Journal of Computational Methods in Sciences and Engineering 8 (2008) 213–224
IOS Press

Support vector machines, Decision Trees and Neural Networks for auditor selection

Efstathios Kirkos[a], Charalambos Spathis[b] and Yannis Manolopoulos[c,*]
[a]*Department of Accounting, Technological Educational Institution of Thessaloniki, PO BOX 141, 57400, Thessaloniki, Greece*
[b]*Division of Business Administration, Department of Economics, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece*
[c]*Department of Informatics, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece*

Abstract. The selection of a proper auditor is driven by several factors. Here, we use three data mining classification techniques to predict the auditor choice. The methods used are Decision Trees, Neural Networks and Support Vector Machines. The developed models are compared in term of their performances. The wrapper feature selection technique is used for the Decision Tree model. Two models reveal that the level of debt is a factor that influences the auditor choice decision. This study has implications for auditors, investors, company decision makers and researchers.

Keywords: Auditor choice, audit quality, Big 4 auditors, data mining, SVM, MLP, C4.5

1. Introduction

By the end of each fiscal year the companies have to publish their financial statements (i.e. balance sheet, income statement and cash flow statement). To assure that the published financial statements reflect the real financial status of the company, external auditors are hired to perform an in-depth examination of the published account values and to issue their opinion regarding the proper disclosure. In this sense, the auditors are called to defend third parties' interests like the shareholders' or the creditors'. However, the auditor is hired by the auditee. This client-supplier relationship may obligate the auditor to serve its client's needs.

Managers, whose salaries and bonuses may be linked to the performance of the company, have the incentive to be engaged in aggressive reporting practices. Moreover, many boards of directors are dominated by managers. Under this scene, managers can significantly influence the decision of hiring or firing an auditor. The fear of losing a customer may motivate an auditor to report favorably to his customer. This way, the independence of the auditor is threatened and the quality of the audit is jeopardized. The quality of an audit reflects the likelihood that an auditor will discover and report any material misstatements – the higher the audit quality, the greater the chance of detection of misstatements.

In the literature it is generally acknowledged that the audit markets are segmented into at least two categories, the first-tier or Big auditors and the non-Big auditors. Big auditors are big international

*Corresponding author. Tel.: + 30 2310 991912; E-mail: manolopo@csd.auth.gr.

auditing firms. Some years ago there were 8 big auditors. After auditing firms' merges and the collapse of Arthur Andersen, there remain 4 Big Auditing firms i.e. KPMG, PricewaterhouseCoopers, Ernst & Young and Deloitte & Touche. All other auditors, which have a national or local reputation, are considered as non-Big auditors.

Big auditors are recognized as providers of higher quality auditing services [1–5]. Francis and Wilson [6] proposed a brand name approach. To assure fee levels, auditing firms invest in their brand name reputation. Reputation capital is costly to gain and easy to destroy. Performing poor quality audits could damage the brand name of the auditing firm and result in loss of future revenues. Big auditing firms also invest more in technology, training and facilities. Finally, big auditing firms with substantial resources and insurance coverage are expected to make significant payments in the event of an audit failure.

Hiring a brand name auditor, and thus signaling the quality of the financial statements, may have significant implications. Because of the perceived integrity and the lower information risk, higher quality audits could drive higher share prices. Vice versa, litigation against a brand name auditing firm could cause negative returns even for the auditor's clients that are not involved in the litigation [7].

The fact that the choice of a proper auditor is driven by several factors implies that it is possible to develop models capable of predicting the type of auditor. Data mining provides several classification methodologies, which can be used to develop predicting models. As opposed to other auditing related topics like bankruptcy prediction [8,9], fraud detection [10,11] or prediction of auditors' qualified opinions [12], these methodologies have not yet been applied for the purpose of predicting the type of external auditor.

In the present study three well known classification methods, derived from the field of data mining, are employed to predict the type of the auditor. The employed methods are the Decision Trees, the Neural Networks and the Support Vector Machines. These three methods are compared in terms of their capability of predicting the class of unknown observations. Significant factors which strongly influence the auditor selection decision are revealed. This study has implications for internal and external auditors, company decision makers, investors and researchers. It can also be used to predict the most probable outcome for the selection of auditor.

The paper proceeds as follows: Section 2 reviews prior research related to auditor selection. Section 3 describes the construction of our sample and the input variables' selection. Section 4 refers to models' results, interpretation and validation. Finally, section 5 presents the concluding remarks.

2. Prior research

Several researchers examined the auditor selection problem and obtained empirical results. Cravens et al. [13] attempted to determine factors which influence the auditor selection process. They used univariate statistics and a standardized z-score to perform comparisons. The results indicate that there are differences between clients of big and non big auditors, but also that there are differences among clients of different big auditors.

