Selecting Bankruptcy Predictors Using A Support Vector Machine Approach

Alan Fan and Marimuthu Palaniswami
Department of EEE, University of Melbourne, VIC 3010, Australia
{apfan@ee.mu.oz.au, swami@ee.mu.oz.au}

Abstract

Conventional Neural Network approach has been found useful in predicting corporate distress from financial statements. In this paper, we have adopted a Support Vector Machine approach to the problem. A new way of selecting bankruptcy predictors is shown, using the Euclidean distance based criterion calculated within the SVM kernel. A comparative study is provided using three classical corporate distress models and an alternative model based on the SVM approach.

1 Introduction

This problem originated from the work of Beaver [4] when he proposed the univariate analysis on financial ratios to predict the bankruptcy of a company. It was followed by Altman's 1968 [1] seminal paper, when the Linear Discriminant Analysis (LDA) was first employed to solve the problem. In the early 1990s, the limitation of "traditional statistical tools" was challenged, as the neural network (NN) based techniques started to appear in the financial literature. These are applied in problems of predicting failure in banking industry [3] [6] , manufacturing and other sectors [14] [8]. It had also induced different methods of data preprocessing such as selecting the suitable bankruptcy predictors.

Recent NN researches has focus on alternatives techniques through the use of statistical learning theory. In particular, Support Vector Machine (SVM) is one of such methods that is receiving increasing attention in recent years [17]. The method implemented the principle of Structural Risk Minimization [17] by constructing an optimal separating hyperplane in the hidden feature space, using quadratic programming to find a unique solution.

We have applied the Support Vector Machine approach to corporate distress prediction using real-life data from Australian companies, firstly by examining the empirical results on different classical corporate distress models. Secondly we attempted to use a filter approach to input selection, where the Euclidean distance criterion is used to choose a subset of input variables to favor the svm kernel used, for the purpose of finding suitable financial indicators.

2 The Problem and the Data set

The bankruptcy prediction was usually formulated as a two-class pattern recognition problem. Assuming that there are N financial ratios in our classifier, the predictor variables of the i^{th} firm can be represented by the vector $\mathbf{x}_i = (x_1, x_2, ..., x_N)$. For simplicity the financial status of the firm will be a binary dependent variable $y_i = \pm 1$, where $+1$ represents healthy and -1 for firms in distress. Therefore a training set of l firms will be the following x/y pairs:

$$\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_l, y_l)\} \subset \mathbf{R}^N \times \pm 1 \tag{1}$$

The learning machine which attempts to learn from the examples can then be regarded as a set of functions mapping the predictor variables to values of y, which ultimately aims to reduce the risk of misclassification by adjusting its parameters. For gradient methods like back-propagation, this is achieved by reducing the empirical risk, whereas SVM aims to reduce the bound on the actual risk instead.

Our data set consists of financial indicators (table 2) obtained from 174 small/medium Australian industrial firms, including 86 firms in financial distress, i.e. bankruptcy or default loan. In order to access

0-7695-0619-4/00 $10.00 © 2000 IEEE

the predictive performance, financial data of distressed firms were obtained at least 1 year before bankruptcy was recorded. The data set used in this paper exhibits the characteristics of most real-life financial problems, of being small in sample size, and noisy in nature. One source of the noise maybe introduced by the various creative accounting practice, another being the errors and missing fields in data. In order to have consistency of data between different classifiers, missing data was calculated by mean imputation (i.e. replacing missing ratios with their mean), and the data was standardized as a whole.

3 Overview of SVM

SVM addresses the problem that minimization of empirical error does not guarantee a small actual error. Therefore rather than reducing the empirical error (Empirical Risk Minimization), it implements the principle of Structural Risk Minimization [17], which aims to reduce the bound of the misclassification risk. For linearly separable patterns, this is achieved by using optimal separating hyperplane which takes the form of eqt.(2) with weights $\mathbf{w}$ and bias b:

$$f(\mathbf{x}_i) = sgn((\mathbf{w}^{'}\mathbf{x}_i) + b) = y_i, \qquad i = 1, ..., l \tag{2}$$

Note that this form is equivalent to the linear discriminant function with a bias and can be re-written as

$$y_i((\mathbf{w}^{'}\mathbf{x}_i) + b) \geq 1, \qquad i = 1, ..., l \tag{3}$$

For the hyperplane f to be optimal, its margin of separation,

$$\min_{\{\mathbf{x}_i|y_i=+1\}} \frac{\mathbf{w}^{'}\mathbf{x}_i + b}{|\mathbf{w}|} - \max_{\{\mathbf{x}_j|y_j=-1\}} \frac{\mathbf{w}^{'}\mathbf{x}_j + b}{|\mathbf{w}|} = \frac{2}{\sqrt{\mathbf{w}^{'}\mathbf{w}}} \tag{4}$$

must be maximized subjected to eqt.(3). In fact this can be transformed into a Lagrangian, and there exist a dual problem which is equivalent to maximizing:

$$Q(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i^{'}\mathbf{x}_j) \quad \text{subject to} \quad \alpha_i \geq 0, \sum_{i=1}^{l} \alpha_i y_i = 0$$

