Combining Cluster Analysis and Neural Networks for the Classification Problem

Kyungsup Kim and Ingoo Han
Management Information Systems, Korea Advanced Institute of Science and Technology,
P.O.Box, 201, Cheongryang, Seoul 130-650, Korea

Abstract

The extensive researches have compared the performance of neural networks(NN) with those of various statistical techniques for the classification problem. The empirical results of these comparative studies have indicated that the neural networks often outperform the traditional statistical techniques. Moreover, there are some efforts that try to combine various classification methods, especially multivariate discriminant analysis with neural networks. While these efforts improve the performance, there exists a problem violating robust assumptions of multivariate discriminant analysis that are multivariate normality of the independent variables and equality of variance-covariance matrices in each of the groups. On the contrary, cluster analysis alleviates this assumption like neural networks. We propose a new approach to classification problems by combining the cluster analysis with neural networks. The resulting predictions of the composite model are more accurate than each individual technique.

1. Introduction

The extensive researches have compared the performance of neural networks(NN) with various statistical models and other artificial intelligent models for the classification problem. The empirical results of these comparative studies have indicated that the neural networks often outperform the traditional statistical techniques. The success of NN is due to the capability of generalized prediction, when new data are entered after NN are trained.

In addition to comparative studies of each classification model, there are some efforts that try to combine various classification methods, especially multivariate discriminant analysis(MDA) and NN. These efforts have resulted in better performance than each separate model[9,10].

According to Kim and Kim, "An integrating approach which makes full use of both statistical and neural networks techniques offers the promise of increasing performance over each method alone"[8]. We were motivated to explore a model combining NN and other statistical techniques except MDA which have some strict assumptions.

The outstanding statistical techniques applied in business classification are MDA and cluster analysis (CA). So far, there has been no study performing classification in business domain with CA. The large part of Hybrid models consists of combination of MDA with NN, especially using backpropagation (BP), a supervised learning algorithm of the NN. Exceptionally, Lee. et. al. proposed a hybrid model combining a MDA model, Kohonen's self-organizing feature map(SOFM), an unsupervised learning algorithm, linear vector quantization (LVQ) and BP, two supervised models. Their models make the best use of every functions selecting input variables of MDA, clustering cases of SOFM, refining uncertainty of unknown data of LVQ and generalizing prediction of BP. They reported that this model outperform all the others[9].

Meanwhile, Clustering techniques are used in engineering domain and unsupervised learning algorithms of NN, for example, SOFM, LVQ and adaptive resonance theory. Adjusting weights used in preceding NNs are quite similar to typical partitioning CA, such as k-means clustering.

We propose a hybrid NN model integrating CA and BP. In this model, we are to examine that the performance of classification can be improved, when information from CA inputs BP. The favorable result are meant to the utilization of CA as a preprocessing.

In regard to domain, We select bankruptcy prediction of the business classification problems, for example, bankruptcy, credit evaluation, bond rating, and so on.

The rest of this paper is organized as follows. The models used in our study are discussed in the next section. Research data and modeling is presented in section 3. Empirical results are provided in section 4. Concluding remarks are presented section 5.

2. The classification models

In this section, we discuss the characteristics of

classification models used in our empirical tests: MDA, CA, NN, a CA-assisted NN.

2.1. Multivariate discriminant analysis

The MDA model is based on Fisher procedure which constructs a discriminant function by maximizing the ratio of within groups variances to between-groups variances. Linear Fisher classifiers derived from the procedure are known to be optimal in minimizing the expected cost of misclassifications, provided the following conditions are satisfied[9]:

1. each group follows a multivariate normal distribution.
2. the covariance matrices of each group are equal.
3. mean vectors, covariance matrices, and prior probabilities of misclassification are known.

In bankruptcy prediction, financial ratios are used as independent variables while the state of bankruptcy or non-bankruptcy is used as dependent variables. The violation of multivariate normality assumptions for independent variables frequently occurs in the financial data[3]. The stepwise method of MDA is used as preprocessing mechanism for selecting important input variables which will be used in the NN model and the CA model.

2.2. Cluster Analysis

The CA model is an alternative statistical classifier. CA separates the component cases into groups with their the computation of a distance or similarity, so as to identify homogeneous groups or clusters.

Although CA and MDA classify objects or cases into categories, MDA requires you to know group membership for the cases used to derive the classification rule. In CA, group membership for all cases is not known and the number of groups is often not known. On the contrary, cluster analysis alleviates this assumption like neural networks. CA is different to MDA as following Table 1.

Table 1. Characteristic of multivariate statistical methods

technique	actual categories	number of categories	number of variables
discriminant analysis	prior knowledge	prior knowledge	prior knowledge
cluster analysis (hierarchical) (k-means)	posterior posterior	posterior prior	prior knowledge

CA is divided to hierarchical and partitioning techniques. Hierarchical techniques cluster component cases into clusters at various levels. Partitioning techniques form clusters by optimizing some specific clustering criterion, so to speak, the number of clusters[4]. The outstanding partitioning technique is k-mean algorithm. Unlike hierarchical clustering techniques, the algorithm, that effect a partition of data do not require that the allocation of an object to a cluster be irrevocable and can be used to cluster large numbers of cases efficiently.