Chaney et al. [14] conducted a survey in non-publicly listed firms in the UK. Privately held firms are not compelled by market pressures to signal the quality of their financial statements. The employed method was OLS Regression. The findings suggest that in the absence of the pressure of the market, the clients choose the lower cost auditor available. Chaney predicted the auditor choice and then he used this information in the fee analysis. Fee analysis suggests that auditors structure their business in a manner appropriate for specific client segments.

Citron and Manalis [15] examined the choice of auditor in publicly listed Greek firms just after the liberalization of the Greek audit market. The auditors were categorized as big and non-big auditors. They applied Binomial Logistic Regression to develop a model capable of differentiating the cases were companies select a big auditor. According to the results, the level of shareholdings held by foreign shareholders is positively associated with the choice of a big auditor.

Velury et al. [16] conducted a survey to examinee a possible relationship between the choice of a specialized auditor and the level of institutional ownership. The method used was two-stage least square regression. The empirical results suggest that firms with relatively greater levels of institutional ownership tend to employ industry specialist auditors and thus assure higher audit quality.

In a following work, Kane and Velury [17] investigated the relationship between the selection of a big auditor and the level of institutional ownership. The method used was Logistic Regression. The independent variables were some financial ratios and a variable indicating the proportion of shares held by institutional investors. According to this work, firms with high level of institutional ownership tend to choose a big auditing firm. The results also indicate a positive relationship between the selection of a big auditor and the variables Size of the auditee and Debt level.

A critical assessment of the collected literature reveals that the conducted research is restricted to the limits of a typical statistical analysis. No attempt has been ever made to develop models capable of predicting the auditor choice for out-of-the-training-sample cases. Moreover, all the employed methods are statistical. Statistical techniques assume arbitrarily the independence of the input variables. Regression assumes a linear relationship between the dependent (or the log of the dependent) variable and the independent variables. Data mining provides numerous assumption free methodologies. These methodologies have been successfully applied to address other auditing related problems like bankruptcy prediction, fraud detection, or the prediction of qualified auditors' opinions. The results of these studies suggest that these methodologies perform at least equally well as the statistical techniques. However, these techniques have not been applied to address the auditor selection problem.

In an attempt to contribute to this direction, Kirkos et al. [18] employed and compared three data mining methods to develop models capable of predicting the selection of a big or a non-big auditor. The employed methods were Decision Trees, Neural Networks and k-Nearest Neighbor. All the methods were validated against unknown observations. According to [18], the Decision Tree outperformed the other two methods, followed by the Neural Network and the k-NN. These results also suggest that companies having a high debt level tend to choose a brand name auditor. The achieved accuracy rates according to a 10-fold cross validation test were 81.97%, 78.45% and 73.21% for the Decision Tree, Neural Networks and k-NN methods, respectively. In an attempt to increase performances, the authors applied bagging to the three classifiers. According to [18], bagging had a significant effect on the Decision Tree model, by improving its accuracy rate.

This study extends the previously mentioned one. New input variables are tested for their contribution to the auditor selection decision. The methods that achieved the highest performances in the previous study (i.e. C4.5 Decision Tree and Backpropagation Neural Networks) are now compared against the relatively more recent method of Support Vector Machines (SVMs). SVMs enjoy a good reputation for their classification capabilities.

