A Framework for Information Mining from Audit Data

Aikaterini Ioannou, Dimitrios Bourlis, Stavros Valsamidis, and Athanasios Mandilas

Abstract Companies seek new technologies to enhance their business processes. As information systems in companies become more complex, the traditional audit trail is diminished or eliminated. The importance of audit automation and the utilization of IT in modern audits has grown significantly in recent years due to both technological developments and changing regulatory environment. Automation of business processes has inevitably led to changes in auditing procedures and standards. Additional drivers of audit automation adoption include the ever growing complexity of business transactions and increasing risk exposure of modern enterprises. Therefore, the audit's purpose, which is namely to examine the true and fair view of financial statements, is heavily increasing in complexity. On the other hand, the prevalence of the data paradigm has manifold impacts on the accounting-relevant processes. To cover the requirements to Audit Information System, we strive for the development of a framework for information mining from audit data. In this paper, we report on the framework we have developed in the department of Accounting and Finance. Our study identifies the management of audit alarms and the prevention of the alarm floods as critical tasks in this implementation process. We develop an approach to satisfy these requirements utilizing the data mining techniques. We analyse established audit data from a well-known data repository considering the dimensions of the data paradigm. This led us to a tentative proposal of a conceptual mechanism for an integrated audit approach. With the increasing number of financial fraud cases, the

A. Ioannou · S. Valsamidis (✉) · A. Mandilas
Department of Accounting and Finance, International University of Greece, Campus of Kavala, Agios Loukas 65404, Kavala, Greece
e-mail: svalsam@teikav.edu.gr

A. Ioannou
e-mail: gikomo92@gmail.com

A. Mandilas
e-mail: smand@teiemt.gr

D. Bourlis
Mercedes-Benz, Perigiali 65402, Kavala, Greece
e-mail: bourlisdimitris@gmail.com

A. Horobet et al. (eds.), *Global, Regional and Local Perspectives on the Economies of Southeastern Europe*, Springer Proceedings in Business and Economics,
https://doi.org/10.1007/978-3-030-57953-1_14

application of data mining techniques could play a big part in improving the quality of conducting audit in the future.

Keywords Framework · Audit data · Data mining techniques

JEL Classification Codes G21 · L86

1 Introduction

The paradigm of audit data has tremendous impacts on both IT and auditing departments (Ghasemi et al. 2011). Financial statements are produced in automated Accounting Information Systems (AIS) and the auditor is faced with risen complexity and risks due to an increasing processing of ever-growing data (Vasarhelyi et al. 2015; Cao et al. 2015; Adamyk et al. 2018). Over the past 30 years, both information systems and auditing have undergone radical changes (Moffitt and Vasarhelyi 2013). Standards and regulations have also become frustratingly complex. But there's a powerful remedy for today's auditing headaches: continuous auditing and reporting (Singleton and Singleton 2005).

Financial statements are not as important to investors as they once were, as technology has changed the way companies create value today (Gallegos et al. 2004). While these changes pose serious threats to the economic viability of auditing, they also create new opportunities for auditors to pursue (Gangolly 2016). With the real-time accounting and electronic data interchange popularizing, Computer-Assisted Audit Tools (CAATs) are becoming even more necessary (Zhao et al. 2004). While they continue to acquire IT technical knowledge and skills, many auditors do not have the time or interest in becoming programmers. In the most based case, auditors in the new millennium need to understand the basics of computerized systems, including the core hardware components of a computer system and the basic concept for every computer program (input-process-output). At the same time, there is a lot more to understanding technology, including the basics of systems development, systems lifecycles, process flowcharting, programming logic, and writing scripts for analytics. These skills should exist in some aspect of the staffing or be outsourced (The Institute of Internal Auditors Research Foundation 2015).

Murphy and Groomer (2004) proposed how information technology (IT) frameworks, such as extensible markup language (XML) and Web services can be utilized to facilitate auditing for the next generation of accounting systems. The alternative architectures for auditing that have been proposed in both the research and practice environments are explored by Kuhn and Sutton (2010). They blend a focus on the practical realities of the technological options and ERP structures with the emerging theory and research on continuous assurance models. The focus is on identifying the strengths and weaknesses of each architectural form as a basis for forming a research agenda that could allow researchers to contribute to the future evolution of both ERP system designs and auditor implementation strategies.