For detailed derivation see [15] or [7]. The solution to this problem will be given by quadratic programming, in the form of $\alpha = \{\alpha_1, ..., \alpha_l\}$ where each $\alpha_i > 0$ corresponds to a support vector, and then the weights $\mathbf{w}$ and bias b can be calculated. To allow for non-separable patterns, it is equivalent to impose extra constraints on eqt.(5) as $\alpha_i \leq C$. This regularization parameter C was usually selected manually to control the trade-off between complexity (hence capacity) and number of non-separable points (hence training error).

However, SVM takes one more step to map the input vectors to a hidden, high dimensional feature space before constructing the optimal hyperplane. This can be done with minimum extra computational cost with the appropiate mappings. Consider the mapping $K : \mathbf{x}_i \rightarrow \mathbf{z}_i$ with the dot product of the transformation as k:

$$\left(K(\mathbf{x})^{'}K(\mathbf{x}_i)\right) = k(\mathbf{x}, \mathbf{x}_i) \tag{5}$$

Then the quadratic programming problem of minimizing Q of eqt.(5) can be re-written as

$$\begin{aligned} Q(\alpha) \quad &= \quad \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \\ \text{subject to} \quad & \quad 0 \leq \alpha_i \leq C, \\ & \quad \sum_{i=1}^{l} \alpha_i y_i = 0 \end{aligned} \tag{6}$$

with decision function

$$f(\mathbf{x}) = sgn\left(\sum_{i=1}^{l} y_i \alpha_i k(\mathbf{x}, \mathbf{x_i}) + b\right) \tag{7}$$

Notice that the whole problem can still be solved in the same way without knowing the exact mapping K, but using the knowledge of dot product $k(\mathbf{x}, \mathbf{x}_i)$ instead. With different kernels, the SVM architexture has the form of different classical classifiers. In this study we used the Radial Basis kernel where $k(\mathbf{x}, \mathbf{y}) = exp(-\|\mathbf{x} - \mathbf{y}\|^2)$. Once the appropiate kernel is used, only the upper bound on the alphas (C) was left to be assigned manually. The parameter selection problem in SVM then became a tradeoff between capacity and generalization. For a noisy problem like the one we have, we expected a small C value to allow for more errors within the training sample.

4 Choosing variables according to the SVM Kernel

One challenging task for predicting bankruptcy is the selection of input variables. For this problem, the number of financial indicators (usually financial ratios) is usually large relative to the total sample size. Although more variables may contain more information, use of too many variables can degrade the generalization performance of the classifier. In classical statistics this was known as a peaking phenomenon whereby training a samples of finite sizes, the performance of a given discriminant rule does not increase monotonically with the number of inputs [12]. Instead the overall true error rate will stop decreasing and start to increase after certain threhold of the number of variables. As in most previous study on bankruptcy, we used a subset of all the variables avaliable.

Furthermore, in a recent work by Boritz and Kennedy [5] , network performance was found to vary greatly with the choice of variables used. Therefore we find it necessary to select an appropiate combination of ratios for the classifer. Usually for blackbox models such as multi-layer preceptron (MLP) this was achieved by using the same variables used in the traditional statistical classifier. If we would like to choose a better subset then we have to use pruning or a wrapper approach based on the network error [9]. However pruning was not applicable in this case because the network performance decreased greatly when all ratios were used as input. A wrapper approach maybe suitable for MLP, but not feasible for SVM due to the high computational cost involved.

Theoretically SVM is a slightly better 'blackbox' than existing neural techniques because of its formulation of the problem in kernels and dot products. Although the underlying feature space is hidden (or maybe even infinite dimensional), we still know the exact kernel which is the dot product of the transformation. Moreover, the discriminat rule within the feature space is linear in nature. We utilized this property and attempted to select input variables for SVM via a filter approach [9]. This was achieved by using the so-called 'kernel trick' as in eqt. (5), to work within the hidden feature space and derive some sort of criterion for class separability.

