An Empirical Modeling of Companies Using Support Vector Data Description

Mohammad Ebrahim Gorgani, Mahdi Moradi and Hadi Sadoghi Yazdi

***Abstract*—this paper introduces a new approach for modeling of company's behavior based on support vector data description (SVDD). Using SVDD behavior of all companies surrounded by hyper sphere in high dimensional space. We encountered appropriate observation from experimental results.**

***Index Terms*— Support vector data description, Behavioral modeling of companies.**

I. Introduction

Going concern is a fundamental concept in accounting. Decision making problems in the area of financial status evaluation have been considered very important. Making incorrect decisions in firms is very likely to cause financial crises and distress. Predicting going concern of factories and manufacturing companies is the desire of investors, auditors, financial analysts, governmental officials, employees and managers. The development of the going concern or bankruptcy prediction model has long been regarded as an important and widely studied issue in the academic and business community. Several classification techniques have been suggested to predict going concern or bankruptcy using financial ratios and data originating from these financial statements.

Beaver (1966) [3] first used univariate analysis to predict corporate bankruptcy and find out that different financial ratios have different discriminant ability. Altman (1968) [1] began to use multiple discriminant analysis (MDA) to identify companies into known categories and concluded that bankruptcy could be explained quite completely by a combination of five financial ratios, which constructed the famous Z-score model. Ohlson (1980) [11] paid attention on the limitation of linear prediction model and attempted to apply Logistic regression (Logit) model to predicting financial distress. Zmijewski, (1984) [17] presented a model by the probit .This model does give a crisp relationship between explanatory and response variables of the given data from a statistical viewpoint and do not assume multivariate normality.The probit model assumed that the cumulative probability distribution must be standardized normal distribution, while the logit model assumed that the cumulative probability distribution must be logistic distribution.

Remm (2004), Watson (1999) yielded a model by Case-based reasoning (CBR) which describes a methodology for problem solving. CBR is a link between statistical method with artificial intelligence ones. CBR involves case storage, retrieval, revise, reuse, and retain (Aamodt & Plaza, 1994; Finnie & Sun, 2003 [7] ; Pal & Shiu, 2004). Jo, Han, and Lee (1997) and Sun and Hui (2006) [12] respectively compared the financial distress prediction (FDP) accuracy based on case-based reasoning (CBR) with other models. The former considered that there was no significant difference between CBR and MDA, and the latter indicated that CBR was more suitable for short-term FDP. Li and Sun (2008) [14] constructed a new algorithm of CBR by employing ranking-order information and applied it into the area of FDP, with the conclusion that the new algorithm could achieve better performance than traditional algorithm. Li and Sun (2009) [9] attempts to investigate whether or not a Multi-CBR system can produce higher predictive performance than commonly used CBR models and statistical models with data from Chinese listed companies. They can draw the conclusion that Multi-CBR–MV is suitable for financial distress prediction of Chinese listed companies. It can achieve high accuracy and is stable at the same time, in the condition that there are no proper methods for data reduction.

Sun and Li (2008) [9] used weighted majority voting combination of multiple classifiers for FDP, and Cho et al. (2009) introduced an integration strategy with subject weight based on neural network for bankruptcy prediction. They all generated diverse classifiers by applying different learning algorithms (with heterogeneous model representations) to a single data set, and concluded that to some degree FDP based on ombination of multiple classifiers was superior to single classifiers according to accuracy rate or stability. The most used machine learning technique is the neural network model (Haykin, 1999), trained by the back-propagation learning algorithm (Wong et al., 1997; Wong & Selvi, 1998; Vellido et al., 1999; Huang et al., 2004), whose prediction accuracy outperforms statistical models including logistic regression (LR), linear discriminant analysis (LDA), multiple discriminant analysis (MDA) and other machine learning models, such as k-nearest neighbor (k-NN) and decision trees. In addition, the back-propagation neural network(BPN) model can be used as the benchmark for financial decision support models.

Chen and Du (2009) found that prediction performance for

Manuscript received May 29, 2010.

M. E. Gorgani is with the Ferdowsi University of Mashhad, Mashhad, Iran.

Mahdi Moradi is with the Ferdowsi University of Mashhad, Mashhad, Iran.