Vasarhelyi et al. (2012) discussed the need for AIS to accommodate business needs generated by rapid changes in technology. It was argued that the real-time economy had generated a different measurement, assurance, and business decision environment. Three core assertions relative to the measurement environment in accounting, the nature of data standards for software-based accounting, and the nature of information provisioning, formatted and semantic, were discussed.

An implementation of the monitoring and control layer for monitoring of business process controls (CMBPC) in the US internal IT audit department of Siemens Corporation is described by Alles et al. (2018). Among their key conclusions is that "formalizability" of audit procedures and audit judgment is grossly underestimated. Additionally, while cost savings and expedience force the implementation to closely follow the existing and approved internal audit program, a certain level of reengineering of audit processes is inevitable due to the necessity to separate formalizable and non-formalizable parts of the program.

Lenz and Hahn (2015) find first, common themes in the empirical literature are identified. Second, the main threads into a model comprising macro and micro factors that influence audit effectiveness are synthesized. Third, promising future research paths that may enhance audit value proposition were derived. The "outside-in" perspective indicates a disposition to stakeholders' disappointment in audit: audit is either running a risk of marginalization or has to embrace the challenge to emerge as a recognized and stronger profession (PWC 2013). The suggested research agenda identifies empirical research threads that can help audit practitioners to make a difference for their organization, be recognized, respected and trusted and help the audit profession in its pursuit of creating a unique identity.

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 and Rodda 2004). Generally, audits are classified into two categories as internal and external auditing (Cosserat 2009). Internal-audit, although is an independent department of an organization, but resides within the organization. These are company-employees who are accountable for performing audits of financial and nonfinancial statements as per their annual audit plan. External audit is a fair and independent regular audit authority, which is responsible for an annual statutory audit of financial records. The external audit company has a fiduciary duty and is critical to the proper conduct of business.

There are many issues related to Audit and Decision Support Systems (Socea 2012; Schaltegger and Burritt 2017). Since 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 high risk of unfair practices, predictive analytics by using data mining techniques could provide actionable insights for the auditing. According to a research by Tysiac (2015), data analytics has benefited internal auditing more as compared to advancements it has contributed for the external audits. One of the most common applications of predictive analytics in audit is the classification of suspicious firm. 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.

Data mining techniques have already been applied for accounting information systems (Gelinas et al. 2017). Data mining techniques are providing great aid in financial accounting fraud detection, since dealing with the large data volumes and complexities of financial data are big challenges for forensic accounting (Sharma and Panigrahi 2013). The authors propose a framework based on data mining techniques for accounting fraud detection. Automated accounting fraud detection is presented also by Wang (2010). He categorizes, compares, and summarizes the data set, algorithm and performance measurement in published technical and review articles in accounting fraud detection. Data mining techniques accomplish the task of management fraud detection that could facilitate the auditors (Kirkos et al. 2007). The applications of data mining techniques in accounting and the proposal of an organizing framework for these applications is explored by Amani and Fadlalla (2017). They create a framework that combines the two well-known accounting reporting perspectives (retrospection and prospection), and the three well-accepted goals of data mining (description, prediction, and prescription). The proposed framework revealed that the area of accounting that benefited the most from data mining is assurance and compliance, including fraud detection, business health and forensic accounting. The ensemble machine learning method is also applied successfully for improving the classification accuracies of the auditing task (Kotsiantis et al. 2006).

The objective is to make the use of data analytics a sustainable, efficient, and repeatable process (Zhang et al. 2015). As with most uses of software technology, it is not a magic bullet. It requires attention to people and process issues, from management's commitment and support through training and the assignment of roles (Lientz and Larssen 2012).

The basic data analysis can be performed using a range of tools, including spreadsheets and database query and reporting systems (Antipova and Rocha 2018). There are certainly risks from using spreadsheets, apparent to any auditor because of the difficulty of ensuring data integrity. General purpose analysis tools also have their own limitations (Henry and Robinson 2009). It is clear that the analytics process must be managed in order to be relied upon by auditing, which is why accounting-specific analysis software should include capabilities such as: (i) Maintaining security and control over data, applications, and findings (ii) logging all activities (iii) analysis techniques designed to support accounting objectives and (iv) automated creation and execution of tests (Bellino et al. 2007).

