2009 Ninth International Conference on Intelligent Systems Design and Applications

Support Vector Machines for Insolvency Prediction of Irish Companies

Anatoli Nachev
Cairnes School of Business & Economics
NUI, Galway
Galway, Ireland
anatoli.nachev@nuigalway.ie

Abstract— **This study explores experimentally the potential of linear and non-linear support vector machines with three kernels to predict insolvency of Irish firms. The dataset used contains selected financial features based on information collected from 88 companies for a period of six years. Experiments show that non-linear support vector machines (SVM) with polynomial kernel gives highest prediction accuracy and outperforms all other techniques used so far with the same dataset. SVM performance is estimated by various metrics, receiver operating characteristics analysis, and results are validated by the leave-one-out cross-validation technique.**

Keywords-supprt vector machines; data mining; insolvency prediction

I. INTRODUCTION

Financial analysis has developed a large number of techniques aimed at helping potential investors and decision makers. To estimate credit risk, investors usually apply scoring systems, which takes into account factors, such as leverage, earnings, reputation, etc. Due to lack of metrics and subjectiveness in estimates, sometimes decisions are unrealistic and not consistent.

Generally, a prediction of firm insolvency can be viewed as a pattern recognition problem, and as such, it can be solved by using one of two approaches: structural, and empirical. The former derives the probability of a company for default, based on its characteristics and dynamics, while the later approach relies on previous knowledge and relationships in that area, learning from existing data or experience, and deploys the statistical or other methods to predict failure.

Empirical techniques used for insolvency prediction can be considered as statistical and intelligent [6]. Statistical techniques include linear discriminant analysis (LDA), multivariate discriminate analysis (MDA), quadratic discriminant analysis (QDA), logistic regression (logit), factor analysis (FA), and some modern, such as support vector machines. Intelligent techniques include different neural network (NN) architectures, such as single-layer perceptron (SLP), multi-layer perceptron (MLP), probabilistic neural networks (PNN), auto-associative neural network (AANN), self-organizing map (SOM), ARTMAP neural networks, learning vector quantization (LVQ), cascade correlation neural network (Cascor), decision trees, case-based reasoning, evolutionary approaches, rough sets, soft computing (hybrid intelligent systems), operational research techniques including linear programming (LP), data envelopment analysis (DEA) and quadratic programming (QP), etc.

LDA, MDA and logistic regression have been the most commonly used statistical models in this type of work. These techniques, however, have been sharply criticized because of assumptions about the linear separability, multivariate normality, and the independence of the predictive variables, as these constraints are incompatible with the complex nature, boundaries, and interrelationships of most of financial ratios [7]. The intelligent techniques have shown themselves to be more appropriate for that task as they do not rely on a-priori assumptions about the distribution of data [3].

The backpropagation NN is one of the most well known and widely used models of supervised NNs, but this approach is totally empirical as no complete theoretical explanation exists to obtain the optimal architecture.

The objective of this study is to explore the potential of both linear and non-linear SMV with various types of kernels to provide insolvency warning signals. We used data collected from 88 Irish companies which allows us to compare results with other studies and different prediction techniques [5], [8], [9].

The paper is organized as follows: support vector machines are presented briefly in section 2; section 3 discusses the dataset and data pre-processing; experiments and analysis are presented in section 4, followed by section conclusions.

II. SUPPORT VECTOR MACHINES

SVM, originally introduced by Vapnik in 1990s [11], provide a new approach to the problem of pattern recognition with clear connections to the underlying statistical learning theory. They differ radically from comparable approaches such as NN because SVM training always finds a global minimum in contrast to NN [12].