Hadi Sadoghi Yazdi is with the Ferdowsi University of Mashhad, Mashhad, Iran. (Corresponding author: +989151738312; fax:+985118763306; e-mail:h-sadoghi@um.ac.ir).

the clustering approach is more aggressively influenced than the BPN model and the BPN approach obtains a better prediction accuracy than the data mining (DM) clustering approach in developing a financial distress prediction model.

The purpose of this paper is to apply support vector machines (SVMs) in going concern prediction model. SVM was introduced from statistical learning theory by Vapnik. SVM lead to increase performance in pattern recognition, regression estimation, and so on; financial time-series forecasting (Kim, 2003 [8]; Mukherjee, Osuna, & Girosi, 1997; Tay & Cao, 2001),text categorization (Joachims, 2002), face detection using image (Osuna, Freund, & Girosi, 1997), hand written digit recognition (Burges & Scholkopf, 1997; Cortes & Vapnik, 1995). SVM is known as the algorithm that finds a special kind of linear model with the maximum margin hyperplane. The maximum margin hyperplane gives the maximum separation between decision classes. The training examples that are closest to the maximum margin hyperplane are called support vectors. Therefore, the aim of this paper is assess SVMs by using two different data sets. Here different training and testing proportions of each data set will be used to train and test SVMs. In addition, the SVM classifier will be trained by different kernel functions in order to compare it with the benchmark of the neural network model. In SVM, Using different kernel functions and the determination of optimal values of the parameters to train SVMs will lead to different results. These issues affect the ability to make the final conclusion reliable.

The paper is organized as follows. In the next section we review the support vector data description (SVDD). The proposed method is explained in Section 3 with some experiments. Final section includes the conclusion.

II. Support Vector Data Description

The support vector data description was presented by Tax and Duin [19] and again in Tax and Duin [18] with extensions and a more thorough treatment. The SVDD is a one-class classification method that estimates the distributional support of a dataset. A flexible closed boundary function is used to separate trustworthy data on the inside from outliers on the outside.

The basic idea of SVDD is to find a minimum hypersphere containing all the objective samples and none of the nonobjective samples. The hypersphere is specified by its center *a* and its radius *R*. The data description is achieved by minimizing the error function:

$$F(R,a) = R^2\,, \qquad (1)$$

$$s.t. \quad \|x_i - a\|^2 \le R^2,\ \forall i\,. \qquad (2)$$

In order to allow for outliers in the training dataset, the distance of each training sample x_i to the center of the sphere should not be strictly smaller than R^2. However, large distances should be penalized. Therefore, after introducing slack variables $\xi_i \ge 0$ the minimization problem becomes:

$$F(R,a) = R^2 + C\sum_i \xi_i\,, \qquad (3)$$

$$s.t. \quad \|x_i - a\|^2 \le R^2 + \xi_i\,, \qquad \forall i\,. \qquad (4)$$

The parameter C gives the tradeoff between the volume of the description and the errors. The constraints can be incorporated into the error function by introducing Lagrange multipliers and constructing the Lagrangian.

$$L(R, a, \alpha_i, \gamma_i, \xi_i) = R^2 + C\sum_i \xi_i - \sum_i \alpha_i \{R^2 + \xi_i - (\|x_i\|^2 - 2a.x_i + \|a\|^2)\} - \sum_i \gamma_i \xi_i\,. \qquad (5)$$

With the Lagrange multipliers $\alpha_i \ge 0$ and $\gamma_i \ge 0$. Setting partial derivatives to 0 gives these constraints:

$$\frac{\partial L}{\partial R} = 0 : \quad \sum_i \alpha_i = 1\,, \qquad (6)$$

$$\frac{\partial L}{\partial a} = 0 : \quad a = \frac{\sum_i \alpha_i x_i}{\sum_i \alpha_i} = \sum_i \alpha_i x_i\,, \qquad (7)$$

$$\frac{\partial L}{\partial \xi_i} = 0 : \quad C - \alpha_i - \gamma_i = 0\,. \qquad (8)$$

From the above equations and the fact that the Lagrange multipliers are not all negative, when we add the condition $0 < \alpha_i < C$, Lagrange multipliers γ_i can be safely removed. So the problem can be transformed into maximizing the following function L with respect to the Lagrange multipliers α_i:

$$L = \sum_i \alpha_i (x_i . x_i) - \sum_{i,j} \alpha_i \alpha_j (x_i . x_j)\,, \qquad (9)$$

$$s.t \qquad 0 < \alpha_i < C\,. \qquad (10)$$

Note that from Eq. (7), the center of the sphere is a linear combination of the training samples. Only those training samples x_i which satisfy Eq. (4) by equality are needed to generate the description since their coefficients are not zero. Therefore these samples are called Support Vectors. The radius can be computed using any of the support vectors:

$$R^2 = (x_k . x_k) - 2\sum_i \alpha_i (x_i . x_k) - \sum_{i,j} \alpha_i \alpha_j (x_i . x_j)\,. \qquad (11)$$