The open source R software has one of the largest libraries of applications available. Free software such as R and Weka are used nationwide in university courses and by some research and technology firms, but are somewhat frowned upon by auditing firms because they are not validated (Appelbaum 2017). These concerns are not without merit, since open source software can be clumsier and less user friendly than proprietary software, but their utility should not be ignored. In addition, while a basic knowledge of statistics and information technology is becoming essential for all auditors; other, more specialized functions can be contracted to other experts, perhaps online.

Proprietary tools such as Audit Command Language (ACL) and Interactive Data Extraction and Analysis (IDEA), as well as generic statistical software such as Statistical Analysis System (SAS) and Statistical Package for the Social Sciences (SPSS), are frequently used by large businesses and large firms (Singleton 2006; Tysiac 2015). Furthermore, the capabilities and scope of these packages are constantly evolving, requiring that accountants and auditors have sufficient knowledge of analytics (Appelbaum et al. 2016). This convergence will likely take place with the emerging statistical and visualization toolsets being developed.

In this paper, we implement the aforementioned data mining techniques on the audit data of an existing audit organization of government firms of India, using the WEKA software package (Weka 2018). The outcomes support the decision-making process regarding the companies it audits (Hooda et al. 2018). The training and testing of a risk detection and management model will contribute to covering an existing research gap. The addressing of the above problems required the use of either specialized software such as ACL and IDEA, or general statistical packages such as SAS and SPSS with difficulty in adjusting and customizing audit data. It is worth noting that all of the aforementioned packages are commercial while WEKA is free software.

2 Background Theory

Data mining is an iterative process of creating predictive and descriptive models, by uncovering previously unknown trends and patterns in vast amounts of data, in order to extract useful information and support decision making (Kantardzic 2003). The most popular techniques for data mining (DM) are clustering, classification and finding association rules (Han et al. 2011).

Classification methods use a training dataset in order to estimate some parameters of a mathematical model that could in theory optimally assign each case from a new dataset into a specific class. In other words, the training set is used to train the classification technique how to perform its classification (Witten et al. 2016). There are various classification methods implemented in WEKA, like ZeroR, OneR, PART etc. The algorithm OneR uses the minimum-error attribute for prediction, discretizing numeric attributes (Holte 1993). In this technique, the attribute/s which best describe (s) the classification will be discovered.

Clustering refers to methods where a training set is not available. Thus, there is no previous knowledge about the data to assign them to specific groups. In this case, clustering techniques can be used to split a set of unknown cases into clusters. The clustering step contains digitalization clustering with the use of the k-means algorithm (MacQueen 1967; Kaufmann and Rousseeuw 1990) for unsupervised learning, called SimpleKMeans in WEKA. K-means is an efficient partitioning algorithm that decomposes the data set into a set of k disjoint clusters. It is a repetitive algorithm in which the items are moved among the various clusters until they reach the desired set of clusters. With this algorithm a great degree of similarity for the items of the

same cluster and a large difference of items, which belong to different clusters, are achieved. Furthermore, the algorithm automatically normalizes numerical attributes when doing distance computations.

According to Linoff and Berry (2011) relationship mining is a technique which discovers relationships between variables, in a data set with a large number of variables. There are four types of relationship mining: association rule mining, correlation mining, sequential pattern mining, and causal data mining. In this paper we focus on association rule mining (Liu et al. 1998). *Association rule mining* is one of the most well studied data mining tasks. It discovers relationships among attributes in databases, producing if-then statements concerning attribute-values (Agarwal et al. 1993). An association rule $X \rightarrow Y$ expresses a close correlation among items in a database, in which transactions in the database where X occurs, there is a high probability of having Y as well. In an association rule X and Y are called respectively the antecedent and consequent of the rule. The strength of such a rule is measured by values of its support and confidence. The confidence of the rule is the percentage of transactions with antecedent X in the database that also contain the consequent Y. The support of the rule is the percentage of transactions in the database that contains both the antecedent X and the consequent Y in all transactions in the database. There are several association rule-discovering algorithms available but Apriori algorithm is preferred as the most popular and effective algorithm for finding association rules over the discretized accounting data table (Agrawal and Srikant 1994). Apriori is the best-known algorithm to mine association rules. It uses a breadth-first search strategy to counting the support of item sets and uses a candidate generation function, which exploits the downward closure property of support. Iteratively reduces the minimum support until it finds the required number of rules with the given minimum confidence.