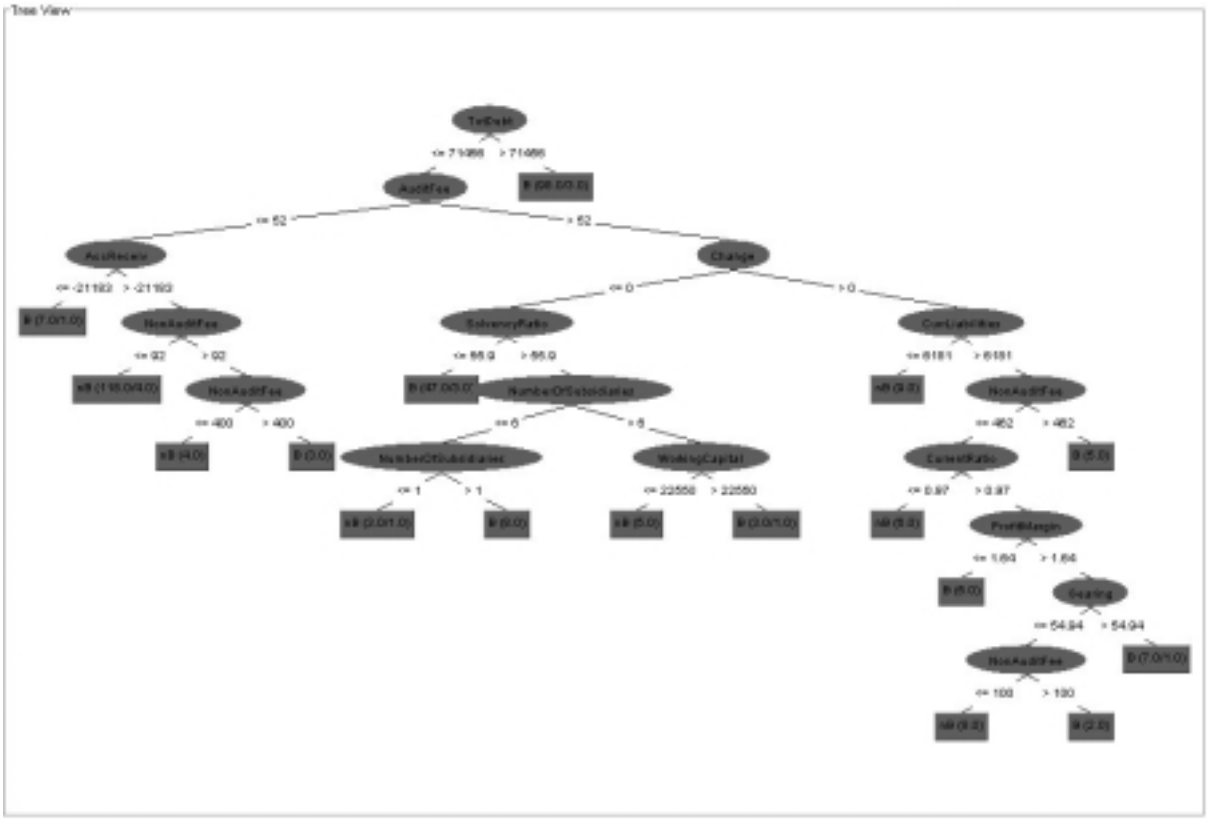

Fig. 1. The Decision Tree model.

3. Research methodology

3.1. Sample construction

The sample used in the present study contains data about 338 UK and Irish firms. The data were drawn from the financial data base FAME (Financial Analysis Made Easy). Initially, we selected the publicly listed companies which selected their auditor during the years 2003-2005. Financial companies were excluded from the sample. Companies with many missing values were also excluded. The remaining companies were matched with equal number of companies that did not select an auditor. The matching criteria were the primary activity of the company and the fiscal year. The final sample contained 181 companies with a big auditor and 157 companies with a non big auditor.

3.2. Variable selection

Here, we review prior relevant research to select the attributes of the companies, which can be related to the auditor choice problem. According to Citron and Manalis [15] companies which choose a big auditor are more profitable. Here, we test the profitability related variables Gross Profit, Operating Profit, Retained Profit, Profit Margin, Return on Shareholders' Funds and Return on Total Assets. Companies with many subsidiaries are more complex and more likely to be audited by a big auditor. We test the variable Number of Subsidiaries.

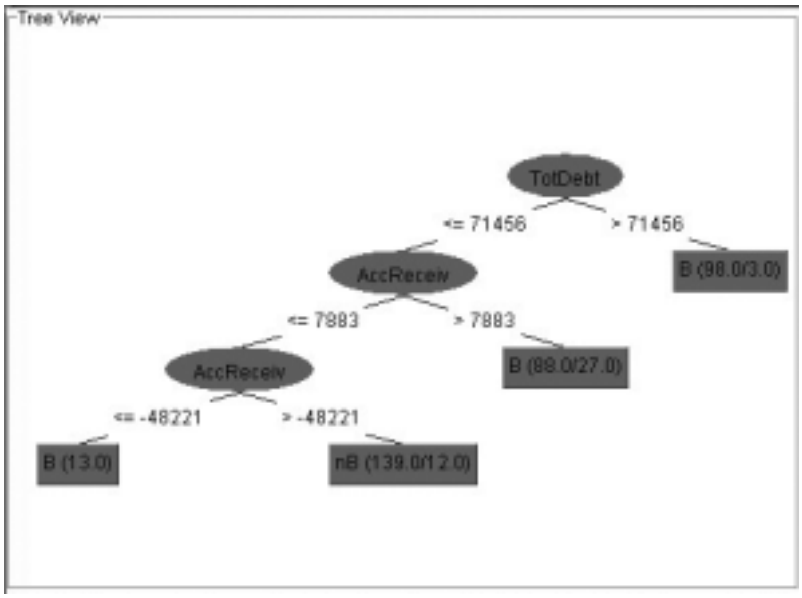

Fig. 2. The Decision Tree model with wrapper feature selection.

Krishnan et al. [19] claims that bigger companies are more likely to hire a big auditing firm. To test the auditee's size we use the variables Turnover, Total Assets, Fixed Assets and Shareholders' Funds. Inventory and accounts receivable are prone to manipulation. Icerman and Hillison [20] suggest that these accounts require audit adjustments. Here, we test the variables Accounts Receivable, Accounts Receivable to Total Assets (ACRTA), Stock & WIP and Inventory to Total Assets (INVTA). Kane and Velury [17] and DeFond [21] associate debt with the selection of auditor. We test the debt related variables Long Term Debt, Total Debt and Gearing.

