Applied Artificial Intelligence
An International Journal

ISSN: 0883-9514 (Print) 1087-6545 (Online) Journal homepage: https://www.tandfonline.com/loi/uaai20

Optimizing Fraudulent Firm Prediction Using Ensemble Machine Learning: A Case Study of an External Audit

Nishtha Hooda, Seema Bawa & Prashant Singh Rana

To cite this article: Nishtha Hooda, Seema Bawa & Prashant Singh Rana (2020) Optimizing Fraudulent Firm Prediction Using Ensemble Machine Learning: A Case Study of an External Audit, Applied Artificial Intelligence, 34:1, 20-30, DOI: 10.1080/08839514.2019.1680182

To link to this article: https://doi.org/10.1080/08839514.2019.1680182

APPLIED ARTIFICIAL INTELLIGENCE
2020, VOL. 34, NO. 1, 20–30
https://doi.org/10.1080/08839514.2019.1680182

Taylor & Francis
Taylor & Francis Group

Optimizing Fraudulent Firm Prediction Using Ensemble Machine Learning: A Case Study of an External Audit

Nishtha Hooda[a], Seema Bawa[b], and Prashant Singh Rana[b]

[a]Computer Science and Engineering Department, Chandigarh University, Mohali, Punjab, India; [b]Computer Science and Engineering Department, Thapar Institute of Engineering and Technology, Patiala, Punjab, India

ABSTRACT

This paper is a case study of utilizing machine learning for developing a decision-making system for auditors before initializing the audit fieldwork of public firms. Annual data of 777 firms from 14 different sectors are collected and a MCTOPE (Multi criteria ToPsis based Ensemble) framework is implemented to build an ensemble classifier. MCTOPE framework optimizes the performance of classification during ensemble building using the TOPSIS multi-criteria decision-making algorithm. Ensemble machine learning is used for optimizing the prediction performance of suspicious firm predictor in the previous work available at https://www.tandfonline.com/doi/full/10.1080/08839514.2018.1451032. After achieving an accuracy of 94.6% and AUC (area under the curve) value of 0.98, this ensemble classifier is employed in a web application developed for auditors using Python and R script for the prediction of suspicious firm before planning an external audit. The performance of an ensemble classifier is validated using K-fold cross validation technique and is found to be better than the state-of-the-art classifiers.

Introduction

Fraud is a critical issue worldwide. Firms that resort to the unfair practices without the fear of legal repercussion have a grievous consequences for the economy and individuals in the society. Auditing practices are responsible for the fraud detection. Audit is defined as the process of examining the financial records of any business to corroborate that their financial statements are in compliance with the standard accounting laws and principles (Cosserat 2009). Data analytics tools for an effective fraud management have become the need of the hour for an audit. The possibilities that how data analytics can improve the quality of process is published in Emerging Assurance Technologies Task Force of the AICPA Assurance Services Executive Committee (ASEC) (AICPA Staff 2014).

When the audits are performed by any external audit company, the risk assessment plays a vital role in deciding the amount of fieldwork that would

CONTACT Nishtha Hooda 27nishtha@gmail.com Thapar Institute of Engineering and Technology, Patiala, India
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/uaai.

be required before actually visiting the official firms. The complete process of risk assessment during audit is explained in detail in the previous work (Hooda 2018). The prime goal of an auditor during an audit-planning phase is to follow a proper analytical procedure to impartially and appropriately identify the firms that resort to a high risk of unfair practices. Identifying fraudulent firms can be studied as a classification problem. The purpose of classifying the firms during the preliminary stage of an audit is to maximize the field-testing work of high-risk firms that warrant significant investigation.

Many researchers have employed algorithms like artificial neural network, logistic regression, decision trees, and bayesian belief networks for detecting management fraud in the financial statements (Fanning 1998; Green 1997; Spathis 2002). Ensemble machine learning method is also applied successfully for improving the classification accuracies of the auditing task (Kotsiantis 2006). Machine learning algorithms like support vector machine, logistic regression, probabilistic neural network, genetic algorithm, etc. are also combined with feature selection methods in order to prove their usability in detecting fraud in the Chinese firms (Ravisankar 2011). During audit-planning, auditors examine the business of different government offices but target to visit the offices with very-high likelihood and significance of misstatements.

