December 2005

Using Support Vector Machines to Evaluate Financial Fate of Dotcoms

Indranil Bose
The University of Hong Kong

Pal Raktim
James Madison University

Using Support Vector Machines to Evaluate Financial Fate of Dotcoms

Indranil Bose
School of Business
The University of Hong Kong
bose@business.hku.hk

Raktim Pal
College of Business
James Madison University
palrx@jmu.edu

Abstract

The success of dotcoms has been short lived. Within a short span of a few years the dotcoms experienced a meteoric rise and suffered a dramatic fall. This research is a support vector machine based investigation of dotcoms from a financial perspective. Data from the financial statements of survived and failed dotcoms are collected and 24 financial ratios are computed. The financial ratios are analyzed using support vector machines to find out whether they can predict the financial fate of companies. The results show that support vector machines can predict the financial fate of dotcoms with at most 76% accuracy. Several numerical experiments are conducted to check the impact of size of training sample, size of testing sample, ratio of training and testing sample size, and balance of sample on the classification accuracy of support vector machines.

Keywords: Classification, Dotcoms, Financial Ratios, Support Vector Machines

1. Introduction

The phenomenal growth in Internet related e-tailing has suffered a major setback in the early years of the new millennium. Within a very short time, several corporations that have seen phenomenal growth in their stock prices in the late nineties have gone out of business. Several factors are thought to be responsible for this demise and are detailed in Sharma (2001). The objective of this research is to use support vector machines to identify if survival or attrition of the dotcoms can de predicted based on financial ratios obtained from financial statements of these companies. The results of this research may offer directives to emerging companies that plan to do business on the Internet in the coming years. It will be able to provide guidelines about which financial factors to monitor closely for long-term competitive advantage in the market. Also it will be useful for investors who plan to invest in similar companies.

2. Literature Review

Due to the importance of predicting the financial health of an organization, this is a widely researched area. Different approaches and techniques have been used for forecasting the likelihood of failures as indicated by Wong et al. (2000) and it is quite common to use financial ratios. The interest is this area has been spurred by Altman (1968) with the use of multiple discriminant analysis with five financial ratios, to predict the risk of failure. In addition to discriminant analysis, several other techniques have been used. All these techniques have used financial ratios as independent variables. Notable examples have included the use of multi-criteria decision aid methodology by Eisenbeis (1977) and Dimitras et al. (1996), logit analysis by Ohlson (1980), probit analysis by Zmijewski (1984), recursive partitioning algorithm by Frydman et al. (1985), expert systems by Messier and Hansen (1988), mathematical programming methods by Gupta et al. (1990), survival analysis by

Luoma and Laitinen (1991), multi-factor model by Vermeulen et al. (1998), rough sets by Dimitras et al. (1999), and neural networks by Tam and Kiang (1992).

Support vector machine is a relatively new technique that has recently caught the attention of researchers for solving classification problems. The use of this technique in business has been limited so far. Support vector machine has been used for forecasting of financial time series. Some examples in this direction have included the work done by Cao and Tay (2001), and Tay and Cao (2001). Use of support vector machines for bankruptcy prediction has been reported by Fan and Palaniswami (2000). They have reported that support vector machines have outperformed neural networks and linear discriminant analysis in terms of generalization performance for identifying bankruptcy predictor variables.

From the above discussion it can be meaningfully concluded that several competitive techniques have been used for studying the problem of prediction of failures of corporations using financial ratio analysis. But to the best of the authors' knowledge, there has been no study on prediction of survival or failure of dotcoms using support vector machines. Dotcoms are a recent phenomenon, and it is conjectured that since dotcoms are formed differently and also function quite differently from traditional corporations, a different set of financial factors may be responsible for their demise than what has been reported in the literature. It is also not known whether it is possible to use techniques like support vector machines for prediction of financial health of such companies. Even it is possible to do so, it is not known whether traditional financial ratios like price to earnings ratio, quick ratio, retained earnings/total assets etc. will be as predictive as they have been for traditional companies.

