International Journal of
Intelligent Systems and Applications in Engineering

ISSN:2147-6799 www.ijisae.org Original Research Paper

Fraud Detection on Financial Statements Using Data Mining Techniques

Murat Cihan Sorkun*[1,2], Taner Toraman[2]

Accepted : 12/06/2017 Published: 30/09/2017

Abstract:This study explores the use of data mining methods to detect fraud for on e-ledgers through financial statements. For this purpose, data set were produced by rule-based control application using 72 sample e-ledger and error percentages were calculated and labeled. The financial statements created from the labeled e-ledgers were trained by different data mining methods on 9 distinguishing features. In the training process, Linear Regression, Artificial Neural Networks, K-Nearest Neighbor algorithm, Support Vector Machine, Decision Stump, M5P Tree, J48 Tree, Random Forest and Decision Table were used. The results obtained are compared and interpreted.

Keywords:*Data mining, fraud detection, financial statements, e-ledger,machine learning*

1. Introduction

"One who handles honey, licks his fingers." [12] This Turkish proverb has become a word that employees use to reduce their conscientious responsibility for the unfair gains they have achieved in areas where the risk of arrest is minimal.

Unfortunately, much more intensely, fraud, irregularity and corruption are being made in order to gain unfair advantage in both corporate and family businesses without any sectoral distinction. In an effort by the Association of Certified Fraud Examiners (ACFE) in 2016, it was reported that businesses lost 5% of their annual income through fraud [3]. According to the same report, it is reported that the total loss at 2,410 cases was $ 6.3 billion and 23% of the cases had lost $ 1 million.

There are various purposes of applying Accounting Fraud, one of the types of financial frauds. It is possible to list them as follows:concealment of embezzlement, partner misleading requests, desire to distribute less profit, concealment of corruption, desire for unintended incentives, tax evasion though [13].

The doors to the outside world of businesses are their financial statements. Financial statements are standardized financial reports that summarize the financial performance of the business. There are two main financial statements, the Balance Sheet and the Income Statement. The balance sheet reports the assets held by the company and the capital and debts that make up these assets. The Income Table shows the operating results of the company's income and expenses in a period. The net profit or loss of the operator is also reported on this statements [10]. Businesses are obliged to present their financial statements annually. Investors, shareholders and banks have information on the financial situation of companies by evaluating these financial statements and make critical decisions. For this reason, businesses can cheat on their financial statements or in their financial ledgers which are the source of these statements. These statements are inspected by the auditors to identify the fraud cases. However, growing data from enterprises makes the audit processes difficult and leads to long periods of time. This has led to the need to resort to innovative methods of fraud.

[1]*Galatasaray University,Istanbul– 34349, Turkey*
[2] *Idea TeknolojiÇözümleri,Istanbul–34398,Turkey*
** Corresponding Author:Email: mcsorkun@gmail.com*

The spread of electronic ledgers and the fact that the financial statements created from these books are kept in a regular structure facilitated the processing and data mining of these tables. Data mining is the discovery of patterns, relations, changes, irregularities, rules and statistically significant structures in data [11]. The studies show that data mining methods have been successful in detecting fraud on financial statements. Terzi examined data mining methods used in cheating control and he mentioned that the use of data mining methods would provide great advantages to businesses in the prevention of fraud and errors [14].

There are different data mining methods and fraud auditing studies on the financial statements in the literature. Kirkos et al. tried to identify firms that published fraudulent financial statements using Decision Trees, Artificial Neural Networks and Bayesian Networks [1]. In the study, 76 financial tables with half of them fraudulent were used. As a result of training and tests with 10 selected features, the best classifier was Bayes Networks with 90.3% success. Ravisankaret al. used data mining techniques such as Multilayer Neural Networks, Support Vector Machines, Genetic Programming, Group Method of Data Handling, Logistic Regression and Probabilistic Neural Network to identify companies that similarly use fraud in their financial statements [2]. The financial statement of 202 Chinese companies labeled as fraudulent in the study was trained with 10 selected features. Ata and Seyrek attempted to identify deceptions in financial statements using Decision Trees and Artificial Neural Networks [6]. Fraudulent financial statement related to 100 Turkish companies was used by the experts. Decision Trees showed 67.92% success in classification, while Artificial Neural Networks achieved 77.36% success. Hoogset al. tried to detect the deceptions in financial statements using Genetic Algorithm in their work [9]. Sharmaand Panigrahi have examined and compared data mining applications and fraud detection systems in the literature [7]. The findings of this compilation show that Logistic Models, Neural Networks, Bayesian Networks and Decision Trees data mining techniques are most widely applied to provide solutions to the problems of identification and classification of fraudulent data.