SVMs are supervised learning methods used for classification and regression. Training data is a set of points of the form

$$D = \{(x_i, c_i) \mid x_i \in \Re^p,\ c_i \in \{-1,1\}\}_{i=1}^{n}, \qquad (1)$$

where the c_i is either 1 or −1, indicating the class to which the point x_i belongs. Each data point x_i is a

978-0-7695-3872-3/09 $26.00 © 2009 IEEE
DOI 10.1109/ISDA.2009.54

IEEE computer society

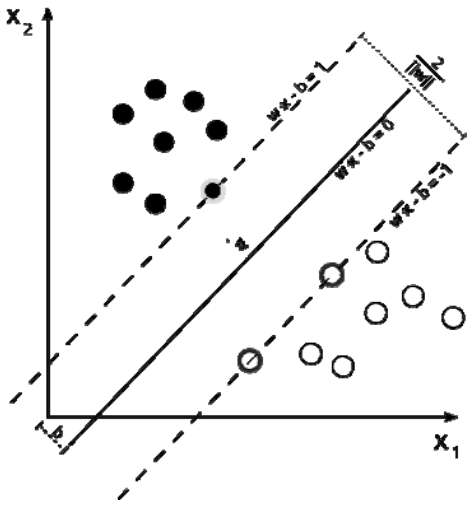

Figure 1. Maximum–margin hyperplane for a SVM trained with samples from two classes. Samples on the margin are support vectors.

real vector. During training a linear SVM constructs a *p-1*-dimensional hyperplane that separates the points into two classes (Fig. 1). Any hyperplane can be represented by $x \cdot w-b$, where w is a normal vector and $\cdot$ denotes dot product. Among all possible hyperplanes that might classify the data, SVM selects one with maximal distance (margin) to the nearest data points (support vectors).

When the classes are not linearly separable (there is no hyperplane that can split the two classes), a variant of SVM, called soft-margin SVM, chooses a hyperplane that splits the points as cleanly as possible, while still maximizing the distance to the nearest cleanly split examples. The method introduces slack variables, ξ_i , which measure the degree of misclassification of the datum x_i. Soft-margin SVM penalizes misclassification errors and employs a parameter (the soft-margin constant C) to control the cost of misclassification. Training a linear SVM classier solves the constrained optimization problem (2).

$$\min_{w,b,\xi_k} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \qquad s.t. \quad w \cdot x_i + b \geq 1 - \xi_i \tag{2}$$

In dual form the optimization problem can be represented by (3).

$$\min_{\alpha_i} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i \cdot x_j - \sum_{i=1}^{n}\alpha_i \qquad s.t. \quad 0 \leq \alpha_i \leq C, \quad \sum_{i=1}^{n}\alpha_i c_i = 0 \tag{3}$$

The resulting decision function $f(x)=w \cdot x+b$ has weight vector (4).

$$w = \sum_{k=1}^{n}\alpha_k y_k x_k \tag{4}$$

Data points x_i for which $\alpha_I>0$ are called support vectors, since they uniquely define the maximum margin hyperplane. Maximizing the margin allows one to minimize bounds on generalization error.

If every dot product is replaced by a non-linear kernel function, it transforms the feature space into higher-dimensional, thus though the classifier is a hyperplane in the high-dimensional feature space it may be non-linear in the original input space. The resulting classifier fits the maximum-margin hyperplane in the transformed feature space. Some common kernels include:

- Polynomial kernel: $k(x,x')=(sx \cdot x'+c)^d$
- RBF kernel: $k(x,x')=exp(-\gamma(x-x')^2)$
- Sigmoid kernel: $k(x,x')=tanh(s(x \cdot x')+c)$

An non-linear SVM is largely characterized by the choice of its kernel, and SVMs thus link the problems they are designed for with a large body of existing work on kernel based methods. Once the kernel is fixed, SVM classifiers have few user-chosen parameters. The best choice of kernel for a given problem is still a research issue. Because the size of the margin does not depend on the data dimension, SVM are robust with respect to data with high input dimension. However, SVM are sensitive to the presence of outliers, due to the regularization term for penalizing misclassification (which depends on the choice of C). The SVM algorithm requires $O(n^2)$ storage and $O(n^3)$ to learn.