3. Research Methodology

In this paper dotcoms and online subsidiaries of brick-and-mortar corporations are studied to find out whether they will survive or cease to exist. The research described in this paper follows four key steps.

3.1 Data Collection

For this study, financial information for two hundred and forty publicly traded corporations for the last decade (1993-2003) is collected from the WRDS (Wharton Research Data Services) database. Corporations are identified to be dotcoms if they have the suffix ".com" in their company name. Financial data for these companies have included details like price to earnings ratio, dividends per share, current assets and liabilities, net revenues etc. Based on the collected financial data, twenty-four financial ratios are calculated for the dotcoms. The list of the twenty-four financial ratios that is used in this paper is given in Table 1.

3.2 Data Cleaning and Preprocessing

The collected data is categorized into two classes – firms that have survived and firms that have failed. A firm is said to have survived it if is publicly traded (as of August 2003) and if the stock price is more than 1 cent and these types of firms are indicated by '1'. If a firm previously has traded publicly but currently is trading at a price below one cent then it is considered to have failed and is indicated by '0'. A balanced data sample is constructed with equal representation of both types of firms. It is found that the data retrieved from WRDS has missing financial data items. In order to remove the 'noise' from the data a '0' value is substituted for the missing financial data items. The next step involves formation of random samples from the 'clean' data. Each random sample is divided into two parts – the 'training' random sample and the 'testing' random sample. The ratio of training to testing random sample size is initially taken as 80:20 and care is taken to include all two hundred and forty dotcoms in either the training sample or the testing sample, so that no data is lost in the process. In this way, ten random training samples and ten random testing samples are created.

Of these samples, two random samples of training and the corresponding two random samples of testing are found to be balanced (equal number of '0' and '1'). The composition of the ten random samples is given in Table 2.

3.3 Data Analysis

This is the most important step in data mining. The problem under study is a classification problem where the inputs consist of the twenty-four financial ratios of the company and the output is a binary output. The goal is to classify the firms with unknown outputs (i.e., the testing sample) as '0' or '1' and to identify the variables that play the most important role in the process. The analysis is conducted using support vector machines (SVM).

SVM is a new pattern recognition method proposed by Vapnik (1995). SVM uses the concept of statistical learning theory to classify an input vector into known output classes. SVM starts with a set of training data with several data points from two linearly separable classes (e.g., in this case '0' and '1') and obtains the optimal hyper-plane that maximizes the separation of the two classes. In general, for linearly separable data there can be many hyper-planes that can separate the data into two classes. The optimal hyper-plane is the one that, in addition to separating the classes without error, maximizes the margin or the sum of the absolute deviations between the closest training data in each class and the hyper-plane. For the linearly separable case, this is equivalent to finding the solution to a quadratic programming problem. For non-linearly separable data, it uses the kernel method to transform the input space into a high dimensional feature space, where an optimal linearly separable hyper-plane can be constructed. Some examples of kernel functions are linear function, polynomial function, radial basis function, and sigmoid function. The advantage of using SVM is that it can yield good accuracy for classification of high dimensional data and at the same time it does not require any assumptions about the distribution of the input data.

3.4 Interpretation and Evaluation

In the last step of data mining, the numerical results obtained are studied to find out what combination of input variables lead to higher classification accuracy. Intuitive explanations of the observed phenomena are provided using basic knowledge of accounting and finance.

4. Numerical Experimentation

For SVM analysis, the LIBSVM software (Chang and Lin, 2001) is used. Using one of the random samples, the best choices for the parameters gamma and cost (for radial basis kernel function) are identified and the SVM analysis is repeated for all random samples using the same values for gamma and cost. Then, type I and type II accuracies are calculated for the training and the testing samples. The type I accuracy is calculated by finding the percentage of cases where a '0' (failed) is correctly identified as a '0' (failed) and the type II accuracy is calculated by finding the percentage of cases where a '1' (survived) is correctly identified as '1' (survived). The values of the tuned parameters gamma and cost are reported in Table 3 as well as the classification accuracies for several choices of variables. It can be seen that SVM provides good training and testing accuracies for all reported models. For most experiments using SVM it is observed that the Type II accuracies are higher than Type I accuracies. Also, it is seen that SVM results in high training accuracies. This is probably due to over-fitting in case of SVM. Model 19 results in moderate testing accuracy for SVM and so this model is chosen for subsequent sensitivity analysis.