Investigations indicate that the financial statements do not have fixed characteristics that can accurately identify the fraud.

Apparaoet al. stated that the selection of features is very important, combined systems with supervised learning methods are inadequate and unsupervised methods are required as result of study [8]. In their work, Zhouand Kapoor examine the identification of different data mining techniques and fraudulent financial statements, and mentions that fraud methods change over time [5]. An adaptive self-updating system has been proposed in these changing methods. Kotsiantis et al. have attempted to identify fraudulent financial statements using hybrid data mining methods and to identify the discriminating factors in their work [4]. A total of 164 financial statements, 41 of them were labeled as fraudulent, were trained by hybrid methods used by Decision Tree, Artificial Neural Networks, Bayesian Networks, K-Nearest Neighbor, Rule Learner and Support Vector Machine. Hybrid methods used; Stack, Voting, Rating and Best Classifier methods. As a result of the studies, hybrid methods have been found to achieve more successful results than classifiers.

This study is organized as follows. The next section describes the preparation and labeling of the data set. In Section 3, the results of applying different data mining methods on the data set are reported with tables and the classification algorithms are compared. Section 4 contains some ideas for the future evaluation of the study findings.

2. Data Description

Businesses share their financial status with other organizations through financial statements. Financial statements are summary reports produced from financial ledgers and financial tables do not contain as much information as ledger. Therefore, the fraud made on the ledger can be hidden on the statements. Review the ledger where the most guaranteed method of making a financial statement can be clearly identified as fraudulent.

Financial statements used in our work are labeled using the e-ledger they are produced. In the labeling process, a rule-based e-ledger control application is used. The used application examines the relevant parts of the ledger and gives a fraud score ranging from 0 to 100 depending on the violation of the specified rules. Using this fraud score, 72 financial statements produced from e-ledger are labeled.

One of the most important points in fraud detection on financial statements is the process of determining the distinguishing characteristics. Apparaoet al.have highlighted the importance of feature selection in his review of studies on financial fraud detection and mentioned that there is no consensus on what best features are [8]. Although there are no precise distinguishing features, there are a number of ratios suggested in the literature. In addition to the suggestions in the literature, our work has been subject to the selection of features by adding the expertise we have obtained. A total of 9 properties obtained as a result of feature selection are shown in Table-1.

Table 1. Selected Features

(L.T. Debts + S.T. Debts) / Total Assets
(Fixed Assets / Equity)
(Current Assets / S.T. Debts)
(Liquid Assets / S.T. Borrowings)
(Liquid Assets + Securities) / S.T. Debts
Stocks / Net Sales
Current Assets / Total Assets
(Net Profit / Average Equity)
(Net Profit / Average Total Assets)

3. Methodology

Financial statements included in data sets used in literature studies are marked as fraudulent or not, and data mining study is based on classification for these two labels [1, 2, 4, 6]. The data used in this study are marked with fraud scores between 0 and 100. The data are trained by different machine learning methods. In the training process, features derived from the Balance Sheet and Income Table produced from 72 e-ledger were used. In the training process, Linear Regression, Artificial Neural Networks (ANN), K- Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Stump, M5P Tree and Decision Table were used. These algorithms were tested with two different test methods and their performance was measured. The test methods used are proportional split (66% Train 34% Test) and folding (10-fold). The successes of the results obtained by these methods are measured by the NRMSE (Normalized Root Mean Squared Error) metric. The obtained results are shown in Table-2 and Table-3.