A straight forward method to select suitable input is to have the subset of inputs that maximize the distance of vectors between different classes, and minimize the distance within same class. For the case of Euclidean distance, it can be the minimization of:

$$J = \sum_{c}^{2} \frac{\sum_i^{l_c} \sum_j^{l_c} \|\mathbf{z}_{ci} - \mathbf{z}_{cj}\|^2}{{l_c}^2} - \frac{\sum_i^{l_1} \sum_j^{l_2} \|\mathbf{z}_{1i} - \mathbf{z}_{2j}\|^2}{l_1 l_2} \tag{8}$$

where $\mathbf{z}_{ci}$ is the ith pattern (in the feature space) of class c, and l_c is the number of patterns in class c. Since the square of the Euclidean distance is simply the dot product, we can take the SVM approach to evaluate J using only the kernel. Using eqt.(5) and by $\mathbf{z} = K(\mathbf{x})$, we get:

$$\|\mathbf{z}_i - \mathbf{z}_j\|^2 = (\mathbf{z}_i - \mathbf{z}_j)'(\mathbf{z}_i - \mathbf{z}_j) = k(\mathbf{x}_i, \mathbf{x}_i) - 2k(\mathbf{x}_i, \mathbf{x}_j) + k(\mathbf{x}_j, \mathbf{x}_j) \tag{9}$$

And we can substitute this into eqt. (8) and then J can be evaluated as:

$$J = \sum_{c}^{2} \frac{\sum_i^{l_c} K_{cc}(i,i)}{l_c} - 2\left(\frac{\sum_i \sum_j K_{11}(i,j)}{l_1^2} - \frac{\sum_i \sum_j K_{12}(i,j)}{l_1 l_2} + \frac{\sum_i \sum_j K_{22}(i,j)}{l_2^2}\right) \tag{10}$$

where K_{cd} is the kernel dot product matrix with $K_{cd}(i,j) = \mathbf{z}_{ci}'\mathbf{z}_{dj} = k(\mathbf{x}_{ci}, \mathbf{x}_{dj})$.

This criterion maybe sub-optimal as it does not guarantee an optimal SVM classification especially when capactiy control is used. However, variables selected according to this criterion by definition tend to separate from each other in the hidden feature space, and this method has a lower computational cost compared to a wrapper SVM approach.

5 Experimental Results

Firstly we compared the predictive performance of SVM approach to some traditional approaches including the Linear Discriminant Classifier (LDA) [1], Multi-layer Preception (MLP), and the Learning Vector Quantization (LVQ) [10]. For the adaptive neural techniques heuristic such as re-initialization and early stopping was used. Since we were more interested to compare the generalization ability of each classifier, we used a 20-fold cross validation to estimate the true test accuracy as it averaged over the 20 runs. During each trial an extra 10% of the training set was used as a validation set to estimate parameters such as the number of hidden units in MLP and LVQ, and the regularization parameter C in SVM.

The comparison was performed using three different sets of financial ratios used in models such as Altman's [2] revised Z-score model for private firms, the Australian Lincoln's [11] model (also known as L-score), and the model by Ohlson [13]. For the listing of variables please refer to table 3. The following table shows the average training and testing prediction results for each set.

	Altman		Lincoln		Ohlson	
	Training	*Testing*	*Training*	*Testing*	*Training*	*Testing*
LDA	65.76	64.31	68.78	62.15	70.87	64.72
MLP	68.07	62.85	81.59	64.90	72.23	68.61
LVQ	69.59	62.71	72.55	66.25	72.55	66.25
SVM	74.05	65.14	87.24	67.22	81.15	69.17

Table 1: The results using classical models (percentage classified correctly)

It can be seen that our results were somewhat empirically consistent with most previous research in the area. In the 3 sets of financial ratios considered, neural models such as MLP and LVQ outperformed LDA with the exception on the Altman's ratio set. Furthermore, SVM outperforms all 3 other techniques in each set of ratios. Note that results significantly differed with the set of inputs used.

In the next step we attempted to select a subset of financial indicators favourable for SVM, using the techniques described in section 4. We combined all three sets of inputs mentioned above with additional 16 manually selected ratios based on availability. On this pool of variables we performed a sequential search, aiming to minimize J of eqt.(10) using the 'plus 2 take away 1' method [16]. The following graph shows the trend of the average testing accuracy as variables were added into the input subset. The best number of variables was 11 ratios (see table 2), and its results were reported in table 2.

	Training	*Testing*
LDA	63.85	61.39
MLP	73.91	66.11
LVQ	69.76	62.71
SVM	81.73	70.97

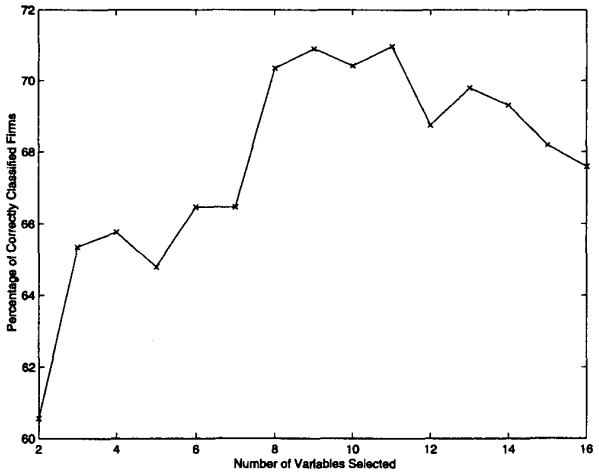

Figure 1: Test accuracy vs number of ratios selected (with results of 11 selected ratios on the left)