To judge a test sample *z* whether is in the target class, its distance to the center of sphere is computed and compared with R, if satisfies Eq. (12), it will be accepted, and otherwise, rejected.

$$\|z - a\|^2 = (z.z) - 2\sum_i \alpha_i (z.x_i) - \sum_{i,j} \alpha_i \alpha_j (x_i.x_j) \le R^2\,. \qquad (12)$$

SVDD is stated in terms of inner products. For more flexible boundaries, therefore, inner products of samples $(x_i.x_j)$ can be replaced by a kernel function $K(x_i,x_j)$, where $K(x_i,x_j)$ satisfies Mercer's theorem [8]. This implicitly, maps samples into a nonlinear space to obtain a more tight and nonlinear boundary. In this context, the SVDD problem of Eq. (9) can be expressed as:

$$L = \sum_i \alpha_i K(x_i . x_i) - \sum_{i,j} \alpha_i \alpha_j K(x_i . x_j)\,. \qquad (13)$$

Several kernel functions have been proposed for the SV classifier. Not all kernel functions are equally useful for the SVDD. It has been demonstrated that using the Gaussian kernel:

$$K(x, y) = \exp\left(-\frac{\|x-y\|}{S^2}\right)^2, \qquad (14)$$

results in tighter description. By changing the value of *S* in the Gaussian kernel, the description transforms from a solid hypersphere to a Parzen density estimator.

III. Data Collection And Preprocessing

The database used in this study was obtained Iranian Stock Exchange. Based on the background of Iranian listed companies, the criteria whether the listed company is specially treated (ST) by Iranian Stock Exchange is used to categorize financial state into two classes, i.e. normal and distressed. The most common reason that Iran listed companies are specially treated by Iranian Stock Exchange is that they have had accumulated loss to Stockholders' equity more than twice. ST companies are considered as companies in financial distress and those never specially treated are regarded as healthy ones. This experiment uses financial data two years before the company is specially treated, which is often denoted as year (t-2) in many literatures.

The data used in this research was obtained from Iran Stock Market and Accounting Research Database. According to the data between 2000 and 2009, 70 pairs of companies listed in Tehran Stock Exchange are selected as initial data set. The preprocessing operation to eliminate missing and outlier data is carried out: (1) Sample companies in case of missing at least one financial ratio data were eliminated. (2) Sample companies with financial ratios deviating from the mean value as much as three times of standard deviation are excluded. After eliminating companies with missing and outlier data, the final number of sample companies is 120.

IV. Feature Selection

This study uses more variables than other authors, who usually do not use more than 20. The ratios initially selected allow for a very comprehensive financial analysis of the firms including financial strength, liquidity, solvability, productivity of labour and capital, various kinds of margins and profitability and returns. Although, in the context of linear models, some of these variables have small discriminatory capabilities for default prediction, the non-linear approaches used here can extract relevant information contained in these ratios to improve the classification accuracy without compromising generalization. Feature selection is an important issue in bankruptcy prediction, as in other problems where a large set of attributes is available, since elimination of useless features may enhance the accuracy of detection while reducing the amount of time for processing the data. Due to the lack of an analytical model, the relative importance of the input variables can only be estimated through empirical methods. A complete analysis would require examination of all possibilities, for example, taking two variables at a time to analyse their dependence or correlation, then taking three at a time, etc. This, however, is both infeasible and not error free since the available data may be of poor quality in sampling the full input space. 20 financial ratios covering profitability, activity ability, debt ability and growth ability are selected as initial features.(see table1)

TABLE 1: Definition of Predictor Variables

Variable	Financial ratios Description
X1	Debt ratio
X2	Interest expenses to total expenses
X3	Sales to total assets
X4	Ordinary income to sales
X5	Net income to sales
X6	Operating income to sales
X7	Costs of sales to sales
X8	Net interest expenses to sales
X9	Ordinary income to total assets
X10	Rate of earnings on total capital
X11	Net working capital to total assets
X12	Current liabilities to total assets
X13	Stockholders' equity to total assets
X14	Total borrowings and bonds payable to total ssets
X15	Total assets turnover
X16	Ordinary income to total assets
X17	Net working capital to sales
X18	Stockholders' equity to sales
X19	Ordinary income to total assets
X20	Earnings before interest and taxes to sales