Chow and Rice [22] associate the size of auditor with the possibility to issue a qualified opinion. Other studies suggest that the receipt of a qualified opinion may lead a company to switch auditor [19,23] and that the direction of the switch is from a larger to a smaller auditor [24]. We use the binary variables Qualifications of Current Year and Qualifications of Previous Year. Moreover, we test if changes in qualifications are related to the selection of Auditor. For two successive years there are four possible combinations i.e. no qualifications for both the previous and the current year, no qualifications for the previous year but qualifications for the current year, qualifications for the previous year but no qualifications for the current year and finally qualifications for both the previous and current year. These combinations are the four possible nominal values of one variable. However, some techniques like Neural Networks can not handle nominal values. For that reason we created four dummy variables. These variables are Unqualified-Unqualified, Unqualified-Qualified, Qualified-Unqualified and Qualified-Qualified. For each observation the variable which depicts the qualifications cases obtains the value 1 where the other three dummy variables obtain the value 0.

Big auditors hire more expensive personnel and invest more in technology and training. This cost translates to higher audit fees. Moreover big auditors are more exposed to litigation risk and, thus, they charge higher fees to compensate for the risk [25]. Here, we test the variables Audit Fees and Audit Fees to Total Assets (AFTA). In UK it is allowed to auditing firms to provide bookkeeping and consultancy services to the audited company. For that reason companies are obligated to disclose the non-audit fees.

Table 1
Descriptive statistics, F values by one way ANOVA

	BIG		NON-BIG			
Variable	Mean	StDev	Mean	StDev	F	p
Turnover	**877229**	**2779232**	**109893**	**894534**	**9.60**	**0.002**
Gross Profit	327944	1479756	31310	234231	4.80	0.029
Operating Profit	**100188**	**390501**	**6241**	**72018**	**8.82**	**0.003**
Retained Profit	28945	131084	2285	37181	6.07	0.014
Audit Fees	**566**	**1362**	**89**	**492**	**17.31**	**0.000**
Non-Audit Fees	**506**	**1237**	**92**	**585**	**12.64**	**0.000**
Fixed Assets	755559	2798495	140425	1546409	5.94	0.015
Current Assets	**359633**	**1022334**	**46337**	**412695**	**12.91**	**0.000**
Current Liabilities	**303331**	**1082573**	**42439**	**415358**	**8.08**	**0.005**
Working Capital	**101180**	**307032**	**16878**	**150894**	**9.41**	**0.002**
Total Assets	**1115192**	**3750319**	**184974**	**1948673**	**7.82**	**0.005**
Long Term Debt	339191	1259099	45161	387991	4.72	0.031
Shareholders Funds	373305	1314165	88565	909180	5.21	0.023
Current Ratio	**1.910**	**1.828**	**3.898**	**6.619**	**15.03**	**0.000**
Liquidity ratio	**1.604**	**1.851**	**3.550**	**6.540**	**14.68**	**0.000**
Solvency Ratio	**42.63**	**27.51**	**54.33**	**35.68**	**11.44**	**0.001**
Gearing	**107.5**	**206.9**	**51.1**	**90.6**	**7.97**	**0.005**
Current Assets Trend	237	2934	76	311	0.45	0.503
Total Assets Trend	352	4394	106	495	0.47	0.494
Current Liabilities Trend	17.2	54.9	49.2	158.7	6.35	0.012
Long Term Liabilities Trend	159.5	1008.0	91.6	638.8	0.40	0.528
Profit Margin	**3.26**	**19.41**	**−6.93**	**26.50**	**14.15**	**0.000**
Return on shareholders Funds	15.9	60.8	−8.5	135.7	4.68	0.031
Return on Total Assets	−3.17	48.26	−18.01	53.38	7.18	0.008
Stock & WIP	85745	297810	21230	156637	3.83	0.051
Market Capitalization	**985**	**2690**	**156**	**1480**	**11.01**	**0.001**
Price / Book value	3.678	4.810	2.931	3.769	2.19	0.140
Number of Subsidiaries	**33.85**	**53.39**	**9.83**	**34.63**	**23.26**	**0.000**
Total Debt	**424392**	**1439138**	**67000**	**478444**	**8.83**	**0.003**
Z Score	12.74	75.11	17.36	63.93	0.36	0.547
Audit Fees to Total Assets (AFTA)	**3.314**	**5.969**	**5.701**	**8.187**	**9.53**	**0.002**
Non-Audit Fees to Total Assets (NAFTA)	5.9196	24.3541	10.2927	24.0037	2.75	0.098
Audit Fees to Total Fees (AFTF)	0.5566	0.2187	0.5740	0.2547	0.4563	0.4998
Accounts Receivable	**273888**	**792423**	**25108**	**291499**	**13.84**	**0.000**
Acc. Receivable to Total Assets	−1.5456	8.6328	−4.3696	13.1253	5.59	0.019
Sales to Total Assets (SALTA)	1.63	4.61	14.47	62.31	7.64	0.006
Quick Ratio	−7.2	43.9	−43.8	171.6	7.65	0.006
Inventory to Total Assets (INVTA)	20.9	86.6	49.3	132.0	5.60	0.019
Working Capital to total Assets (WCTA)	1.74	1.44	6.06	32.34	3.23	0.073

Note: Bold characters indicate the significant input variables $p \leqslant 0.005$.