III. DATASET

The dataset contains financial information for a period of six years for a total of 88 Irish firms, of which 44 are insolvent and 44 are solvent. The dataset consists of Altman's [1] financial ratios (features) as they have been the most widely and consistently used to date by both researchers and practitioners. The ratios are:

- R1: Working Capital / Total Assets;
- R2: Retained Earnings / Total Assets;
- R3: Earnings Before Interest and Taxes (EBIT) / Total Assets;
- R4: Market Value of Equity / Book Value of Total Debt;
- R5: Sales / Total Assets.

The working capital is current assets minus the current liabilities, which is an indication of the ability of the firm to pay its short term obligations. A firm's total assets are sum of the firm's total liabilities and shareholder equity. It can be viewed as an indicator of its size and therefore can be used as a normalizing factor. The retained earnings is the surplus of income compared to expenses, or total of accumulated profits since the firm commencement. The firm's earnings before interests and taxes is also an important indicator. Low or negative earnings indicate that the firm is losing its competitiveness, and that endanger its survival. Market capitalization relative to the total debt indicates that a firm is able to issue and sell new shares in order to meet its liabilities. Total sales of a firm, relative to the total assets, is

an indicator of the health of its business, but without certainty as it can vary a lot from industry to industry.

This study uses the Altman's ratios with two changes necessitated: operating profit was used instead of profit before interest and tax and so may contain a negligible amount of interest receivable; total shareholder funds was used as a proxy for market value of equity because not all of the companies used were quoted.

Two feature sets were used for experiments: one with all Altman's ratios and one with a reduced number of variables. Reduction of variables has a potential to improve the abilities of a classifier to alleviate the effect of the curse of dimensionality problem that appears with small datasets. This is because a classifier with fewer inputs has fewer adaptive parameters to be determined, and these are more likely to be properly constrained by a data set of limited size, leading to a classifier with better generalization properties. In addition to that, a classifier with fewer weights may be faster to train. A variable selection also helps to avoid the overfitting phenomenon that makes a classifier to adjust to very specific random features of the training data that have no causal relation to the target function and makes the classifier to lose its ability to generalize.

An F-ratio analysis shows that variables can be scored by their discriminatory power and that the set of variables [R1, R2, R3, R4] is a good selection. The same selection was used by Serrano [10] with an American dataset and Jones [5] with the dataset we use in this study. This also allows our results to be compared with those from the other studies.

The dataset was also preprocessed by transformation (5) and (6) that eliminates the effect of unbalanced variable values.

$$\tilde{x}_i^n = \frac{x_i^n - \overline{x}_i}{\sigma_i} \tag{5}$$

where

$$\overline{x}_i = \frac{1}{N}\sum_{n=1}^{N} x_i^n \ , \ \sigma_i^2 = \frac{1}{N-1}\sum_{n=1}^{N}(x_i^n - \overline{x}_i)^2 \tag{6}$$

This is necessary because different variables have values which differ significantly because of different units of measurements, which can reduce the predictive abilities of the model as some of the variables can dominate over others. Each of the input variables xi was treated independently. Transformed variables have zero mean and unit standard deviation over the transformed training set.

IV. Experiments and Discussion

A series of experiments with SVM was conducted in order to optimize the SVM parameters for the classification task. Linear and non-linear models were tested where options for non-linear were polynomial, RBF and sigmoid kernels. Results show that using all Altman's ratios [R1, R2, R3, R4, R5] best performer is SVM with polynomial kernel (7)

$$k(x, x') = (8.317 x \cdot x' + 4.296)^3 \tag{7}$$

Optimization of SVM parameters with reduced set of features [R1, R2, R3, R4], as discussed above, showed that best performance can be obtained by polynomial kernel (8)

$$k(x, x') = (1.435 x \cdot x' + 1.522)^3 \tag{8}$$

SVM output the distance of each test datapoint from the hyperplane, but in contrast to soft classifiers, such as back-propagation NN, it does not require mapping of real-valued ranks into crisp true/false values by threshold function or other techniques. The SVM algorithm optimizes the hyperplane location in order to minimize the misclassification error and the sign +/- of the distance can be considered as class labels true/false. Classification outcomes were counted in terms of true positive (TP) or positive hits; true negative (TN) or correct rejections; false positive (FP) or type I error; and false negative (FN) or type II error. Fig. 2 and Fig. 3 show the classification results by distance from the hyperplane.