$$\text{Root Mean Squared Error (RMSE)}=\sqrt{\frac{\sum_{i=1}^{n}(y_i-y)^2}{n}} \quad (1)$$

$$\text{Normalized Root Mean Squared Error}=\frac{RMSE}{y_{max}-y_{min}} \quad (2)$$

Table 2. Splitting Results

Method	NRMSE
Linear Regression	0,4
ANN	0,51
KNN	0,5
SVM	0,53
Decision Stump	0,28
M5P Tree	0,32
Decision Table	0,33

Table 3. Folding Results

Method	NRMSE
Linear Regression	0,34
ANN	0,46
KNN	0,5
SVM	0,4
Decision Stump	0,25
M5P Tree	0,29
Decision Table	0,33

Table 4. Classification Results

Method	TP-Rate	F-Measure
Random Forest	0,77	0,76
ANN	0,65	0,64
KNN	0,37	0,38
SVM	0,6	0,56
Decision Stump	0,81	0,8
M5P Tree	0,8	0,78
Decision Table	0,8	0,75

In addition to these studies, the labels were separated into 3 groups and tested with classification algorithms. The labels are grouped into low, medium and high fraud rates according to their grades. The data between 0-40 is low, between 41-80 is medium and between 81-100 are included in the high fraud rate group. Different classification algorithms have been tested with the newly tagged data set resulting from this grouping. In this study, Random Forest and J48 classifiers were used instead of Linear Regression and M5P trees. The success rates of the results obtained with these methods are measured by the True Positive Rate (TP-Rate) and the F-Measure metrics. The obtained results are shown in Table-4.

Comparison of different machine learning algorithms using both ungrouped and grouped data are shown in Figure-1 and Figure-2.

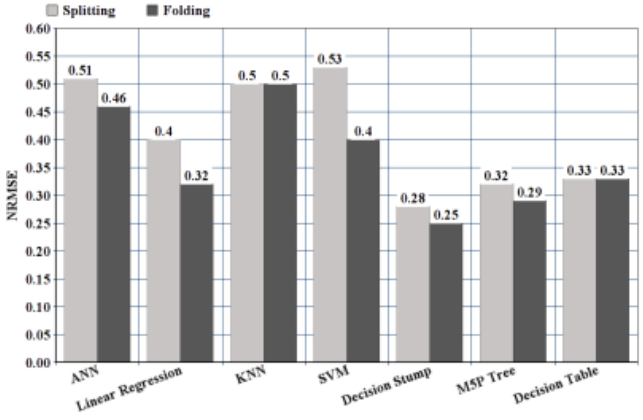

Fig. 1. Comparison of different machine learning algorithms

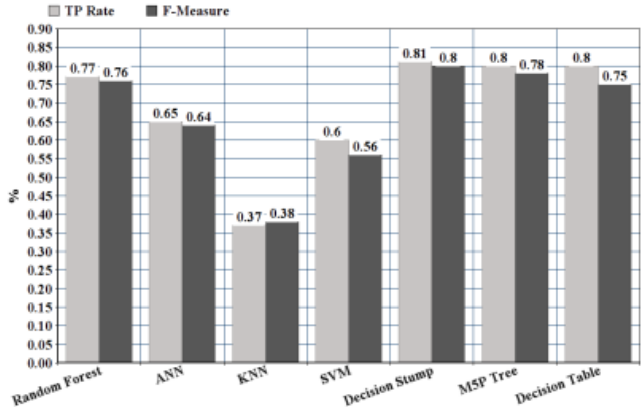

Fig. 2. Comparison of different classification algorithms using grouped data

4. Results and Discussion

As a result of the regression tests, the most successful algorithm has been Decision Stump for both the classification of the fraud on the financial statements and for the error checking. Other algorithms that succeeded in the classification tests were J48, Random Forest and Decision Table. These results show that decision trees are successful in detecting fraud on financial statements. In the regression tests, M5P, Decision Table and Linear Regression showed higher success than the other algorithms.