There are different techniques of categorization for association rule mining. Most of the subjective approaches involve user participation in order to express, in accordance with his/her previous knowledge, which rules are of interest. One technique is based on *unexpectedness* and *actionability* (Liu and Hsu 1996; Liu et al. 2000). Unexpectedness expresses which rules are interesting if they are unknown to the user or contradict the user's knowledge. Actionability expresses that rules are interesting if users can do something with them to their advantage. The number of rules can be decreased to unexpected and actionable rules only. Another technique proposes the division of the discovered rules into three categories (Minaei-Bidgoli et al. 2004). (1) *Expected and previously known*: This type of rule confirms user beliefs, and can be used to validate our approach. Though perhaps already known, many of these rules are still useful for the user as a form of empirical verification of expectations. (2) *Unexpected*: This type of rule contradicts user beliefs. This group of unanticipated correlations can supply interesting rules, yet their interestingness and possible actionability still requires further investigation. (3) *Unknown*: This type of rule does not clearly belong to any category, and should be categorized by domain specific experts. The Weka system has several association rule-discovering algorithms available (Hipp et al. 2000). The Apriori algorithm will be used for finding association rules over discretized data (Agrawal and Srikant 1994).

3 Approach

The proposed approach consists of five steps (Fig. 1):

1. Target data finding.
2. Data pre-processing.
3. Classification.
4. Clustering.
5. Association rule mining.

3.1 Dataset

The dataset in which the methodology will be applied is in the world-wide known machine learning repository UCI. 463 datasets are included in a wide range of applications (UCI1 2018). In particular, for Audit, there is a set of data to be used in the