A. Preformance Metrics

Most common metric to estimate a classifier performance is accuracy (ACC). It represents the total number of correctly classified instances divided by the total number of all available instances.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

The factor C in optimization problem (2) is a parameter that controls the trade-off between training error and model complexity. The lower the value of C, the more training error is tolerated. The best value of C depends on the data and must be determined experimentally.

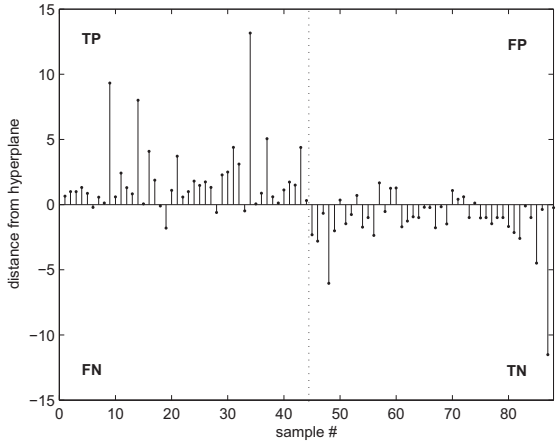

Figure 2. Classification of feature set [R1, R2, R3, R4, R5] by SVM with kernel (7).

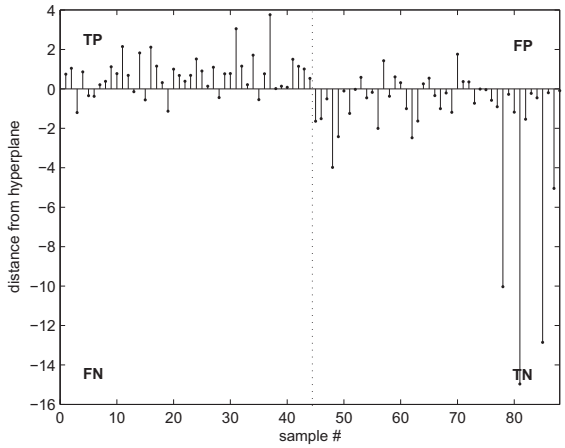

Figure 3. Classification of feature set [R1, R2, R3, R4] by SVM with kernel (8).

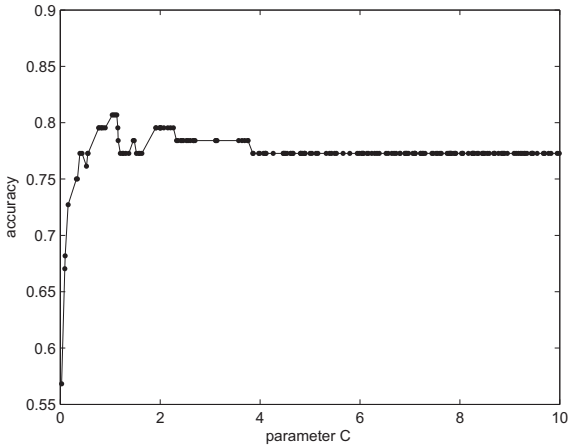

Figure 4. Prediction accuracy of SVM with kernel (7) and feature set [R1, R2, R3, R4].

A series of experiments sought the optimal values of C in terms of accuracy for both full and reduced feature set and results are shown in Figures 4 and 5. Optimal values for C are C=19.127 (ACC=84.09%) for the full feature set and C=1.067 (ACC=80.68%) for the reduced set.