As stated by Wolpert, there is no single best algorithm which is applicable for all the possible cases of problems (Wolpert and Macready 1997). Furthermore, a lot of research efforts have been made for improving the performance of machine learning models by developing an ensemble-classifier which is constructed from diverse machine learning models (Dietterich 2000; Qi 2012). From a practical point of view, multiple-opinions are unfailingly better than a single opinion in any decision-making process. Ensemble learning serves as a powerful tool in machine learning as it employs multiple classifiers and works on optimizing the performance of base classifiers separately. Although it cannot always guarantees a success, but generally, it reduces variance and offers better performance than a single classifier solution (Dietterich 2000; Qi 2012; Zhang and Ma 2012). By choosing a specific aggregation technique like majority voting, boosting, bagging, etc., an ensemble classifier aids to scrutinize the risk of obtaining poor results from a single classifier system.

This research work is a case study of an external government audit company which is also an external auditor of government firms of India. Complex audit data is collected and an ensemble classifier is built using MCTOPE (Multi criteria ToPsis based Ensemble) framework, which implements the Technique for Order of preferences by similarity to Ideal Solution (TOPSIS) Multi-Criteria assessment algorithm (Majid 2012). The performance of the built ensemble is tested using K-fold cross validation technique and is also compared with other state-of-the-art methods. This ensemble is

employed in a *Fraudulent Firm Prediction* Web Application (built in Django Python Framework) for predicting the high-risk firms.

The goal of the research is to help the auditors by building a classification model that can predict the fraudulent firm on the basis of the present and historical risk factors. The validity of proposed framework is tested using the K-fold cross validation method, and then the proposed method is applied to suspicious firm prediction problem.

The rest of this paper is structured as follows. Section 2 describes the audit data, features, and experimental setup. The proposed framework and ensemble model building technique are presented in Section 3. Performance analysis, results and analysis are discussed in Section 4. Finally, Section 5 concludes this paper and points out the scope of further research.

Data and Experimental Setup

Data Collection

Exhaustive 1 year and 6 months non-confidential data firms is collected from the Audit General Office (AGO) of India. There are total 776 firms from 46 different cities of a state that are listed by the auditors for targeting the next field-audit work. The target-offices are listed from 14 different sectors. Detailed description of data and features is presented in the previous paper (Hooda 2018)

Experiment Setting

Ten state-of-the-art classification methods namely decision tree (DT) (Quinlan 1986), adaboost (AB)(Schapire 1999), random forest (RF) (Liaw et al. 2002), support vector machine (SVM) (Keerthi et al. 0000), probit linear model (PLM) (Chambers 1977), neural network (NN)(Russell 2003), decision stump (DSM)(Iba et al. 1992), J48 (Ross Quinlan 1996), Naive Bayes (NB) (Rish 2001), and Bayesian (BN) are employed to make the pool of classifiers called model-pool for ensemble building. The R 'caret' package is used to implement the various classification models. The models are available in R open-source software. R is licensed under GNU GPL. The complete description of model's parameter setting and required packages is summarized in the previous work (Hooda 2018). The purpose is to measure the prediction performance of the model when it is up- and running and then predicting the suspicious firm class of the new samples without the benefit of knowing the true risk-class of the samples.

Prediction Model

The prediction engine automates the process of ensemble building. The complete process is presented in Figure 1.