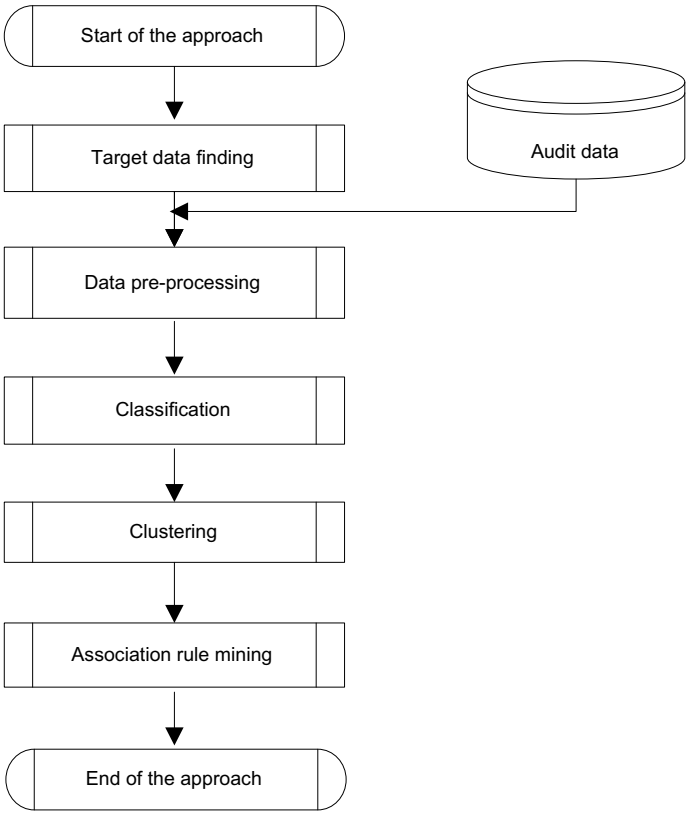

Fig. 1 Approach of five steps

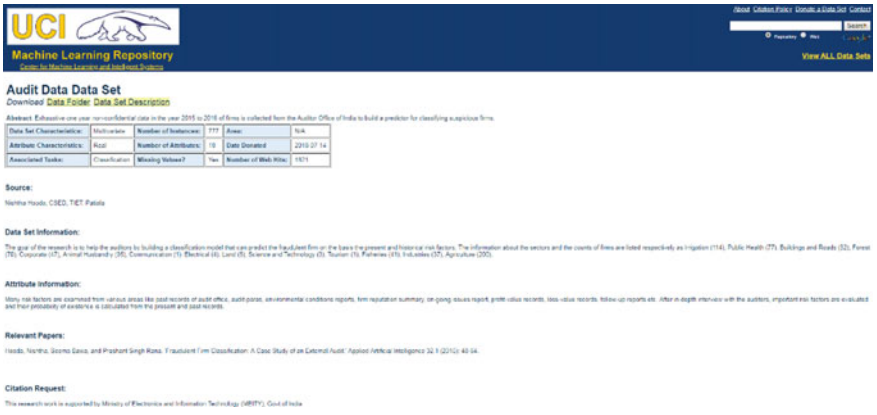

Fig. 2 Audit data from the repository UCI

study (UCI2 2018). The general information for that particular dataset is shown in Fig. 2.

Comptroller and Auditor General (CAG) of India is an independent constitutional body of India. It is an authority that audits receipts and expenditure of all the firms that are financed by the government of India. While maintaining the secrecy of the data, exhaustive one year non confidential data in 2015 and 2016 of firms is collected from the Auditor General Office (AGO) of CAG. There are total 777 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. The information about the sectors and their counts are summarized in Table 1.

Many risk factors are examined from various areas like past records of audit office, audit-paras, environmental conditions reports, firm reputation summary, on-going issues report, profit-value records, loss- value records, follow-up reports etc. After an in-depth interview with the auditors, important risk factors are evaluated and their probability of existence is calculated from the present and past records. Tables 2 and 3 describe the various examined risk-factors that are involved in the case study. Various risk factors are categorized, but combined audit risk is expressed as one function called an Audit Risk Score (ARS) using an audit analytical procedure. At the end of risk assessment, the firms with high ARS scores are classified as "Fraud" firms, and low ARS score companies are classified as "No-Fraud" firms.

3.2 Tool

The WEKA (Waikato Environment for Knowledge Analysis) computer package was used in order to apply classification, clustering and association rule mining methods to the dataset (Witten et al. 2016). WEKA is open source software that provides a collection of machine learning and data mining algorithms. Figure 3 shows the basic

Table 1 Target sectors

Sector ID	Target sector	Information	Number of target firms
1	IR	Irrigation	114
2	P	public health	77
3	BR	Buildings and roads	82
4	FO	Forest	70
5	CO	Corporate	47
6	AH	Animal husbandry	95
7	C	Communication	1
8	E	Electrical	4
9	L	Land	5
10	S	Science and Technology	3
11	T	Tourism	1
12	F	Fisheries	41
13	I	Industries	37
14	A	Agriculture	200

Table 2 Risk factors classification and other features in model

Feature	Information	Feature	Information
Para a value	Discrepancy found in the planned-expenditure of inspection and summary report A in Rs (in crore).	Sector score	Historical risk score value of the target-unit in the Table 1 using analytical procedure
Para B value	Discrepancy found in the unplanned-expenditure of inspection and summary report B in Rs (in crore).	Loss	Amount of loss suffered by the firm last year
Total	Total amount of discrepancy found in other reports Rs (in crore).	History	Average historical loss suffered by firm in the last 10 years
Number	Historical discrepancy score	District score	Historical risk score of a district in the last 10 years
Money value	Amount of money involved in misstatements in the past audits		

Table 3 Other features

Feature	Information	Feature	Information
Sector ID	Unique ID of the target sector	Location ID	Unique ID of the city/province
ARS	Total risk score using analytical procedure	Audit ID	Unique Id assigned to an audit case
Risk class	Risk class assigned to an audit-case		