We check the variables Non-Audit Fees, Non-Audit Fees to Total Assets (NAFTA) and Audit Fees to Total Fees (AFTF) as possible predictors.

Fast growing firms may require additional financing. To attract creditors these firms may hire a brand name auditor [16]. We test several trends related ratios like Total Assets Trend, Current Assets Trend, Current Liabilities Trend and Long Term Liabilities Trend. Kane and Velury [17] use the variable Market Value of Equity to model the selection of auditor. We also use the variable Market Capitalization.

In this study we test also some additional accounts and ratios. The binary variable Change Auditor indicates that the company changed auditor in the examined fiscal year. Other variables used are Current Assets, Current Liabilities, Working Capital, Current Ratio, Liquidity Ratio, Solvency Ratio, Quick

Table 2
The high level splitters of the Decision Tree model

Variable
Total Debt
Audit Fees

Ratio, Sales to Total Assets (SALTA), Working capital to Total Assets (WCTA), Price to Book Value and finally the Altman's Z-score as a proxy for financial distress.

In total we compiled 39 accounts and financial ratios. We ran one way Analysis of Variance to initially estimate the significance of each variable. This method estimates the variation of the values for each class. For each variable a level of significance (the p-value) is calculated. If p-value is small then there are significant differences in the values of that variable according to the class each observation belongs. Table 1 depicts the F-values, the p-values and descriptive statistics for each selected variable. We selected 18 variables with the lowest $p$values ($p \leqslant 0.005$). These variables together with the four dummy variables and the binary variable Auditor Change participate in the final input vector. The rest of the variables were discarded.

Descriptive statistics and the p values provide initial indications concerning features relevant to auditor selection. Companies selecting a big auditor are considerably bigger in terms of Turnover and Total Assets. The mean values of the debt related variables reveal that companies with a high debt structure tend to choose a big auditor. Liquidity seems to be significant since three out of four liquidity related variables were selected to participate in final input vector. Fees matters are also significant. Companies with a big auditor have higher mean values for the variables Audit Fees and non-Audit Fees. However, this can be attributed to their bigger size. The mean values of the variables Audit Fees to Total Assets and non-Audit Fees to Total Assets indicate that companies which choose a non big auditor pay more fees relatively to their size. Trends and financial distress seem to be irrelevant, since all related variables had high p-values and were discarded.

4. Empirical results

After the final sample definition, we applied the three alternative data mining classification methods. The software used to develop all models was WEKA 3.4.11. Initially we developed a Decision Tree model. In particular, the chosen classifier was the J48 Decision Tree. J48 is an implementation of the Quinlan's [26] C4.5 algorithm. The tree was built with confidence factor 0.25. Pruning was enabled. Figure 1 depicts the obtained decision tree.

As can be seen in Fig. 1 the tree uses as first level splitter the variable TotDebt (Total Debt). This means that the algorithm selects this variable as the variable that best separates the two classes. By defining a cut-off value of 71,456,000£ the tree differentiates 98 companies which have a Total Debt higher that the cut-off value. The vast majority (95 out of 98) of these firms have a big auditor. The tree demonstrates that companies with a high debt structure tend to choose a brand name auditor and thus seek after audit quality. As second level splitter the tree selects the variable Audit Fees. Table 2 exhibits the high level splitters of the Decision Tree.

The second applied classification method was a Neural Network. We developed a Multilayer Perceptron Backpropagation network. A well known problem with neural networks is that the empirical definition of several parameters is required. We tested numerous alternative topologies to assure the proper architecture selection. In particular, we tested all the topologies with one to twenty nine hidden nodes

Table 3
The weights of the linear polynomial SVM model

Variable	Weight
Auditor Change	1.9193
Current Ratio	1.0982
Liquidity Ratio	1.0035
Total Debt	−0.9409
Gearing	−0.7412
Number of Subsidiaries	−0.7191
Non Audit Fee	−0.7065
Profit Margin	−0.4393
Audit Fee	−0.3626
Market Capitalization	−0.2441
Audit Fees to Total Assets (AFTA)	−0.1948
Total Assets	0.1558
Current Liabilities	0.1280
Current Assets	−0.0957
Solvency Ratio	0.0734
Operating Profit	−0.0535
UnqualifiedUnqualified (UnqUnq)	−0.0509
Working Capital	0.0492
Quall. Current Year	0.0364
Quall. Previous Year	0.0352
Accounts Receivable	−0.0284
Turnover	−0.0282
QualifiedQualified (QQ)	0.0207
UnqualifiedQualified (UnqQ)	0.0157
QualifiedUnqualified (QUnq)	0.0145

arranged in one hidden layer. For each of these topologies, several alternatives have been tested regarding the learning rate, momentum and training epochs. The best performance, according to a 10-fold cross validation test, was achieved by a network with two hidden nodes, learning rate $= 0.4$, momentum $=$ 0.3, and training epochs $= 500$.