In the first step, the data is prepared using data preparator. Data preparation works on cleaning, feature extraction and feature selection process. In the second step, random samples are generated to train the randomly selected classifiers in the model pool. The decision file is generated to collect the predictions of the randomly selected classifiers from the model-pool. The sample decision file is presented in Table 1. The sample decision file contains the predictions of six selected classifiers, chosen randomly from the model pool. D represents the ensemble decision by majority voting. In the third step, initially, a preliminary-ensemble is generated by combining m($1 \leq m \leq 10$) different machine learning model's decision by random using the majority voting aggregation technique. TOPSIS multi-criteria decision score of preliminary-ensemble is evaluated using six different performance measures. In the next iteration, a new ensemble is produced using different training samples, and with a new set of models in model-pool. The performance of

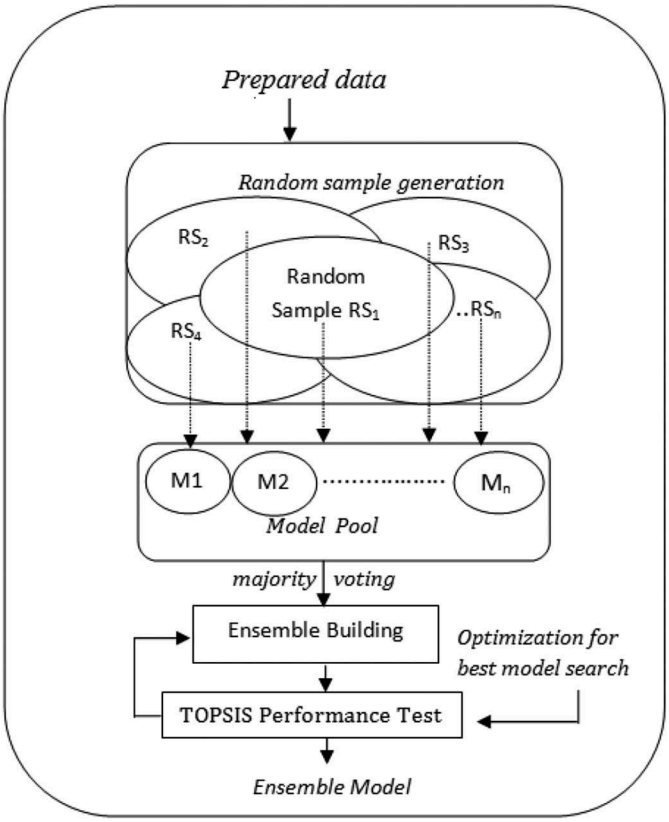


Figure 1. Architecture of prediction engine in MCTOPE Architecture; M: Machine learning model; RS: Random sample of training dataset.

this ensemble is compared with the preliminary ensemble, and the model with the higher TOPSIS score is saved. This phase is called *optimization for the best model search*. At the end *n* iterations (say n = 5000 for small dataset), a final ensemble with the highest performance score is declared as the Winning-Ensemble.

Topsis Performance Evaluation

Accuracy of the classifier is a popular approach to compare the level of competence among several classifiers (Bradley 1997). Using a single evaluation metric for comparing the performance of machine learning model is error-prone approach. In this work, the proposed ensemble is no longer evaluated only the accuracy of classification, which is quite different from the available techniques. For comprehensive evaluation of ensemble model performance, multiple performance metrics should be considered. Among the several solutions available, Multi-Criteria Decision making is the most prevalent approach. Technique for Order Preference by Similarity to an Ideal Solution (TOPSIS) is a popular multi-criteria analysis technique (Triantaphyllou 2000). TOPSIS score does not only calculates the accuracy but also focuses on sensitivity, specificity, and area under the ROC curve (AUC) of the classifier. Choosing the evaluation-criteria that suits the goal of improving the classification performance is an important step of this process. The proposed framework considers multi-criteria performance metrics of confusion matrix (Table 2) presented in Table 3 for testing the performance of suspicious firm classification.

A TOPEES assessment score function (Triantaphyllou 2000) is introduced for comprehensive performance assessment. It works on the principle of finding an alternative closest to the ideal solution and farthest from the negative-ideal solution. Ideal solution i is the set of evaluation-criteria solution with maximum benefit and can be described as

$$i = \{max(accuracy), min(error), max(sensitivity), max(specificity), max(MCC), max(Fscore), max(AUC)\} \quad (1)$$

Table 1. Decision file sample.