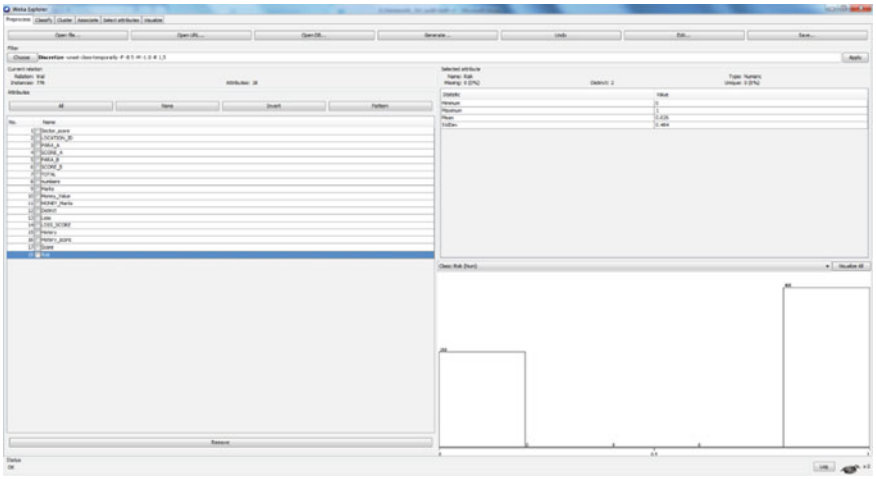

Fig. 3 WEKA environment

Graphical User Interface (GUI) of WEKA. One of the main objectives of WEKA is to mine information from existing datasets; the main reason for choosing Weka is that provides a collection of machine learning and data mining algorithms for data pre-processing, classification, regression, clustering, association rules, and visualization (Hall et al. 2009).

4 Results

As it is depicted in Fig. 2, the dataset contains 777 instances. There are no missing values for all the attributes.

In WEKA environment data is depicted as in Fig. 4.

4.1 Pre-processing

The first step before applying the described data mining techniques is the pre-processing of the data in order to prepare them for data analysis.

Certain filters were applied on the data. Firstly, the filter Remove was applied on the attributes PARA_A, PARA_B, Money_Value, Loss, History and Score, since they obviously are dependent on the attributes SCORE_A, SCORE_B, Money_Marks Loss_Score, History_Score and Risk respectively (Fig. 5).

The filter *NumericalToNominal* was applied on the attributes SCORE_A, SCORE_B, Marks, MONEY_Marks, District, LOSS_SCORE, History_score and

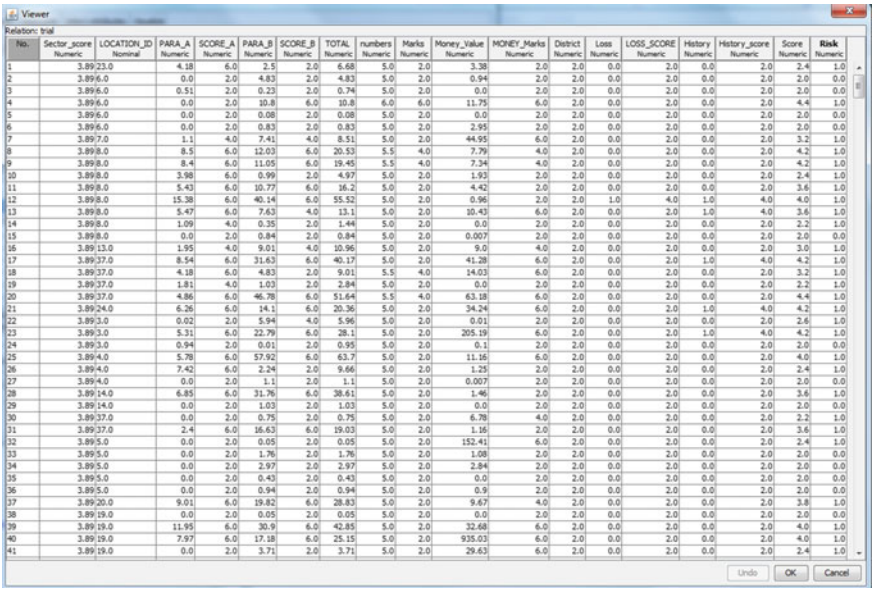

Fig. 4 The dataset in WEKA environment

Risk in order to convert numeric variables and their values to nominal. The attributes number 3, 4, 7–12 are converted to nominal (Fig. 6).