Another well known problem with neural networks is that they act as black boxes, since their decision making process is not understandable by humans. Several methods have been proposed to interpret a neural network model. Unfortunately, WEKA does not provide any methodology to interpret the model and estimate the contribution of each input variable. For that reason, we could not estimate the significance of the input variables by using the backpropagation model.

The third applied classification method was a Support Vector Machine (SVM). We chose the Sequential Minimal Optimization (SMO) classifier provided by WEKA, which implements the Platt's [27] sequential minimal optimization algorithm for training a support vector classifier. Support vector machines require the definition of the kernel function. We tested three alternatives, one linear polynomial function, one quadratic polynomial function and one radial base function. For the case of quadratic polynomial functions we also tested several alternatives for the c variable and for the allowance of lower order terms. According to a 10-fold cross validation test, the best performance was achieved by the quadratic polynomial classifier with $c = 2$ and allowing lower order terms.

To estimate the input variables' contribution for the Support Vector machine classifier we considered the attribute weights of the linear polynomial classifier. Table 3 exhibits the attribute weights provided by WEKA.

As can be seen in Table 3 the SVM classifier recognizes as significant the variables Change Auditor, Current Ratio, Liquidity Ratio, Total Debt and the debt related variable Gearing. Notably both the

Table 4
The significant variables of the linear polynomial SVM model

Variable
Auditor Change
Current Ratio
Liquidity Ratio
Total Debt
Gearing

Table 5
10-fold cross validation performances

Model	Big Auditor %	Non-Big Auditor %	Total %
C4.5	86.19	80.89	83.73
C4.5 (Wrapper)	91.71	80.25	86.39
MLP	73.48	77.71	75.44
SVM (linear)	69.61	72.61	71.01
SVM (quadr.)	82.32	75.80	79.29
SVM (RBF)	71.82	72.61	72.19

decision tree and the SVM classifiers agree that the variable Total Debt is significant in auditor selection. This conclusion complies with the findings of Kane and Velury [17] and DeFond [21].

Table 4 summarizes the significant variables according to the SVM classifier.

4.1. Models' validation

As stated in the prior research section, up to now most of the researchers apply methods to detect significant factors associated with the auditor selection problem. Little effort has been directed towards a model evaluation in terms of its ability to predict the auditor choice for unknown observations. Estimating the performance of a model against the training sample may introduce a bias. In many cases the models tend to memorize the sample instead of learning. This phenomenon is called data over-fitting. The effect of data over-fitting is that the model describes in detail the sample, but is unable to satisfactory predict the class of out-of-the-sample observations.

In this study, we validate our models against previously unseen patterns. Data mining framework provides several validation methodologies. The simplest approach is to split the sample in two subsets, one used for training and another used for validation. An alternative, more elaborated approach is the 10-fold cross validation, according to which the sample is divided in ten equal randomly selected folds. Nine folds are used to train the model and the tenth fold is used for validation. This process iterates ten times, each time using a different fold for validation purposes. Finally, the average performance is calculated. This way the whole sample is actually used for validation. The 10-fold cross validation approach introduces lower bias and variance. For that reason this approach is the proposed one in the literature [28]. We applied the 10-fold cross validation method to the three models. Table 5 summarizes the models' performances.

As can be seen in Table 5 the highest classification accuracy is achieved by the Decision Tree model. The DT model predicts correctly 86.19% of the big auditor cases and 80.89% of the non big auditor cases. The total performance of the model is 83.73%. The Support Vector Machine model with the quadratic polynomial kernel function follows, with total accuracy rate 79.29%. Accuracies per class for this model are 82.32% and 75.80% for the big auditor and non big auditor classes respectively. The

Neural Network model comes third with average accuracy rate of 75.44% followed by the SVM model with the radial base kernel function and the SVM model with the linear polynomial kernel function.