4.1 Impact of Training to Testing Ratio

In the numerical experiments the training to testing ratio of 80:20 is used. Next, the training to testing ratio is varied by using three additional choices 75:25, 70:30 and 60:40. The

experiments are run for a randomly chosen model (model 19 in this case) consisting of ten variables TD/TA, OI/TA, CF/S, RE/TA, S/TA, GP/TA, OI/S, OI/MC, C/S, and NI/(TA-TL). The results are reported in Table 4. It is observed that the best total testing accuracy is obtained for the 75:25 ratio. A possible explanation of this may be that the 80:20 ratio tends to over-train and hence SVM does not perform as well as that in the case of 75:25. It is noticed that for all choices of ratios, the type II accuracy is always reported to be higher than the type I accuracy for both training and testing samples.

4.2 Impact of Size of Training Sample

Using the training to testing ratio of 80:20, the effect of changing the size of the training sample on the training and testing accuracies is studied. The training sample is reduced by 10% and 20% and the results are reported in Table 5. While the best total classification testing accuracy is obtained when the training sample is reduced by 10%, the best total classification training accuracy is obtained in the case where the training sample is reduced by 20%. The accuracy increases with decrease in training sample size indicating that overtraining is present in the analysis.

4.3 Impact of Size of Testing Sample

Using the training to testing ratio of 80:20, the sample size for training is kept fixed and the sample size for testing is reduced by 10% and 20% successively; and the results are reported in Table 6. The highest total testing accuracy is reported when the testing sample is reduced by 10%. Also, type I and type II testing accuracies show the same pattern of change as the size of the testing sample is reduced. The total testing as well as training accuracies is not affected much by the change in the size of the testing sample.

4.4 Impact of Balance of Sample

A sample is said to be balanced if it has equal number of 'survived' and 'failed' firms. When using the 80:20 ratio, there are two balanced samples among the ten random samples that are created. The average classification accuracies for the balanced and the unbalanced samples are listed in Table 7. From Table 7, it can be observed that type I, type II, and total testing accuracies are slightly more for the unbalanced samples. On the other hand while Type II training accuracies are slightly more for unbalanced samples, type I training accuracies are less for the same samples.

5. Conclusion

In this paper SVM is used to find out whether it is possible to predict the survival or failure of dotcoms based on financial ratios. It is observed that in most cases the average type II accuracy is higher than the average type I accuracy indicating that it is easier to detect a survived firm than a failed firm. SVM suffers from overtraining to an extent and the testing accuracies are mostly lower than the training accuracies. It is also interesting to note that the best case overall testing accuracy obtained using SVM is 76.04% and it is higher than the best case testing accuracy of 69.17% reported for failure prediction of traditional companies using SVM (Fan and Palaniswami, 2000). The impact of various factors such as training to testing ratio, size of training sample, size of testing sample, balance of sample on classification accuracy is studied and it is found that they may affect the result to some extent. This research also supports the conjecture that it is possible to predict survival or failure of dotcoms using traditional financial ratio analysis to an extent. Please note that most of the ten variables in model 19 are traditional financial ratios.

In future more extensive studies can be conducted using other methods of data mining so that their performance can be compared to those of SVM. SVM is a promising method and further

studies need to be conducted to ascertain if it is possible to identify the most predictive input variables from the SVM models and also to find out if the identified variables are similar to those reported in the literature.

6. References

Altman, E. I. "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," *The Journal of Finance* (23), 1968, pp.589-609.

Cao, L. J., and Tay, F. E. H. "Financial forecasting using support vector machines," *Neural Computing Applications* (10), 2001, pp. 184-192.

Chang, C-C., and Lin, C-J. *LIBSVM: A library for support vector machines,* 2001. Software available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm

Dimitras, A. I., Slowinski, R., Susmaga, R., and Zopounidis, C. "Business failure prediction using rough sets," *European Journal of Operational Research* (114), 1999, pp. 263-280.

