Prediction of financial information manipulation by using support vector machine and probabilistic neural network

Hulisi Öğüt a, Ramazan Aktaş a, Ali Alp a, M. Mete Doğanay b

a Department of Business Administration, TOBB University of Economics and Technology, Söğütözü Caddesi No. 43, Ankara 06560, Turkey
b Department of Business Administration, Çankaya University, Öğretmenler Caddesi No. 14, Balgat, Ankara 06530, Turkey

1. Introduction

Corporate governance has received much attention recently. One of the basic pillars of good corporate governance is transparency and disclosure. OECD corporate governance principles state that, “the corporate governance framework should ensure that timely and accurate disclosure is made on all material matters regarding the corporation, including the financial situation, performance, ownership, and governance of the corporation” (OECD, 2004). This important principle states that all related parties must receive timely and accurate information about the company so that they can make rational decision. A decision is rational if all the relevant information is considered. If the stakeholders of a company cannot receive this information, they cannot be expected to make rational decisions. Rational decisions are very important for the economy as scarce financial resources can be diverted to effective and efficient companies only by way of rational decisions. There are different ways to disseminate information to the stakeholders. The most common way to communicate financial information is through the publication of periodic financial statements. These statements contain information about the financial position, performance, cash flows, and related party transactions of a company. Financial statements must be comparable (all the companies must use the same accounting principles) and free of any material misstatements to serve their intended purpose. In this respect, OECD corporate governance principles stipulate that high quality accounting standards should be used and the financial statements must be audited by independent auditors to ensure fair representation. If financial statements are comparable and fair (free of material misstatements) then they can be relied upon by the decision makers. Decision makers can compare the financial position and the performance of different companies and at the same time they can be sure that the information they use to do this comparison is true. Investors are interested in the firms all over the world as globalization accelerates. Capital is flowing freely from one country to another. So, this situation makes comparable and reliable financial information more important. That is why an investor interested in a foreign company requires that its financial statements should be prepared in accordance with common accounting standards (such as IFRS) and should be audited by an independent and competent party.

The most important deviation from disclosure and transparency principle is the publication of misleading financial statements, those that do not present the financial position and the performance of a company fairly. In order to present their companies more favorably, managers can distort information in the financial statements. By doing this, managers may pursue personal gains. But, the decision makers who base their decisions on this information are misled and deceived. One of the main purposes of stating disclosure and transparency among corporate governance principles is to protect inventors and discourage companies from publishing financial statements that are not fair. Recent accounting scandals such as Enron and Worldcom have drawn further attention to this principle. The act of distorting information in financial
statements intentionally is called financial information manipulation. The main purpose of financial information manipulation is to deceive the investors by publishing financial statements that do not present the financial position and the performance of a company fairly.

As explained in the preceding paragraph financial information manipulation prevents decision makers from making rational decisions. So, it is of utmost importance to detect financial information manipulation and prevent it before the financial statements are made public. Researchers have developed models to predict whether the financial information manipulation occurred or not. One such recent study used the data complied in Turkey and aimed at predicting financial information manipulation in Turkey (Aktas, Alp, & Doğanay, 2007). Multivariate statistical methods and neural networks were used to predict financial information manipulation. While the prediction powers of the multivariate statistical methods were 53.3%, neural network classifier had 33.3% classification accuracy in the test sample. Although multivariate statistical methods provide useful information in detecting financial information manipulation, their performances were not very satisfactory. So, the authors of this paper have tried to find other methods whose performances are higher than those of the multivariate statistical methods. We have noticed that probabilistic neural network and support vector machine methods have not been used for this purpose yet. So, we decided to use these methods to predict financial information manipulation and compare their performances with those of the multivariate statistical methods.

This paper proceeds as follows: Section 2 discusses the results of the preceding studies briefly. Section 3 explains the data collection process. Section 4 gives information about probabilistic neural network and support vector machine. Section 5 presents the result of classifier and last section concludes the paper.