In the literature prevails the allegation that methods like Support Vector Machines and Artificial Neural Networks tend to perform better when dealing with continuous features, where logic-based systems (e.g. decision trees) tend to perform better when dealing with discrete/categorical features (Kotsiantis et al. [29]). Although this allegation is generally valid, it is not an inviolable rule. Several studies which use data sets with continuous features report better results for the Decision Trees. Koh [30] employed Decision Trees, Neural Networks and Logistic Regression to predict Going Concern qualifications. The input vector contained six financial ratios. According to the results the Decision Tree model outperformed the other two models. Doumpos and Zopounidis [31] tried to assess Country Risk by employing six methods. Among these methods were Decision Trees and Neural Networks. They created two alternative input vectors, both containing financial ratios. They calculated the classification accuracy for five different years. In six out of ten tests Decision Trees have been found to prevail over Neural Networks. Kotsiantis et al. [10] in a study aiming to detect Fraudulent Financial Statements, compared several methods including RBF Neural Networks, Support Vector Machines and Decision Trees. The input vector contained financial ratios. The authors report better results for the Decision Tree Model. Michie et al. [32] in an exhaustive study compared 22 classifiers over numerous data sets. The data sets contained Credit Risk data, Image Related data, Medical and Biological data, Industrial Data etc. Breiman [33] derived from the Michie's study the average rank of the classifiers. Decision Trees have been found to perform better that Neural Networks (Average Rank of C4.5 DTs = 9.3, Average Rank of NNs = 12.3).

4.2. Wrapper feature selection for C4.5

Since the Decision Tree succeeded the best accuracy rate we wished to improve further its performance. One technique for improving the performance is to employ a wrapper approach for feature selection. According to wrapper feature selection the same technique which is used for classification is also used for the determination of the attributes that will form the input vector [28]. This methodology could also reduce the size of the produced decision tree and thus increased comprehensibility could be achieved.

Weka provides a Wrapper Subset evaluator. We employed the evaluator in combination with the classifier J48 with five folds. The Wrapper evaluator selected two attributes i.e. Total Debt and Accounts Receivable. We applied the J48 classifier to the reduced data set which had two input variables. The tree managed to classify 91.71% of the Big Auditor cases and 80.25% of the non-Big Auditor cases. The overall accuracy was 86.39%, which is considerably higher than the accuracy of J48 without wrapper feature selection.

Figure 2 depicts the new Decision Tree. As can be seen, the variable Total Debt is again used as first level splitter.

5. Conclusions

Managers play still an important role in the decision of hiring or firing an auditor. The fear that the auditor may lose the customer may motivate the auditor to report favorably, thus jeopardizing the audit quality. Brand name auditors are recognized in the literature as providers of higher quality auditing services. Prior research efforts addressing the auditor selection problem perform mainly typical statistical analysis. Data mining classification methodologies have not yet been applied to develop models capable

of predicting the type of the hired auditor and to reveal factors that influence the auditor selection decision.

In the present study we employed and compared three classification methods to predict the choice of an auditor. These methods are the C4.5 Decision Trees, the Backpropagation Neural Networks and the Support Vector Machines. All these methods are known for their classification capabilities. The sample we used contains data about 338 UK and Irish firms. The input vector contains financial ratios and account values, as well as qualitative variables indicating the qualification cases and the auditor change.

According to a 10-fold cross validation evaluation, the Decision Tree model outperformed the remaining methods, achieving an average accuracy rate of 83.73%. The performances of the quadratic polynomial Support Vector Machine and the Neural Network were 79.29% and 75.44% respectively. Finally, the linear polynomial and RBF Support Vector Machines models came behind in terms of performance. In an attempt to increase the Decision Tree's performance we applied wrapper feature selection. This method improved the performance of the Decision Tree.

Two methods provided insights to the significance of the input variables. The Decision Tree model and the Support Vector Machine model agree that the level of debt is a significant factor that influences the selection of an auditor. Companies with a high debt structure tend to choose a brand name auditor.

The results were encouraging in that we have developed reliable models capable of predicting the auditor choice. However, our findings can stimulate additional research. The auditors' categorization can be further analyzed as international auditors, national reputation auditors and local auditors. Such a particularization could add to the understanding of the factors that influence the auditor selection decision. The auditor choice is also influenced by managerial issues. Enriching the input vector with variables related to managerial characteristics could further improve the models' performances.

Acknowledgments

We would like to thank two anonymous reviewers for their valuable comments and suggestions that resulted in significant improvements in the current version.

References

[1] L. DeAngelo, Auditor size and auditor quality, *Journal of Accounting and Economics* **1** (1981), 113–127.
[2] J. Mutchler, Empirical evidence regarding the auditor's going-concern opinion decision, *Auditing: a Journal of Practice and Theory* **6** (1986), 148–163.
[3] Z. Palmrose, An analysis of auditors litigation and audit service quality, *The Accounting Review* **63** (1988), 55–73.
[4] E. Bartov, F.A. Gul and J.S.L. Tsui, Discretionary-accruals models and audit qualifications, *Journal of Accounting and Economics* **30** (2001), 421–452.
[5] A. Craswell, D.J. Stokes and J. Laughton, Auditor independence and fee dependence, *Journal of Accounting and Economics* **33** (2002), 253–275.
[6] J.R. Francis and E.R. Wilson, Auditor changes: a joint test of theories relating to agency costs and auditor determination, *The Accounting Review* **63**(4) (1988), 663–682.
[7] D.R. Franz, D. Crawford and E.N. Johnson, The impact of litigation against an audit firm on the market value of nonlitigating clients, *Journal of Accounting, Auditing & Finance* **13**(2) (1998), 117–134.
[8] M.J. Kim and I. Han, The discovery of experts' decision rules from qualitative bankruptcy data using genetic algorithms, *Expert Systems with Applications* **15**(4) (2003), 637–646.
[9] S. Kotsiantis, D. Tzelepis, E. Koumanakos and V. Tampakas, Selective costing voting for bankruptcy prediction, *International Journal of Knowledge-Based & Intelligent Engineering Systems* (*KES*) **11**(2) (2007), 115–127.
[10] S. Kotsiantis, E. Koumanakos, D. Tzelepis and V. Tampakas, Forecasting fraudulent financial statements using data mining, *International Journal of Computational Intelligence* **3**(2) (2006), 104–110.