V. The Proposed System

The proposed system include following two stages:

1) Feature extraction from company's behavior
2) Modeling of behavior using SVDD

Feature extraction from company's behavior

Financial ratio, return on owners equity, return on total assets, earnings/sale ratio, sale/total assets ratio, current ratio, liability ratio, earning/total assets ratio, interest expense/current liability ratio, …. more than 20 features are applied for behavioral modeling.

Modeling of behavior using SVDD

First experiment using SVDD is performed by applying to features for each company. In this example as shown in Fig 1, SVDD describes all data in High Dimensional Space (HDS) by one hyper-sphere but in feature space all data surrounded as Fig 1.

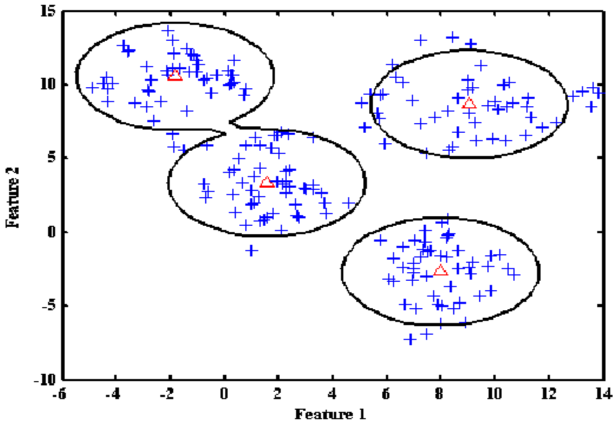

Fig.1: Description of Companies behavior using SVDD

One problem in companies' behavior is noisy samples captured by expert man. This work is performed by applying suitable C coefficient in (3). If C is selected large data as 100 or more all data are surrounded by hyper sphere. By decreasing C in (3) we found surface as Fig 2 which some noisy samples are interpreted as outliers.

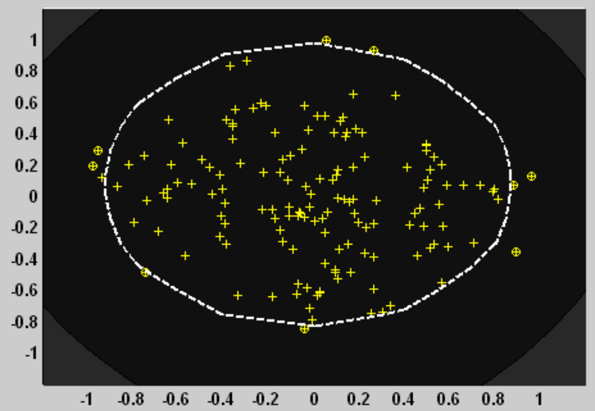

Fig.2: Outliers detection by selection of suitable C in (3)

SVDD can present flexible surface using training procedure that is presented in section II. Following figure (Fig 3) shows this ability of companies' samples.

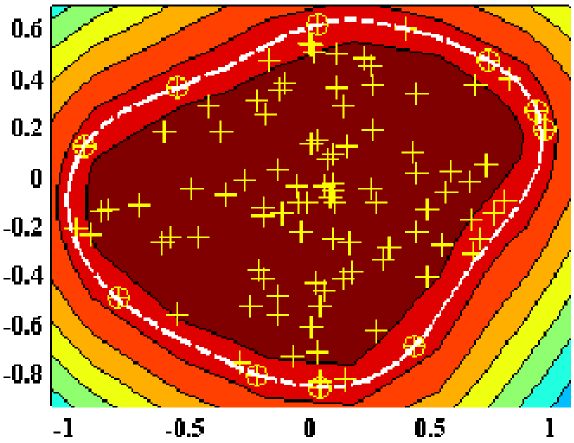

Fig.3: Perfect modeling of companies' behavior using SVDD: A flexible surface

Changing C in (3) is caused an obtained surface change which is shown in Fig 4.