2. Preceding studies

There are different definitions of financial information manipulation (refer to Beneish, 2001 for these definitions). Financial information manipulation in this context is misreporting, in other words intentional fraudulent financial reporting. Fraudulent financial reporting is the violation of accounting standards, the omission of existing amounts or the inclusion of fictitious amounts (Arens & Loebbecke, 2000), DeAngelo (1986), Defond and Jiambalvo (1994), Degeorge, Patel, and Zeckhauser (1999), Erickson and Wang (1999), Jones (1991), Trussel (2003), and Shivakumar (2000) investigated why managers distort information in the financial statements in different organizations. DeAngelo (1986), Defond and Jiambalvo (1994), Healy (1985), Jones (1991), and Shivakumar (2000) used accrual-based methods to detect financial information manipulation whereas Beneish (1999) used probit which is a multivariate statistical method for the same purpose. Aktas et al. (2007) applied discriminant analysis, logistics regression (logit), and probit, all of which are multivariate statistical methods, and neural networks to detect financial information manipulation in Turkey. For this purpose, indices proposed by Beneish are used as independent variables. In their study, they took the multicollinearity into account and eliminated its effects by using stepwise or forward conditional estimation procedures as well. As explained before, they found that the performance of multivariate statistical methods were low and neural networks did not produce useful results.

Beneish (1999) used eight independent variables that are calculated from the financial statements data. Seven out of eight variables are indices. Beneish explains that he preferred indices because he wanted to capture distortions arising from manipulation by comparing financial statement information in the year of the first reporting violation with those in the year prior. Aktas et al. (2007) used the same independent variables in their study. Independent variables used by Beneish and the rational why he chose them are explained below:

Days’ sales in receivables index(X1) = \( \frac{\text{Receivables}_t}{\text{Sales}_t} \)

Beneish states that disproportionate increase in receivables in relation to sales may suggest revenue inflation

Gross margin index(X2) = \( \frac{\left( \frac{\text{Sales}_t - \text{COGS}_t}{\text{Sales}_t} \right)}{\text{COGS}_t} \)

(COGS: Cost of goods sold)

Deterioration of gross margin signals bad prospects for the company and Beneish states that companies with bad prospects are more likely to distort earnings.

Asset quality index(X3) = \( \frac{1 - \left( \frac{\text{Currentassets}_t + \text{PP&E}_t}{\text{Totalassets}_t} \right)}{1 - \left( \frac{\text{Currentassets}_{t-1} + \text{PP&E}_{t-1}}{\text{Totalassets}_{t-1}} \right)} \)

(P&P&E: Plant, property and equipment)

Beneish claims that an increase in asset quality index signals the possibility of cost deferral through capitalization

Sales growth index(X4) = \( \frac{\text{Sales}_t}{\text{Sales}_{t-1}} \)

Beneish says that growth companies are viewed by professionals as more likely to engage in financial information manipulation because of the pressure on the managers of such companies to achieve earnings targets.

Depreciation index(X5) = \( \frac{\text{Depreciation}_{t-1}}{\text{Depreciation}_{t-1} + \text{PP&E}_{t-1}} \)

Depreciation index greater than one indicates possible manipulation of depreciation expense

Sales, general and administrative expenses(SGA) index(X6) = \( \frac{\text{SGA}_t}{\text{Sales}_t} \)

Beneish uses this variable to detect disproportionate increase in sales, general and administrative expenses in relation to sales.

Leverage index(X7) = \( \frac{\text{TotalDebt}_t}{\text{TotalAssets}_t} \)

Beneish uses this variable to capture incentives in debt covenants for financial information manipulation

Total accruals to total assets(X8) = \( \frac{\text{TotalAccruals}_t}{\text{TotalAssets}_t} \)

(Total accruals = Δ Current assets – Δ Cash – Δ Current liabilities – Δ Current maturities of LTD – Δ Taxes payable – Depreciation and amortization)

Higher positive accruals are a sign of a possible financial information manipulation.

3. Data

The same data set in Aktas et al. (2007) is used in this study as well. Companies that distorted their yearly financial statements (financial statements as of 31 December) are compiled from Istanbul Stock Exchange Daily Bulletins and Capital Markets Board Weekly Bulletins. Financial information manipulations uncovered during the audits or during the investigations of the Capital Mar-
kets Board are published in those bulletins. Seventy-five firms that distorted their year-end financial statements were identified. We excluded financial services firms from our study because they have different accounting procedures.

Seventy-five manipulator companies were matched with 75 non-manipulator companies. The non-manipulators were selected from the same industry that the manipulators belong to and their sizes (market capitalizations) are similar. This matching process is accomplished by using Excel. For each year, all non-manipulator companies are listed broken down by industries and a non-manipulator from the industry that the manipulator belongs to is selected automatically. When more than one manipulator exists in an industry in a given year, attention is paid not to select the same firm as non-manipulator. So, a company is chosen as a non-manipulator in a given year only once. By using this method, we try to select non-manipulators that have similar properties with the manipulators. Final sample consists of 150 companies equally divided as manipulators and non-manipulators.