[11] E. Kirkos, C. Spathis and Y. Manolopoulos, Data mining techniques for the detection of fraudulent financial statements, *Expert Systems with Applications* **32** (2007), 995–1003.
[12] C. Gaganis, F. Pasiouras and M. Doumpos, Probabilistic neural networks for the identification of qualified audit opinions, *Expert Systems with Applications* **32**(1) (2007), 114–124.
[13] K.S. Cravens, J.C. Flagg and H.D. Glover, A comparison of client characteristics by auditor attributes: implications for the auditor selection process, *Managerial Auditing Journal* **9**(3) (1994), 27–36.
[14] P. Chaney, D. Jeter and L. Shivakumar, Self selection of auditors and audit pricing in private firms, *The Accounting Review* **79** (1996), 51–72.
[15] D. Citron and G. Manalis, The international firms as new entrants to the statutory audit market: an empirical analysis of auditor selection in Greece, 1993 to 1997, *The European Accounting Review* **10** (2001), 439–459.
[16] U. Velury, J. Reish and D. O'Reilly, Institutional ownership and the selection of industry specialist auditors, *Review of Qualitative Finance and Accounting* **21** (2003), 35–48.
[17] G. Kane and U. Velury, The role of institutional ownership in the market for auditing services: an empirical investigation, *Journal of Business Research* **57** (2004), 976–983.
[18] E. Kirkos, C. Spathis and Y. Manolopoulos, Applying data mining methodologies for auditor selection, *Proceedings 11th Pan-Hellenic Conference in Informatics* (*PCI*), Patras, Greece, 2007, pp. 165–178.
[19] J. Krishnan, J. Krishnan and R. Stephens, The simultaneous relation between auditor switching and audit opinion: an empirical analysis, *Accounting and Business Research* **26** (1996), 224–236.
[20] R. Icerman and W. Hillison, Disposition of auditor–detected errors: some evidence on evaluative materiality, *Auditing: a Journal of Practice and Theory* **10** (1991), 22–34.
[21] M. DeFond, The association between changes in client firm agency costs and auditor switching, *Auditing: a Journal of Practice and Theory* **11** (1992), 16–31.
[22] C. Chow and S. Rice, Note: Qualified audit opinions and auditor switching, *The Accounting Review* **37** (1982), 326–335.
[23] D. Citron and R. Taffler, The audit report under going concern uncertainties: an empirical analysis, *Accounting and Business Research* **22** (1992), 337–345.
[24] B. Johnson and T. Lys, The market of auditor services – Evidence from voluntary auditor changes, *Journal of Accounting and Economics* **12** (1990), 281–308.
[25] M. Firth, Auditor-provided consultancy services and their associations with audit fees and audit opinions, *Journal of Business Finance and Accounting* **29** (2002), 661–693.
[26] R. Quinlan, *C4.5: Programs for machine learning*, Morgan Kaufmann Publishers, San Mateo, CA, 1993.
[27] J. Platt, Fast training of support vector machines using sequential minimal optimization, in: *Advances in Kernel Methods – Support Vector Learning*, B. Schoelkopf, C. Burges and A. Smola, eds, MIT Press, Cambridge, MA, 1999.
[28] J. Han and M. Camber, *Data mining concepts and techniques*, Morgan Kaufman, San Diego, 2000.
[29] S. Kotsiantis, D. Kanellopoulos and V. Tampakas, On Implementing a Financial Decision Support System, *International Journal of Computer Science and Network Security* **6** (2006), 103–112.
[30] H.C. Koh, Going concern predictions using data mining techniques, *Managerial Auditing Journal* **19** (2004), 462–476.
[31] M. Doumpos and C. Zopounidis, On the Use of a Multi-criteria Hierarchical Discrimination Approach for Country Risk Assessment, *Journal of Multi-criteria Decision Analysis* **11** (2002), 279–289.
[32] D. Michie, D.J. Spiegelhalter and C.C. Taylor, *Machine Learning, Neural and Statistical Classification*, Ellis Horwood Limited, 1994.
[33] L. Breiman, Bagging Predictors, *Machine Learning* **24** (1996), 123–140.