Iteration	Machine learning models						Decision
	RF	AB	PLM	SVM	NN	NB	D
i	f	t	t	f	t	t	t

Table 2. Confusion matrix.

	True Reference	
Predicted Condition	**Suspicious**	**Non Suspicious**
Suspicious	True positive X	False Negative Z
Non Suspicious	False Positive Q	True Negative Y

Table 3. Performance evaluation metrics.

Performance Metric	Formula
Type-I error	Q
Type-II error	Z
Sensitivity	$X/(X+Z)$
Specificity	$Y/(Q+Y)$
Accuracy	$(X+Y)/(X+Z+Q+Y)$
F Score	$(2*X)/(2*X)+(Q+z)$
MCC	$(X*Y)-(Q*Z)/SQRT((X+Q)+(X+Z)+(Y+Q)+(Y+Z))$

Negative ideal solution j is the solution with the maximum loss and can be described as

$$j = \{min(accuracy), max(error), min(sensitivity), min(specificity), min(MCC), min(Fscore), min(AUC)\} \quad (2)$$

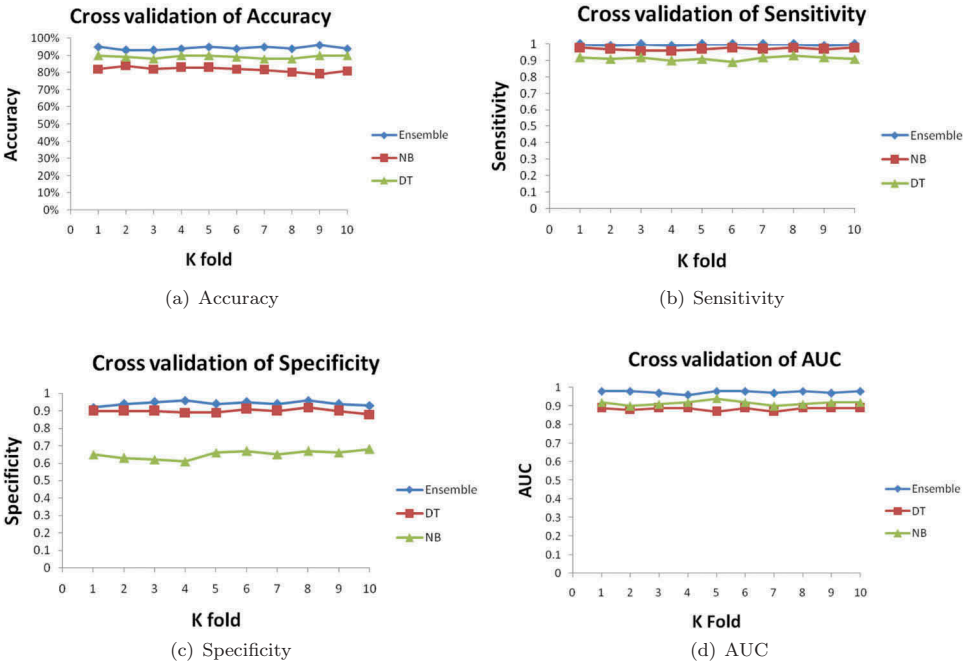
Experimental Results and Discussion

MCTOPE framework builds an ensemble of naive bayes and decision tree classifiers with the highest performance score for the collected audit data. For testing, the experiments are designed to use 10-fold cross validation method. The data set is divided into 10 equal size subsets. Ensemble algorithm is built to train the nine subset folds and testing on the last subset fold. To test the robustness of designed ensemble classifier, the process is iterated. To evaluate the performance of the proposed framework, different performance parameters namely accuracy, sensitivity, specificity, F measure, MCC and Area under curve (AUC) are used.

The performance of the built ensemble using Naive Bayes (NB) and Decision Tree (DT) as base classifiers is graphically depicted on each fold of 10-cross validation method in Figure 2. It can be observed that the accuracy, sensitivity, and specificity of the built ensemble is better than the base classifiers and it is quite robust as value of accuracy is not changing abruptly. It can be observed that the AUC of the ensemble is better than the base classifiers and it is quite robust as values are stable.

