural Science).Vol.44(1),pp. 29-32, Jan.2005.

[3] Gexiang Zhang, Weidong Jin, Laizhao Hu."Statisfactory feature selection and its applications," Control Theory and Applications, Vol.23(1),pp.19-24, Feb. 2006.

[4] Fan Jin, Fei Hu."Principles of satisfactory-solution for fuzzy neural computing," J of the China Railway Society, Vol.18(2),pp.102-107,1996.

[5] Yugeng Xi."Satisfactory control of complex industrial process,"Information and Control, Vol.24(1), pp.14-20, 1995.

[6] Yun Zhao, Weiyi Liu. "Research on Feature Selection Using Genetic Algorithms," Computer Engineering and Applications, Vol.15, pp.52-54, 2004.

[7] Joachims T."Estimating the generalilzation performance of an SVM efficienly,"Proceeding of the 17th International Conference on Machine Learning. San Francisco: Morgan, pp.431-438, 2000.

[8] E. Osuna, R. Freund, F. Girosi. "Support vector machines: Training and applications." AI Memo 1602, Massachusetts Institute of Technology, 1997b.

[9] Yiwei Chen, Chihjen Lin. "Combining SVMs with various feature selection strategies," http://www.csie.ntu.edu.tw/~cjlin/papers/features.pdf