Dimitras, A. I., Zanakis, S. H., and Zopounidis, C. "A survey of business failures with an emphasis on prediction methods and industrial applications," *European Journal of Operational Research* (90), 1996, pp. 487-513.

Eisenbeis, R.A. "Pitfalls in the application of discriminant analysis in business and economics," *The Journal of Finance* (32), 1977, pp. 875-900.

Fan, A., and Palaniswami, M. "Selecting bankruptcy predictors using a support vector machine approach," in *Proceedings of the International Joint Conference on Neural Networks,* 24-27 July 2000, Como, Italy.

Frydman, H., Altman. E. I., and Kao, D-L. "Introducing recursive partitioning for financial classification: the case of financial distress," *The Journal of Finance* (40:1), 1985, pp. 269-291.

Gupta, Y. P., Rao, R. P., and Bagghi, P. K. "Linear goal programming as an alternative to multivariate discriminant analysis: a note," *Journal of Business Finance and Accounting* (17:4), 1990, pp. 593-598.

Luoma, M., and Laitinen, E. K. "Survival analysis as a tool for company failure prediction," *Omega*, (19:6), 1991, pp. 673-678.

Messier, W. F., and Hansen, J. V. "Inducing rules for expert system development: an example using default and bankruptcy data," *Management Science* (34:12), 1988, pp. 1403-1415.

Ohlson, J. A. "Financial ratios and the probabilistic prediction of bankruptcy," *Journal of Accounting Research*, 1980, pp. 109-131.

Sharma, A. "Dot-coms' coma," *The Journal of Systems and Software* (26), 2001, pp. 101-104.

Tam, K.Y., and Kiang, M.Y. "Managerial applications of neural networks: the case of bank failure predictions," *Management Science* (38:7), 1992, pp. 926-947.

Tay, F. E. H., and Cao, L. "Application of support vector machines in financial time series forecasting," *Omega* (29), 2001, pp. 309-317.

Vapnik, V. N. *The nature of statistical learning theory*, Springer-Verlag, New York, 1995.

Vermeulen, E. M., Spronk, J., and Van der Wijst, N. "The application of the multi-factor model in the analysis of corporate failure," in *Operational tools in the management of financial risks*, Zopounidis, C. (ed.), Kluwer Academic Publishers, Dordrecht, 1998, pp. 59-73.

Wong, B. K., Lai, V. S., and Lam, J. "A bibliography of neural network business applications research: 1994-1998," *Computers and Operations Research* (27), 2000, pp. 1045-1076.

Zmijewski, M. E. "Methodological issues related to the estimation of financial distress prediction models," *Studies on Current Econometric Issues in Accounting Research*, 1984, 39-82.

Variables	Symbols	Description
1	WC/TA	Working Capital/Total Assets
2	TD/TA	Total Debt/Total Assets
3	CA/CL	Current Assets/Current Liabilities
4	OI/TA	Operating Income/Total Assets
5	NI/TA	Net Income/Total Assets
6	CF/TD	Cash Flow/Total Debt
7	QA/CL	Quick Assets/Current Liabilities
8	CF/S	Cash Flow/Sales
9	RE/TA	Retained Earnings/Total Assets
10	S/TA	Sales/Total Assets
11	GP/TA	Gross Profit/Total Assets
12	NI/SE	Net Income/Shareholders' Equity
13	C/TA	Cash/Total Assets
14	I/S	Inventory/Sales
15	QA/TA	Quick Assets/Total Assets
16	P/E	Price Per Share/Earnings Per Share
17	S/MC	Sales/Market Capitalization
18	CA/TA	Current Assets/Total Assets
19	LTD/TA	Long Term Debt/Total Assets
20	OI/S	Operating Income/Sales
21	OI/MC	Operating Income/Market Capitalization
22	C/S	Cash/Sales
23	CA/S	Current Assets/Sales
24	NI/(TA-TL)	Net Income/(Total Assets – Total Liabilities)