Comparison with State-of-the-Art Methods

In order to perform the comparison analysis, the state-of-the-art classifiers are explored in the area of auditing and finance. The results of the proposed framework is compared with the outstanding classical classifiers like support vector machine (SVM), random forest, neural network, C4.5, adaboost, naive bayes, etc. The results are presented in Table 4. The best values of the performance metrics are highlighted. The performance of an ensemble is quite better than the other classifiers. If the accuracy parameter is compared, performance of an ensemble classifier is closer to the random forest and



(a) Accuracy (b) Sensitivity

(c) Specificity (d) AUC

Figure 2. Accuracy, sensitivity, specificity and AUC results using K fold cross validation on audit testing dataset using MCTOPE ensemble classifier.

Table 4. Comparison the performance of state-of-the-art classifiers with built ensemble classifier.

Metric	Ensemble	SVM	Rf	J48	NN	Aboost	NB	BN	DT	PLM	DS
Accuracy(%)	**94.6**	65.63	93	92	79.15	92.65	82	91	90	92	87.68
Sensitivity	**1.00**	0.50	0.95	0.97	0.89	0.96	0.99	0.95	0.92	0.95	0.98
Specificity	**0.92**	0.81	0.91	0.89	0.63	0.89	0.65	0.87	0.90	0.88	0.75
F measure	**0.94**	0.64	0.93	0.93	0.79	0.92	0.81	0.91	0.91	0.92	0.87
MCC	**0.87**	0.32	**0.87**	0.86	0.59	0.85	0.69	0.84	0.83	0.84	0.77
AUC	**0.98**	0.65	0.96	0.93	0.86	0.93	0.92	0.95	0.89	0.93	0.86

adaboost classifiers. This can be due to the fact that random forest and adaboost models are also ensemble-based classifiers. Sensitivity of the built ensemble is closer to the naive bayes classifier. Sensitivity of the built ensemble is closer to the decision tree classifier. If F metric is compared, values of the built ensemble is closer to the random forest classifier. MCC value of built ensemble is lower than random forest, J48, adaboost, bayesian network, decision tree, and PLM classifiers. The AUC value of the built ensemble is the best and quite closer to 1. To check the overall performance of the built ensemble, multi-criteria based TOPSIS performance scores can be compared.

For a comprehensive performance check, TOPSIS scores of the base classifiers are compared in Figure 3. It is clear in the figure that the overall

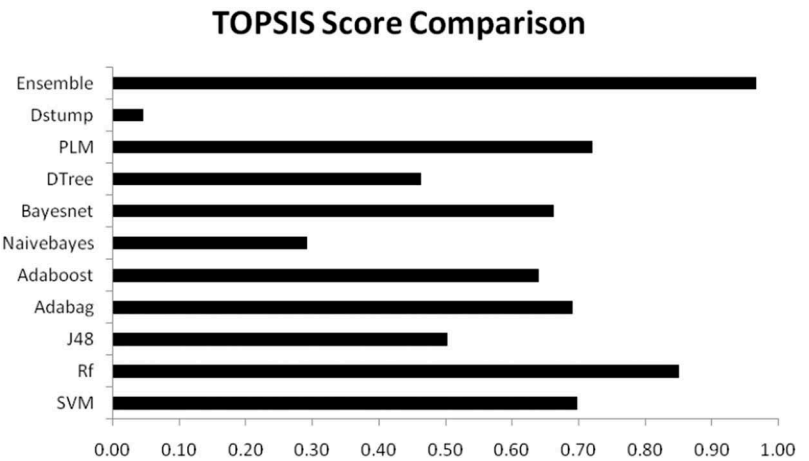


Figure 3. TOPSIS performance comparison analysis.

performance of output ensemble is exceptionally good. It demonstrates that the MCTOPE framework builds an ensemble classifier, which is better than the other base classifiers on the same audit sample dataset.