Balance sheets and the income statements of the manipulators and non-manipulators are obtained from Istanbul Stock Exchange. Beneish’s variables were used and their values are calculated for manipulators and non-manipulators. Year t is taken as the year in which the financial information manipulation occurred. Even though the financial information manipulation may occur in inter-year financial statements, we only use distorted year-end financial statements because only these year-end financial statements are published in the newspapers and only they are used widely by the investors.

4. Methodology

4.1. Probabilistic neural network

Probabilistic neural network (PNN) is first developed and proposed by Specht (1990). The structure of PNN is composed of input layer, pattern layer, summation layer and decision layer. The input layer is solely responsible of distribution of neurons to pattern layer. Probability density function is estimated by using multi-dimensional kernels in the pattern layer. Gaussian kernel is one of the most widely used kernels as convergence of the neural network is guaranteed (Rutkowski 2004). By summing and averaging, posterior probability density function is computed for each class in the summation layer. Depending on the information taken by all summation layers neurons and using Bayes’s decision rule, classification of pattern is done in the decision layer.

4.2. Support vector machine

Based on neural network and statistical theory, support vector machine (SVM) is used for classification, regression and density estimation. SVM has a number of successful applications in business such as marketing (Cui and Curry, 2005), bankruptcy prediction (Min & Lee 2005), customer loan evaluation (Li, Shue, & Huang, 2006), customer churn prediction (Coussement & Van den Poel, 2008). It has been found that SVM has outperformed other classifiers in many areas. SVM is implemented using quadratic programming. A quadratic program finds hyper-plane that minimizes misclassification error and maximizes margin between hyper-plane and nearest point. SVM classifier can map input space into higher feature dimensional space via nonlinear transformation. In this high dimensional space, there can be many hyper-plane that separates the data. However, only one of them achieves maximal separation.

We will briefly discuss details of SVM technique. For a detailed discussion, refer to (Burges 1998 and Vapnik 1995) Training examples can be defined as \( [x_i, y_i] \) where \( x_i \in \mathbb{R}^p \) and \( y_i \in [-1, 1] \) is the input vector, \( n \) is the dimension of input vector and \( y_i \) is the output vector. SVM finds optimal hyper-plane that separates one class from the other based on quadratic programming technique. This quadratic programming can be written mathematically as,

\[
\min_{w,b} \frac{1}{2}w^Tw + C \sum_{i=1}^{n} \xi_i
\]

subject to \( y_i(w^T\phi(x_i) + b) + \xi_i - 1 \geq 0 \)

\( \xi_i \geq 0 \)

where \( \phi(x_i) \) maps training data into high dimensional feature space, \( w \) is weight vector, \( b \) is the bias term, \( C \) is the penalty for the error term and \( \xi_i \) is the slack variable. (Vapnik 1995). This formulation can be solved by using Lagrange technique. Once optimal hyper-plane that separates one class from the other is constructed, classification decision is given by the following equation:

\[
f(y) = \text{sign} \left( \sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b \right)
\]

where \( \text{sign} \) is the sign function, \( \alpha_i \) is the parameter and \( K(x_i, x_j) = \phi(x_i)^T \phi(x_j) \) is Kernel function. Kernel function make implementation easier by considering only inner product rather than high dimensional feature space. Examples of kernel function are linear kernel function (\( K(x_i, x_j) = x_i^T x_j \)), polynomial kernel function (\( K(x_i, x_j) = \langle x_i^T x_j + r \rangle^d \)), radial basis function (\( K(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2) \)) and the sigmoid kernel function (\( K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r) \)) (Hsu, Chang, & Lin 2004).

5. Results

We used MATLAB for the implementation of classifier techniques. Eight attributes (indices) defined in Section 2 are used as input attributes. The output attribute takes the value of 1 for the manipulators and 0 for the non-manipulators. For comparison purposes, we used the same set of training and test data as in Aktaş et al. (2007). Training data is composed of 120 data and 60 of them belong to manipulated instances. Testing data is composed of 30 data and half of them belong to manipulated instances.