Table 1: List of financial ratios used in data analysis

Sample No.	Training			Testing		
	No. of '0'	No. of '1'	Total No.	No. of '0'	No. of '1'	Total No.
1	99	93	192	21	27	48
2	98	94	192	22	26	48
3	100	92	192	20	28	48
4	93	99	192	27	21	48
5	96	96	192	24	24	48
6	100	92	192	20	28	48
7	96	96	192	24	24	48
8	97	95	192	23	25	48
9	93	99	192	27	21	48
10	98	94	192	22	26	48

Table 2: Composition of the ten random samples

Model No.	Parameters for SVM		SVM					
			Training			Testing		
	Cost	Gamma	Type I	Type II	Total	Type I	Type II	Total
1	64	0.0020361	80.69	84.47	82.60	73.30	73.01	73.13
2	128	0.0030436	81.72	80.91	81.35	71.76	72.12	71.88
3	1024	0.000681	77.18	90.72	83.91	68.62	82.92	76.04
4	64	0.0195052	85.34	92.32	88.80	60.57	75.71	68.13
5	64	0.0009766	73.37	97.15	85.16	61.59	89.05	75.63
6	4096	0.0001402	76.04	88.56	82.29	66.85	78.27	72.50
7	512	0.0001327	81.22	70.36	75.89	79.54	63.10	71.04
8	256	0.000874	79.47	84.17	81.82	74.87	73.03	73.96
9	4096	0.0003221	77.20	90.50	83.80	67.98	78.81	73.54
10	128	0.0015864	80.49	89.22	84.84	69.63	74.79	72.08
11	128	0.0017003	82.14	86.28	84.22	74.09	74.41	74.38
12	4096	0.0001221	79.33	76.33	77.92	76.55	68.82	72.50
13	64	0.0010467	72.25	97.25	84.64	62.53	88.00	75.63
14	128	0.0206173	85.76	93.99	89.84	60.69	73.58	67.29
15	4096	0.0001221	77.57	83.28	80.47	74.15	73.94	73.96
16	256	0.0006905	74.19	91.75	82.92	63.74	82.61	73.33
17	256	0.0019531	81.94	83.75	82.86	73.02	71.15	72.08
18	128	0.0118415	84.82	82.17	83.54	68.87	70.05	69.58
19	32	0.0015864	71.54	96.62	83.96	58.82	88.77	74.38
20	512	0.0009112	75.95	90.92	83.39	68.02	80.70	74.58

Table 3: Analysis using SVM

Ratio	80:20	75:25	70:30	60:40
Training				
Type I	71.54	78.21	75.66	80.06
Type II	96.62	95.94	93.71	93.31
Total	83.96	86.94	84.58	86.88
Testing				
Type I	58.82	72.58	68.39	65.18
Type II	88.77	82.75	84.03	80.4
Total	74.38	77.83	76.39	72.5

Table 4: Effect of changing the training to testing ratio

Percentage Reduction of Training Sample	0%	10%	20%
Training sample size	192	173	154
Type I	71.54	75.95	91.71
Type II	96.62	95.56	97.64
Total	83.96	85.72	94.61
Testing sample size	48	48	48
Type I	58.82	65.64	66.62
Type II	88.77	83.33	72.55
Total	74.38	74.58	69.79

Table 5: Effect of changing the size of the training sample

Percentage Reduction of Testing Sample	0%	10%	20%
Training sample size	192	192	192
Type I	71.54	71.02	71.02
Type II	96.62	96.94	96.94
Total	83.96	83.85	83.85
Testing sample size	48	43	38
Type I	58.82	63.57	60.23
Type II	88.77	89.1	88.03
Total	74.38	76.98	74.74

Table 6: Effect of changing the size of the testing sample

Nature of Samples	Balanced	Unbalanced
Number of Training samples	2	8
Type I	72.92	71.19
Type II	95.83	96.81
Total	84.38	83.85
Number of Testing samples	2	8
Type I	58.33	58.94
Type II	87.50	89.09
Total	72.92	74.74

Table 7: Effect of changing the balance of the sample