Implementation

An ensemble classifier with the highest performance score is implemented as a Fraudulent Firm Web Application using Python Django Web framework. Figure 4 presents the main page of the web application. A user can submit the values of the features of any firm to classify the firm as Suspicious or Non Suspicious. In order to predict a new firm for next year fraud risk, this web-application takes input of the important features shown in Figure 4 and predicts the probability of risk using ensemble model, working in the back-end.

Conclusion

In this paper, a case study of an external audit company is studied and an ensemble classifier-based web application is built to help in the decision-making process of predicting the suspicious firm before an external audit. Different from the traditional approach of classification, a MCTOPE framework builds an efficient ensemble classifier by optimizing the overall performance of the classifier using TOPSIS algorithm, a multi-criteria decision-making technique. After more than 1000 iterations, the performance of the final ensemble using Naive Bayes and Decision Tree as base classifiers with the highest TOPSIS performance score is depicted on each fold of the 10-fold cross validation method. The classifier is found to be robust with an accuracy, sensitivity, specificity,

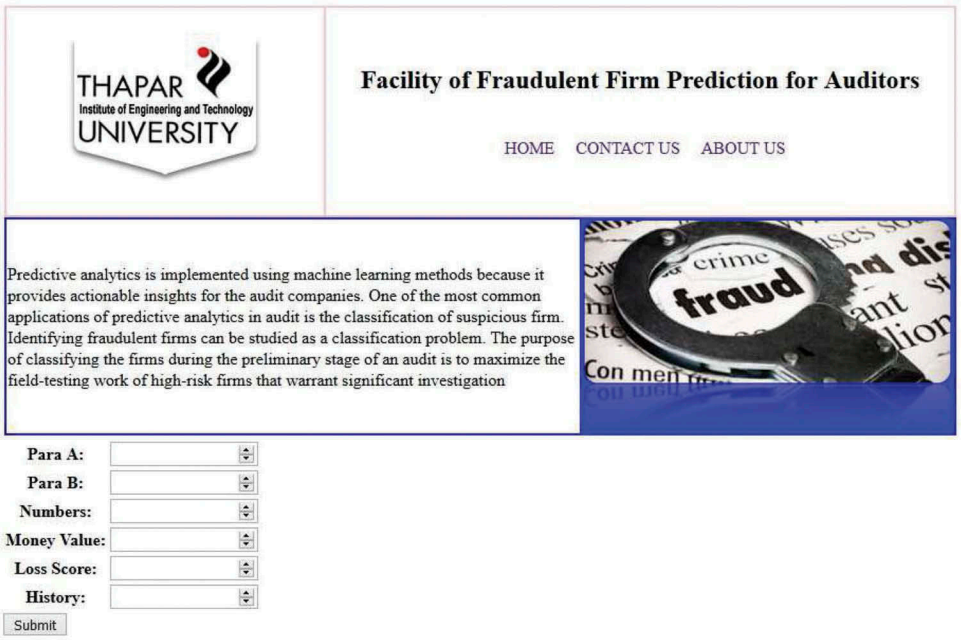


Figure 4. Fraudulent firm web application.

F measure, MCC, AUC of 94, 1, 0.92, 0.94, 0.83, and 0.98, respectively. When compared with the state-of-the-art classifiers, it is found to be better than the available methods, serving as a proof of eligibility of classifiers to perform an efficient fraudulent firm prediction in the audit fieldwork decision-making process.

For future works, we are targeting to offer the auditors to handle the last 10-year data of firms on the top of advance big data techniques like Hadoop, Spark, etc. This research work is supported by Ministry of Electronics and Information Technology (MEITY), Govt. of India (Grant No. DoRSP/1633). The authors wish to thank the auditors of audit office for their assistance, time, and continued support. The authors are grateful for their helpful feedbacks and comments on early version of this work.

Acknowledgments

The authors wish to thank the auditors of audit office for their assistance, time, and continued support. The authors are grateful for their helpful feedbacks and comments on the early version of this work.