Recall that there were limitations on the criterion. Also sequential search does not necessary give the optimal results. Nevertheless the selected inputs showed improvements on SVM performance over the

Altman	Working Capital / Total Assets	Retained Earnings / Total Assets
	Earning before Income Tax / Total Assets	N.W. (book value) / Total Liabilities
	Sales / Total Assets	
Lincoln	Cash Flow before Tax / Current Liabilities	Current Assets / Total Assets
	Current Liabilities / Total Liabilities	Quick Liabilities / Current Liabilities
	Retained Profits / Total Assets	Total Liabilities / Total Assets
Ohlson	log(Total Assets / GDP index)	Total Liabilities / Total Assets
	Working Capital / Total Assets	**Current Liabilities / Current Assets**
	Total Liabilities > Total Assets	**Net Income / Total Assets**
	Funds by Operations / Total Liabilities	**Negative Income for last 2 years**
	Change in Net Income level	
Additional Ratios	Income Before Tax / Total Asset	**Income Before Tax / Owner's Equity**
	Income Before Tax / Sales	Owner's Equity / Total Asset
	Fixed Asset / Owner's Equity	Fixed Asset / Equity and LT Liability
	Total Liability / Owner's Equity	Current Liability / Owner's Equity
	Fixed Liability / Owner's Equity	Retained Earning / Total Asset
	Total Asset Turnover	Owner's Equity Turnover
	Income Before Tax / Interest Expenses	Cash After Operation /
	Stock Days Outstanding	Interest Expenses + Long Term Debt

Table 2: Variables employed in this study (Highlighting first 11 selected variables)

original sets of variables. It should be noted that applying this set of inputs to other classifiers did not increase their performance, as our criterion was kernel specific.

6 Conclusions

We examined the practicality and performance of the Support Vector Machine approach to predict Australian business failure. Empirical results showed that SVM was competitive and outperformed other classifiers in terms of generalization performance. Further improvements of the SVM results was achieved by selecting an input subset suitable for the kernel. We also proposed an Euclidean distance based input selection criterion, which can provide a selection of variables that tends to discriminate within the SVM kernel used.

References

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

[2] E. Altman. *Corporate Financial Distress and Bankruptcy : a complete guide to predicting & avoiding distress and profiting from bankruptcy.* John Wiley & Sons, New York, 1993.

[3] E. Altman, G. Marco, and F. Varetto. Corporate distres diagnosis: Comparisons using linear discriminant analysis and neural networks (the italian experience). *Journal of Banking and Finance*, 18:505–529, 1994.

[4] W. Beaver. Financial ratios as predictors of failure. *Empirical Research in Accounting: Selected Studies*, 1966.

[5] J. Boritz and D. Kennedy. Effectiveness of neural network types for prediction of business failure. *Expert Systems With Applications*, 9(4):503–512, 1995.

[6] H. Etheridge and R. Sriram. A neural network approach to financial distress analysis. *Advances in Accounting Information Systems*, 4:201–222, 1996.

[7] S. Haykin. *Neural Networks : a comprehensive foundation.* Prentice Hall, N.J., 1999.

[8] G. Klersey and M. Dugan. Substantial doubt: Using artifical neural networks to evaluate going concern. *Advances in Accounting Information Systems*, 3:137–159, 1995.

[9] R. Kohavi. *Wrappers for Performance Enhancement and Oblivious Decision Graphs.* PhD thesis, Computer Science Department, Standford University, 1995.

[10] T. Kohonen. *Self-Organization and Associative Memory, 2nd edition.* Springer-Verlag, Berlin, 1987.

[11] M. Lincoln. *An empirical study of the usefulness of accounting ratios to describe levels of insolvency risk.* PhD thesis, University of Melbourne, 1982.

[12] G. McLachlan. *Discriminant Analysis and Statistical Pattern Recogntion.* John Wiley & Sons, 1992.

[13] J. Ohlson. Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research,* 18(1), 1980.

[14] P. Pompe and A. Feelders. Using machine learning, neural networks and ststistics to predict corporate bankruptcy: A comparative study. In *Proceedings on the 4th International Workshop on Machine Learning,* 1996.

[15] B. Schlkopf, C. Burges, and V. Vapnik. Extracting support data for a given task. In *Proceedings of First International Conference on Knowledge Discovery & Data Mining,* 1995.

[16] S. Stearns. On selecting features for pattern classifiers. *Third International Conference on Pattern Recognition,* pages 71–75, 1976.

[17] V. Vapnik. *The Nature of Statistical Learning Theory.* Springer-Verlag New York, Inc., 1995.