5. Conclusion

Fraud detection on financial data is matter always keeping up to date. Growing evidence of businesses makes it difficult to detect fraud on financial documents. The data mining methods enable the control of fraud on financial data, which identifies the relationships, patterns and irregularities on the data. The purpose of this study is to investigate the detection of fraud related to e-ledger over financial statements using data mining methods. Two different methods have been used in the fraud detection process. In the first method, financial statements are labeled with fraud scores and determined by regression. In the second method, fraudulent statements were labeled as low, medium and high and classified. The methods used in the fraud detection process; Linear Regression, ANN, KNN, SVM, Decision Stump, M5P Tree, Random Forest and J48. In the training process, Balance Sheet and Income Statements produced from 72 e-ledgers were used.

The results from the experiments have shown that data mining methods can detect the fraud factors between the financial statements and the e-ledger. On a performance basis, the Decision Stump Algorithm, which gives the best results for both experiments, found the fraud score with an error margin of 0.25 and classified the fraudulent groups 81% correctly.

Experiments made have also shown that feature selection is important. It is planned to extend the feature set and to use the Deep Learning methods in order to determine the characteristics that can detect the fraud in future studies.

Acknowledgements

This study was supported by TÜBİTAK TEYDEB under the project number 3150156 "Eldora: Electronic DocumentArchivingAppliance".

References

[1] Kirkos, E., Spathis, C., and Manolopoulos, Y. (2007). Data Mining Techniques for the Detection of Fraudulent Financial Statements. Expert Systems with Applications Vol. 32.4 Pages.995-1003.

[2] Ravisankar, P., Ravi, V., Rao, G. R., and Bose, I. (2011). Detection of Financial Statement Fraud and Feature Selection Using Data Mining Techniques. Decision Support Systems, Vol. 50, Pages. 491-500.

[3] Association of Certified Fraud Examiners (2016). The Staggering Cost of Fraud [Online]. Available: http://www.acfe.com/rttn2016/docs/Staggering-Cost-of-Fraud-infographic.pdf

[4] Kotsiantis, S., Koumanakos, E., Tzelepis, D., and Tampakas, V. (2006). Forecasting Fraudulent Financial Statements Using Data Mining. International Journal of Computational Intelligence, Vol. 3,Pages. 104-110. Asdasdas

[5] Zhou, W., and Kapoor, G. (2011). Detecting Evolutionary Financial Statement Fraud. Decision Support Systems, Vol. 50, Pages. 570-575.

[6] Ata, H. A., and Seyrek, I. H. (2009).The Use of Data Mining Techniques inDetecting Fraudulent Financial Statements: An Application on Manufacturing Firms. The Journal of Faculty and Economics and Administrative Sciences, Vol. 14. Pages.157-170.

[7] Sharma, A., and Panigrahi, P. K. (2013). A review of financial accounting fraud detection based on data mining techniques.International Journal of ComputerApplication, Vol. 39, Pages. 37–47.

[8] Apparao, G., Singh, A., Rao, G. S., Bhavani, B. L., Eswar, K., and Rajani, D. (2009). Financial Statement Fraud Detection by Data Mining. International Journal of AdvancedNetworking and Applications, Vol. 1. Pages. 159-163.

[9] Hoogs, B., Kiehl, T., Lacomb, C., and Senturk, D. (2007). A Genetic Algorithm Approach to Detecting Temporal Patterns Indicative of Financial Statement Fraud. Intelligent Systems in Accounting, Finance and Management, Vol. 15. Pages. 41-56.

[10] Çömlekçi, F. (2004). Muhasebe Denetimi ve Mali Analiz. Anadolu University Publication.

[11] Özkul, F. U., and Pektekin, P. (2009). Muhasebe Yolsuzluklarının Tespitinde Adli Muhasebecinin Rolü ve Veri Madenciliği Tekniklerinin Kullanılması. World of Accounting Science, Vol. 11.

[12] Turkish Cultural Foundation, Proverbs [Online]. Available: http://www.turkishculture.org/literature/literature/turkish-proverbs-133.htm

[13] Denetim İlkeve Esasları (2004). Maliye Hesap Uzmanları Derneği, Vol.1. Page. 151.

[14] Terzi, S., (2012). Hileve Usulsüzlüklerin Tespitinde Veri Madenciliğinin Kullanımı. Journal of Accounting & Finance, Vol. 54. Pages. 51-65