Funding

This research work is supported by Ministry of Electronics and Information Technology (MEITY), Govt. of India (Grant No. DoRSP/1633).

References

AICPA Staff. 2014. *Reimagining auditing in a wired world*. USA: University of Zurich, Department of Informatics.

Bradley, A. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* Elsevier. 30 (7):1145–59. doi:10.1016/S0031-3203(96)00142-2.

Chambers, J. M. 1977. *Computational methods for data analysis*, 152–89. New York: Wiley.

Cosserat, G. 2009. *Accepting the engagement and planning the audit. Modern auditing*, 73436. John Wiley Sons, Kingston University, the University of Technology, Sydney.

Dietterich, T. G. 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier systems 2000 Jun 21 (pp. 1-15)*. Springer, Berlin, Heidelberg.

Fanning, K. 1998. Neural network detection of management fraud using published financial data. *International Journal of Intelligent Systems in Accounting, Finance & Management* 7 (1):21–41. John Wiley & Sons, Ltd. doi:10.1002/(SICI)1099-1174(199803)7:1<21::AID-ISAF138>3.0.CO;2-K.

Green, B. 1997. Assessing the risk of management fraud through neural network technology. *Auditing* 16 (1):14. American Accounting Association.

Hooda, N. 2018. Fraudulent firm classification: A case study of an external audit. *Applied Artificial Intelligence* 32 (1):49–51. doi:10.1080/08839514.2018.1451032.

Iba, W. and Langley, P. 1992. Induction of one-level decision trees, Proceedings of the ninth international conference on machine learning, Aberdeen, Scotland, United Kingdom, 233–40

Keerthi, S. S and Gilbert E. G. 2002. Convergence of a generalized SMO algorithm for SVM classifier design. *Machine Learning* Springer. 46(1–3):351–60. doi:10.1023/A:1012431217818.

Kotsiantis. 2006. Forecasting fraudulent financial statements using data mining. *International Journal of Computational Intelligence* 3 (2):104–10.

Liaw, A., et al. 2002. Classification and regression by randomForest. *R News* 2 (3):18–22.

Majid, B. 2012. A state-of the-art survey of TOPSIS applications. *Expert Systems with Applications* 39 (17):13051–69. doi:10.1016/j.eswa.2012.05.056.

Qi, Y. 2012. Random Forest for Bioinformatics. *Ensemble Machine Learning Methods and Applications* 1:307–23.

Quinlan, J. R. 1986. Induction of decision trees. *Machine Learning* 1 (1):81–106. doi:10.1007/BF00116251.

Ravisankar. 2011. Detection of financial statement fraud and feature selection using data mining techniques. *Decision Support Systems* 50 (2):491–500. doi:10.1016/j.dss.2010.11.006.

Rish, I. 2001. An empirical study of the naive Bayes classifier. *IJCAI 2001 Workshop on Empirical Methods in Artificial intelligence,IBM New York* 3 (22):41–46.

Ross Quinlan, J. 1996. Improved use of continuous attributes in C4. 5. *Journal of Artificial Intelligence Research* 4:77–90. doi:10.1613/jair.279.

Russell, S. 2003. *Artificial intelligence: A modern approach*, 134–81. Upper Saddle River: Prentice hall.

Schapire. 1999. A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence* 14 (771–780):1612.

Spathis, C. T. 2002. Detecting false financial statements using published data: Some evidence from Greece. *Managerial Auditing Journal* 17 (4):179–91. doi:10.1108/02686900210424321.

Triantaphyllou, E. 2000. Multi-criteria decision making methods. In *Multi-criteria decision making methods: A comparative study 2000 (pp. 5-21)*. Springer, Boston, MA.

Wolpert, D. H., and W. G. Macready. 1997. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation* 1 (1):67–82. doi:10.1109/4235.585893.

Zhang, C., and Y. Ma. 2012. Ensemble machine learning methods and applications. In *Ensemble learning*, 11–34. China: Springer Publishing Company.