Furthermore, the filter *Discretize* was applied in order to discretize numeric variables Sector_score and TOTAL and make them nominal. Figure 7 depicts all the variables used in our analysis.

The Discretization Options are portrayed in Fig. 8.

By visualizing all, it is possible to display the graphical representations of each attribute in relation to any other attribute as portrayed below (Fig. 9).

4.2 Classification

In the classification step, the algorithm OneR is applied. The attribute "Risk" is used as a class. Figure 10 presents the overall accuracy of the model computed from the training dataset and is equal to 84.4072%. The worst performance for the Precision on the class 0 and equals 70.6%, whereas the best performance is also for the Precision but on the class 1 and equals 100%. Confusion matrix validates that the precision for class 1 (variable b) is 100%. On the other hand, 121 instances were faulty not classified in class 0.

The results indicate that the attribute which describes the classification is variable SCORE_A. This means that variable Risk is more closely related to the variable SCORE_A than the other variables.

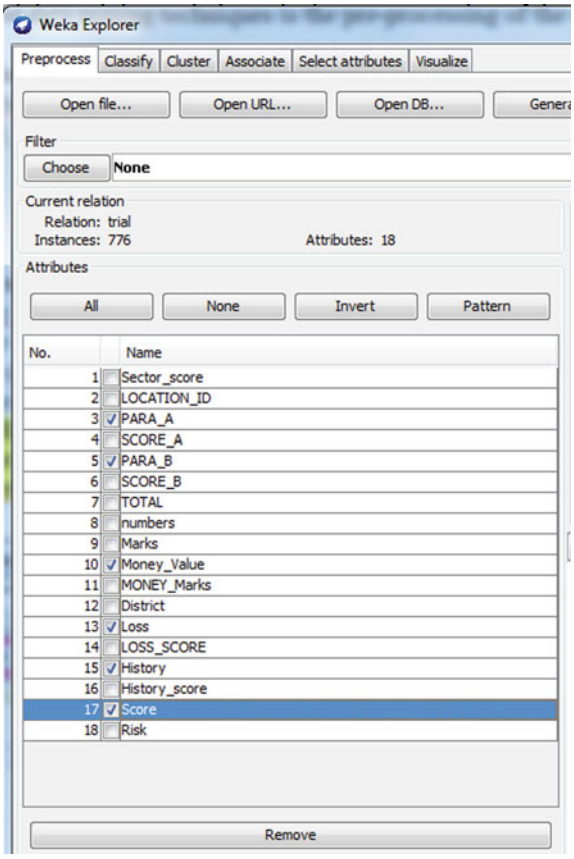

Fig. 5 The filter remove

4.3 Clustering

The clustering step was performed using the k-means algorithm (SimpleKmeans in the context of WEKA). The number of clusters is set to 2, since the variable "Risk" was used to compute the accuracy of the clustering and inspect the audit data. Figure 11 shows the results of the clustering based on variable "Risk". The clustered instances are 433 (56%) and 343 (44%) respectively. It is also evident from the cluster centroids that "Risk" has value 0 in the first cluster and value 1 in the second cluster.

The differences between the two clusters are focused on attributes: Sector_score, LOCATION_ID, SCORE_A, TOTAL and Risk.