The performance of PNN classifier depends on the smoothing factor parameter. We used 4-fold cross validation technique in order to determine the smoothing factor parameter. The advantage of the cross validation technique is to prevent over fitting of training data. In an n-fold cross validation technique, training data is divided into n equal parts at first. Model is developed in n − 1 parts and it is tested in the remaining part. This process is repeated until all subset of data is tested with the model developed by the remaining part of training data. Once all subset of data is tested, model is retrained with the smoothing parameter having highest average classification accuracy and performance of PNN classifier is validated on the test data.

We varied smoothing parameter from 0 to 1 by 0.02 interval. Fig. 1 shows cross validation accuracy of smoothing parameters. As it can be seen, best accuracy is achieved when smoothing parameter is 0.56. The test and training performance of PNN using the smoothing parameter as 0.56 is summarized in Tables 2 and 3. We choose radial basis function (RBF) as a kernel for support vector machine. Although there is no established procedure for determining best kernel function, the advantages of using RBF as kernel are following: (i) sigmoid kernel performs similar to RBF for certain parameters (Lin & Lin 2003). (ii) It has been shown that performance of the linear kernel with parameter \( C \) is the same as RBF kernel with parameter \( C \) (Keerthi & Lin, 2003). Besides, while RBF kernel can nonlinearly map input space into higher dimensional feature space, linear kernel cannot do this (Hsu et al., 2004).
It is difficult to know which parameter combination performs better in the test data. A number of technique for choosing C and γ parameter is proposed by researcher. We used grid search technique proposed by Hsu et al. (2004). In the grid search technique, parameter space is searched with the combination of (C, γ). Although this method is fairly simple, computational time required for this search is not so much different than other heuristic or advanced methods. (Hsu et al., 2004). For the parameters, we consider exponential sequence of C = {2^{-5}, 2^{-3}, 2^{-1}, 2, 2^3, 2^5, 2^7, 2^9, 2^{11}, 2^{13}} and γ = {2^1, 2^3, 2^5, 2^7, 2^9, 2^{11}, 2^{13}}. Then every combination of C and γ parameters are tried in the training data using 4-fold cross validation as in PNN. Once parameter pair having best cross validation accuracy is identified, SVM is retrained with optimal parameters and performance of SVM classifier is validated on the test data. As it can be seen in Table 1, we chose (C, γ) = 2^{11}, 2^{-7} as our optimal parameter pairs. The performance of SVM classifier on the test and training data are reported in Tables 2 and 3.

We compare test performance of the classifiers based on the following statistics defined as:

- **Specificity**: Number of correctly classified non-manipulated instances/number of total non-manipulated instances.
- **Sensitivity**: Number of correctly classified manipulated instances/number of total manipulated instances.
- **Total classification accuracy**: Total number of correctly classified instances/total number of instances.

Aktaş et al. (2007) applied Logit, Probit, Discriminant Analysis classifiers on the same training and test dataset. For the comparison purpose, the findings of that study are presented on Tables 2–4 as well. In this analysis, it has been found that while classification accuracy of SVM and PNN classifier for test data are %70 and %60, respectively, Logit, Probit and Discriminant Analysis classifiers have total classification accuracy of %53.33. However, test performance comparison of the classifiers based on total classification accuracy may be misleading if errors have different misclassification costs. Two types of error may be encountered in our case. While type-1 error occurs when the model classifies manipulator as non-manipulator, type-2 error occurs when the model classifies non-manipulator as manipulator. If investor classifies manipulator as non-manipulator, she invests in a company having manipulated financial information. When the financial information manipulation is discovered, the values of the investment decrease and investor incurs loss. When investor classifies non-manipulator as manipulator, she misses a profitable investment opportunity. But cost of this type of error may be alleviated by investing in other alternatives. Consequently, higher sensitivity statistics is more important than total classification accuracy and specificity statistics in our context. As shown in Table 4, SVM and PNN have done much better than other classifiers in terms of sensitivity statistics. Thus, whether misclassification costs are taken into account or not, SVM and PNN classifiers show much better performance than other classifiers. Besides, specificity statistic of SVM is %60 on the test data compared to %46.7 obtained by other classifiers.
6. Conclusion

The aim of this research is to detect financial information manipulation by using some variables obtained from the financial statements. For this purpose, we use support vector machine and probabilistic neural network classifiers to estimate a suitable model and compare their performance with those of probit, logit and discriminant analysis which were applied by Aktas et al. (2007). We have found that SVM outperforms other models and the performance of PNN is satisfactory. However, there is one disadvantage of using SVM and PNN as a classifier: while it is not possible to derive probability values from SVM and PNN for classification, one can get probabilities from probit and logit models.

References