2.3. NN

NN does not impose any kind of strict assumptions on input variables such as MDA. In this studies, we use back-propagation model of a supervised learning model.

2.4. CA-assisted NN

CA-assisted NN also does not impose any kind of strict assumptions on input variables such as MDA. The information from CA procedure, cluster memberships and the Euclidean distance between individual cases and centroids (means) are the input variables to NN.

3. Research data and modeling

3.1. Research data

Data used in our study are benchmark test data and Korean middle firms data in domain of bankruptcy prediction.

The first data sample come from Odom and Sharda's bankruptcy classification study using NN. They applied NN with Altman's 1968 study. Altman's MDA model a famous comparison to subsequent bankruptcy classification studies, especially using DA and NN. They also used the same financial ratios that Altman used in his 1968 study. These ratios are[1]:

- X1 Working capital to Total assets
- X2 Retained earnings to Total assets
- X3 EBIT to Total assets
- X4 Market value of equity to Total Debts
- X5 Sales to Total assets

Odom and Sharda selected financial ratios of sample firms that went bankrupt between 1975 and 1982. Total 129 firms data obtained from Moody's industrial manuals, consisted of 65 bankrupt firms and 64 non-bankrupt matched on industry and year.

They separated sample data to 74 firms (bankrupt 38, non-bankrupt 36) as a training data set and 55 firms (bankrupt 27, non-bankrupt 28) as validation data set. Data used for the bankrupt firms were from the last financial statements issued before the firms declared bankruptcy[11].

We also adopt their methods including variables and data. We complemented Levine's test for equality of variances to clarify the differences between bankrupt and non-bankrupt groups of each data set. The results implicated that training data set could be more classified than validation data set because the F-ratio (variance between groups to variance within group) of training set is greater than the one of validation set. Each F-ratio are at the following Table 2.

Table 2. The result of equality of variances between bankrupt and non-bankrupt samples

variables	training set	validation set
Working capital to Total assets (F-ratio, level of significance)	F=3.077 p=.084	F=0.487 p=.488
Retained earnings to Total assets	F=3.966 p=.05	F=4.512 p=.038
EBIT to Total assets	F=12.487 p=.001	F=0.008 p=.93
Market value of equity to Total Debts	F=9.246 p=.003	F=5.906 p=.019
Sales to Total assets	F=1.503 p=.224	F=0.212 p=.647

The second data consists of Korean middle size firms that failed in the period 1993-1995. Data were obtained from the lists of Korea Credit Assurance Funds. We selected total 440 firms, consisting of 220 bankrupt firms and 220 non-bankrupt matched on industry and year. We divided sample data to 310 firms (bankrupt 155, non-bankrupt 155) as training data and 130 firms (bankrupt 65, non-bankrupt 65) as validation data.

After a stepwise selection of input variables in MDA, Six financial ratios are remained. These ratios are:

X1 Increase rate of Total assets
X2 Increase rate of Stockholder's equity
X3 Ordinary Earning to Total assets
X4 Earning after Tax and Interest to Stockholder's equity
X5 Retained Earning to Total assets
X6 Increase rate of Operating Capitals

3.2. Research Models

Talking about benchmark test, we selected training data set to find generalized MDA model to predict the categories new input data in MDA. We used k-means clustering algorithm in CA. In NN and CA-assisted NN, we use 3-layer multi-layered perceptron models with BP as a learning method.

About the Korean bankruptcy test, earliest of all, we screened 6 input financial variables from 66 full variables using stepwise method of MDA. The rest is similar to benchmark test. The details of each research model are as follows.

Table 3. Research Models

model	input and output	structure of NN used
MDA	- 5/66 financial ratios - actual categories	
CA	- 5/6 financial ratios	
NN	- 5/6 financial ratios - actual categories	5-8-1(benchmark) 6-20-1(Korean)
CA-assisted NN	- 5 6 financial ratios - actual categories - membership, Euclidean distance from CA	6-12-1(benchmark) 8-23-1(Korean)

4. Empirical results

4.1. Benchmark test

In the benchmark test, we acquired promising results. CA-assisted NN performed best and its hit ratio scored highest in comparison with other existing studies[11,12,13]. On the other hand, CA reached the level of the other statistical classifier, MDA, in validation data.

Table 4. Hit-ratio of benchmark test

	hit ratio	validation
hit ratio of our study	MDA	41/55
	CA	40/55
	NN	44/55
	CA-assisted NN	46/55
hit ratio of established study	BP by Odom and Sharda	44/55
	Perceptron by Rahimian	45/55
	Athena by Rahimian	46/55
	MLP by Serra-Cinca	46/55

4.2. Test with medium-sized firms in Korea.

In Korean middle size firms test, .CA-assisted NN performed best as well. On the other hand, CA could not reach the level of MDA in validation data. But, the information from CA helped to improve the performance of CA-assisted as expected. It was because clustering is vulnerable to outliers and missing values and is dependent on initial means.