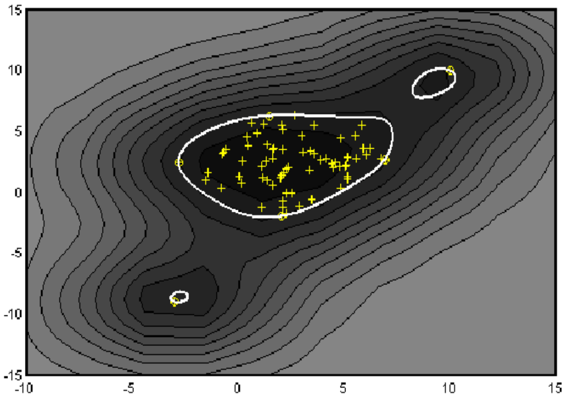

Fig.4: Changing surface by C coefficient

Obtained results from more than 120 Iranian companies shows suitable ability of SVDD in behavioral modeling of financial works for companies.

VI. Conclusion

For the first time, this paper paid to modeling of companies' behavior using SVDD. Flexibility of the proposed method with noise rejection capability was shown in obtained results.

References

[1] Altman, E.(1968).Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. Journal of Finance, 23(4), 589–609.

[2] Altman, E. (1984). A further empirical investigation of the bankruptcy cost question. Journal of Finance, 39(4), 1067–1089.

[3] Beaver, W. H. (1968b).Market prices, financial ratios, and the prediction of failure. Journal of Accounting Research (Spring): 179-192.

[4] Chen,W. and Du,Y.(2009).Using neural networks and data mining techniques for the financial distress prediction model. Expert Systems with Applications 36, 4075–4086

[5] Cortes, C., & Vapnik, V. N.(1995). Support vector networks. Machine Learning, 20, 273–297.

[6] Ding, Y.-S., Song, X.-P., & Zen, Y.-M. (2008). Forecasting financial condition of Chinese listed companies based on support vector machine.Expert Systems with Applications, 34(4), 3081–3089.

[7] Finnie,G. & Sun, Z. (2003).R5 model for case-based reasoning. Knowledge-Based Systems, 16(1), 59–65.

[8] Kim, K. J. (2003). Financial time series forecasting using support vector machines. Neurocomputing, 55(1/2), 307–319.

[9] Li,H and Sun,J . (2009).Majority voting combination of multiple case-based reasoning for financial distress prediction.Expert Systems with Applications 36, 4363–4373

[10] Li, H., Sun, J., & Sun, B.-L.(2007).Financial distress prediction based on OR-CBR in the principle of k-nearest neighbors.Expert Systems with Applications.

[11] Ohlson, J.(1980).Financial ratios and the probabilistic prediction of bankruptcy. Journal of Accounting Research, 18(1), 109–131.

[12] Sun, J., & Hui, X.(2006). Financial distress prediction based on similarity weighted voting CBR.Advanced Data Mining and Applications, 4093, 947–958.

[13] Sun, J., & Li, H.(2007). Listed companies' financial distress prediction based on weighted majority voting combination of multiple classifiers.Expert Systems with Applications. doi:10.1016/j.eswa.2007.07.045.

[14] Sun, J., & Li, H,(2008).Data mining method for listed companies' financial distress prediction. Knowledge-Based Systems, 21(1), 1–5.

[15] Tay, F. E. H., & Cao, L,(2001). Application of support vector machines in financial time series forecasting. Omega, 29, 309–317.

[16] Tsai,Ch.(2008) . Financial decision support using neural networks and support vector machines. Expert Systems, Vol. 25, No. 4

[17] Zmijewski, M. E.(1984).Methodological issues related to the estimated of financial distress prediction models. Journal of Accounting Research, 22(1), 59–82.

[18] D. M. J. Tax, R. P. W. Duin, "Support Vector Data Description". Kluwer Academic Publishers, Machine Learning 54, pp. 45-66, 2004.

[19] David M.J. Tax, Robert P.W. Duin, "Support vector domain description". Pattern Recognition Letters 20, pp. 1191-1199, 1999.

Hadi Sadoghi Yazdi received the B.S. degree in electrical engineering from Ferdowsi Mashad University of Iran in 1994, and then he received to the M.S. and Ph.D. degrees in electrical engineering from Tarbiat Modarres University of Iran, Tehran, in 1996 and 2005, respectively. He works in Department of Computer Engineering as Associate Professor at Ferdowsi University of Mashhad. His research interests include pattern recognition, and optimization in signal processing.