ACC is most often used in applications, however, it can be misleading estimator if class distribution is skewed or if errors of type I and type II can produce different consequences and have different cost. The full picture of performance estimation have to include metrics such as sensitivity or true positive rate (TPR) (10), specificity or true negative rate (TNR) (11), and fall-out or false positive rate (FPR) (12).

$$TPR = TP / (TP + FN) \tag{10}$$

$$TNR = TN / (FP + TN) \tag{11}$$

$$FPR = FP / (FP + TN) \tag{12}$$

Table 1 shows those values obtained with optimal parameter values discussed above.

The two SVMs (with full and reduced feature set) can also be estimated by ROC (Receiver Operating Characteristics) analysis. A ROC space is defined by FPR and TPR as x and y axes respectively, which depicts relative trade-off between true positive (benefits) and false positive (costs). The perfect classification would yield a point in the upper left corner or coordinate (0,1), representing 100% sensitivity (all true positives are found) and 100% specificity (no false positives are found).

A completely random guess would give a point along the no-discrimination line from the left bottom to the top right corner. As the two SVMs are discrete classifiers, each plots a single point in the ROC space with coordinates (0.1861, 0.8) and (0.125, 0.8125) respectively.

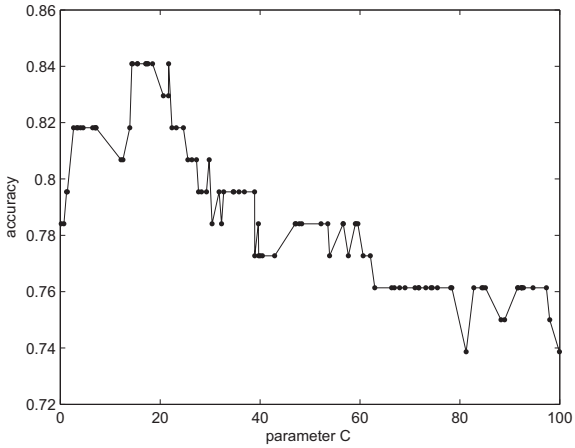

Figure 5. Prediction accuracy of SVM with kernel (6) and feature set [R1, R2, R3, R4, R5].

It can be noticed that the later point is more 'northwest' as it is more distant from the no-discrimination line, therefore the SVM with full feature set is better in the terms of ROC analysis. These results suggest that reduction of the Altman's variables worsen performance of SVM with this dataset.

Finally, Table 2 compares SVM with other techniques for insolvency prediction experimented with the same dataset [10], [11]. Non-linear SVM shows best prediction accuracy of 84% followed by ARTMAP NNs, SOFM, and backpropagation NNs. Other advantages of using SVM are that the local minima problem of NNs is not an issue; SVM requires fewer parameters to tune in contrast to NNs; finding appropriate NN architecture experimentally is not a problem with SVM.

B. Validation

The most popular method for estimating the generalization error of a classification rule is cross-validation [4].

TABLE I. SVM PERFORMANCE: SENSITIVITY (TPR), SPECIFICITY (TNR), FALL-OUT (FPR), AND ACCURACY (ACC).

Feature set	TPR	TNR	FPR	ACC
[R1,R2,R3,R4]	80%	81.39%	18.61%	80.68%
[R1,R2,R3,R4,R5]	81.25%	87.5%	12.5%	84.09%

TABLE II. INSOLVENCY PREDICTION ACCURACY OF TECHNIQUES APPLIED TO THE SAME DATASET.

Method	ACC
Backpropagation neural networks	72.7%
SOFM (Kohonen neural networks)	77.27%
Fuzzy ARTMAP neural networks	79.5%
Default ARTMAP neural networks	83.91%
Support vector machines	84.09%

While there are several versions of cross-validation estimator, most theoretical results concern leave-one-out cross-validation (LOOCV) estimator. Instead of dividing the dataset into training and testing datasets, from the training sample (1) the first example (x_1,c_1) is removed and the resulting sample $D^{\setminus 1}$ is used for training, leading to a classification rule $h_l^{\setminus 1}$. The classification rule is tested on the held-out example (x_1,c_1). This process is repeated for all training examples. The number of misclassifications divided by n is the LOOCV estimate.

$$Err_{LOOCV}^{n}(h_L) = \frac{1}{n}\sum_{i=1}^{n} L_{0/1}(h_L^{\setminus i}(x_i), c_i) \qquad (13)$$

All experimental results discussed in this study were carried out according to the LOOCV procedure. This technique was also used in other studies, such as [5] and [2]. LOOCV is suitable for small size datasets as it allows the greatest possible amount of data to be used for training. It is also a deterministic technique as no random sampling is involved, in contrast of a *k*-fold cross-validation ($1<k<n$).