Concerning prediction error, CA-assisted NN brought about total error, 23.8%, as a whole.

Individually, with regard to type I error (classifying non-bankrupt as bankrupt), NN is superior to the others. About type II error (classifying bankrupt as non-bankrupt), MDA overtakes CA-assisted NN slightly. Type II error is more important than type II error because of the possibility incurring loss of investment.

Table 5. Hit-ratio and Type I, II errors of Korean middle firms test

MDA	72.3% (94/130)
CA	57.7% (75/130)
NN	73.8% (96/130)
CA-assisted NN	76.2% (99/130)

(MDA)

predict / actual	bankrupt	non-bankrupt	total
bankrupt	49	16 (type II)	65
non-bankrupt	20 (type I)	45	65
total	69	61	130

(CA)

predict / actual	bankrupt	non-bankrupt	total
bankrupt	46	19 (type II)	65
non-bankrupt	36 (type I)	29	65
total	82	48	130

(NN)

predict / actual	bankrupt	non-bankrupt	total
bankrupt	43	22 (type II)	65
non-bankrupt	12 (type I)	53	65
total	55	75	130

(CA-assisted NN)

predict / actual	bankrupt	non-bankrupt	total
bankrupt	48	17 (type II)	65
non-bankrupt	14 (type I)	51	65
total	62	68	130

5. Concluding remarks

Our study is an exploratory study that combines NN and other statistical models except for MDA, because of the strict assumptions pertaining to MDA. We found that CA has potential to be applied to business classification by the empirical results.

The results also showed that CA-assisted NN model outperform the other models as well as the existing studies. The reason that CA-assisted NN is superior to the other models is explained by the capability of CA, especially k-means algorithm. The algorithm have some favorable features of (1) clustering cases in the way of minimizing the distance between cases and centroid, (2) rearranging inappropriate cluster, and so on. The similar features are also found in SOFM and LQV, unsupervised learning NNs. Inferring from this relation, the results of SOFM-assisted BP or LVQ-assisted BP is expected to show the same level of performance, CA-assisted BP. We expect the comparative studies among CA-assisted, SOFM-assisted, and LVQ-assisted approach to business classification. We also expect the further studies on control techniques for outliers and missing values, and preparation of proper initial centroid.

REFERENCES

1. Altman, E. "Financial Ratios, Discriminant Analysis, and the Prediction of Corporate Bankruptcy", *Journal of Finance*, September, 1968, 589-609.

2. Berry, R. and D. Trigueiros. "Applying Neural Networks to the Extraction of Knowledge from Accounting Report : A Classification Study, " In R. R. Trippi and E. Turban(Eds.). *Neural Networks in Finance and Investing*, 1993. 103-124.

3. Deakin, E.D., "A Discriminant Analysis of Predictors of Business Failure, " *Journal of Accounting Research*(Spring 1976) 167-179.

4. Dillon, W. R. and Goldstein, M.(1984), *Multivariate Analysis : Methods and Applications*, JW&S, New York.

5. Han, I., J. Chandler, and T. Liang, "The Impact of Measurement Scale and Correlation Structure on Classification Performance of Inductive Learning and Statistical Methods," *Expert Systems with Applications*, 209-221.

6. Han, I., Kwon, Y. & Jo, H. "A Review of Artificial Intelligent Models in Business Classifications." *Journal of Expert Systems*, Vol. I(1), 23-41.

7. Kim, S. H. and D. Kang. "Implicit versus Explicit Forecasting: Case Study in Intraday Stock Index Prediction," *Working Paper*, KAIST.

8. Kim, S. H. and K. Kim, "Integrating Multivariate Statistics and Neural Networks for Financial Prediction : Case Study in Interest Rate Forecasting," *Working Paper*, KAIST.

9. Lee, K.C., I. Han, and Y. Kwon, "Hybrid neural network models for bankruptcy prediction, *Decision Support Systems*, 18(1996) 63-72.

10. Markham, I. S. and C. T. Ragsdale, "Combining Neural Networks and Statistical Predictions to Solve the Classification Problem in Discriminant Analysis." *Decision Science*, Vol. 26, No. 2, 229-242.

11. Odom, M. D. and R. Sharda, "A Neural Network for Bankruptcy Prediction," *International Joint Conference on Neural Networks*, June, 1990, Vol. II, 163-168

12. Rahimian, E., S. Singh, T. Thammancote and R. Virmani, "Bankruptcy Prediction by Neural Networks," In R. R. Trippi and E. Turban(Eds.), *Neural Networks in Finance and Investing*, 1993.

13. Serrano-Cinca, C., Self- organizing Neural Networks for Financial Diagnosis," *Decision Support Systems*, 17(1996) 227-238.