V. CONCLUSIONS

Support vector machines provide an approach to the problem of pattern recognition, related to the statistical learning theory. This study explores experimentally the potential of classifiers based on support vector machines to predict insolvency of Irish firms. The dataset used contains selected financial features based on information collected from 88 companies for a period of six years and represented as Altman's ratios. Experiments show that non-linear SVM give a better accuracy than linear ones and that polynomial kernel outperform RBF and sigmoid kernel. Results also show that SVM with polynomial kernel outperform backpropagation NN, Fuzzy and Default ARTMAP NN, and Kohonen NN, all tested with the same dataset. The experiments also showed that reductions of Altman's ratios does not enhance prediction abilities of SVM (in contrast to backpropagation NN) which is an indication that in this classification task SVM are more resistant to curse of dimensionality and overfitting problems. The SVM performance was estimated by accuracy, sensitivity, specificity, and fall-out, and analyzed by ROC. Results were validated by LOOCV technique.

Using SVM for classification gives other advantages over NNs, such as: training always finds a global minimum; SVM requires fewer parameters to tune; finding appropriate architecture experimentally is not issue.

REFERENCES

[1] Altman, E. Financial Ratios, "Discriminant analysis, and the prediction of corporate bankruptcy", Journal of Finance vol. 23(4), 1968, 598-609.

[2] Charitou, A. Neophytou, E.,& Charalambous, C. "Predicting corporate failure: empirical evidence for UK", European Acc. Review, vol. 13(3), 2004, pp.465-497.

[3] Gallinari, P., Thiria, S., Badran F., and Fogelman-Soulie, F. "On the relations between discriminant analysis and multilayer perceptrons", Neural Networks, vol 4, 1991, pp. 349-360.

[4] Joachims, T. "Learning to classify text using support vector machines", Kluwer Academic Publishers, Boston, 2002.

[5] Jones, M., "Financial diagnosis of irish companies using self-organising neural networks", In Proceedings of the 9-th Irish Academy of Management Annual Conference, September, 7-9, Galway, Ireland, 2005.

[6] Kumar, P. and Ravi, V. "Bankruptcy prediction in banks and firms via statistical and intelligent techniques", European Journal of Operational Research vol. 180 (1), 2007, pp. 1-28.

[7] Lacher, R., Coats, P., Sharma, S., L.F. Fantc, L. "A neural network for classifying the financial health of a firm", European Journal of Operational Research vol. 85, 1995, pp. 53–65.

[8] Nachev, A., "MLP and default ARTMAP neural networks for business classification", In Proceedings of 4th International Conference of Intelligent Systems & Knowledge Engineering ISKE'09, Hasselt, Belgium, in press, 2009.

[9] Nachev, A., Hill, S., Stoyanov, B. "Insolvency prediction of irish companies using backpropagation and fuzzy artmap neural networks" in J. Filipe & J. Cordeiro (Eds.), Lecture Notes in Business Information Processing, LNBIP 24, Springer-Verlag, Berlin Heidelberg pp. 287–298, 2009.

[10] Serrano-Cinca, C., "Self organizing neural networks for financial diagnosis". Decision Support Systems vol. 17, 1996, pp. 227–238.

[11] Vapnik, V., "The nature of statistical learning theory". Springer, New York, 1995.

[12] Vapnik, V., Kotz, S., "Estimation of dependences based on empirical data", Springer, New York, 2006.