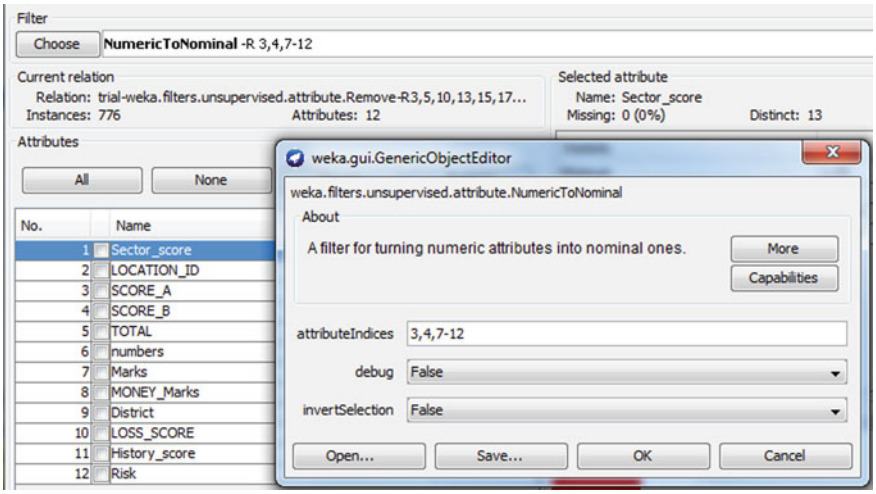

Fig. 6 The filter numerical to nominal

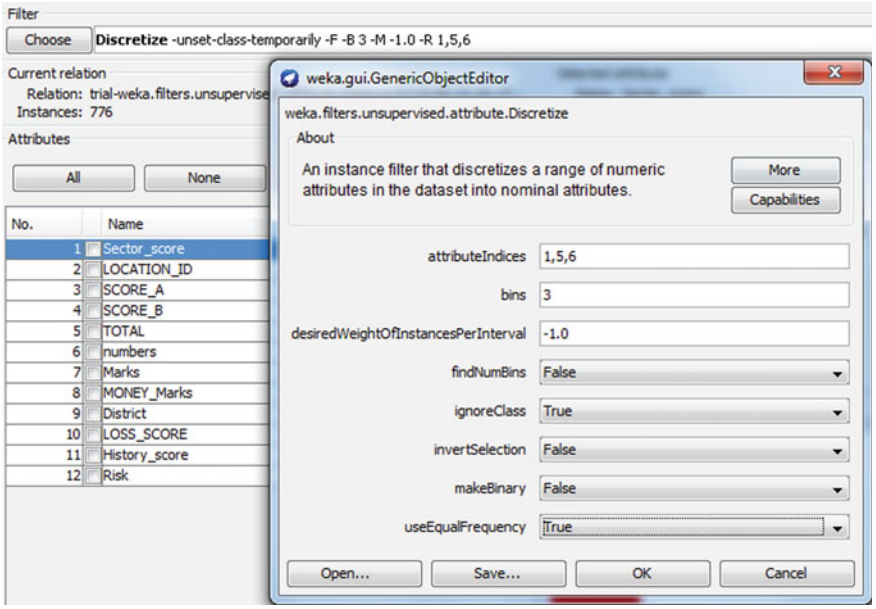

Fig. 7 The filter discretize

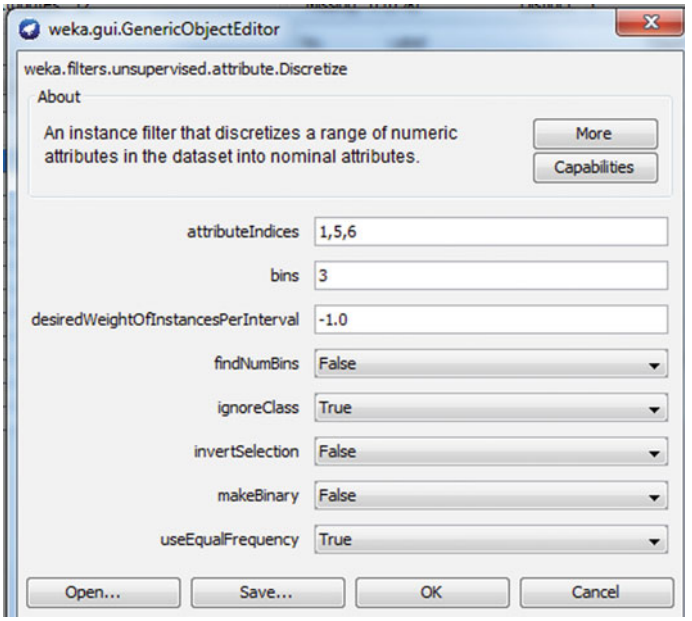

Fig. 8 Discretization options

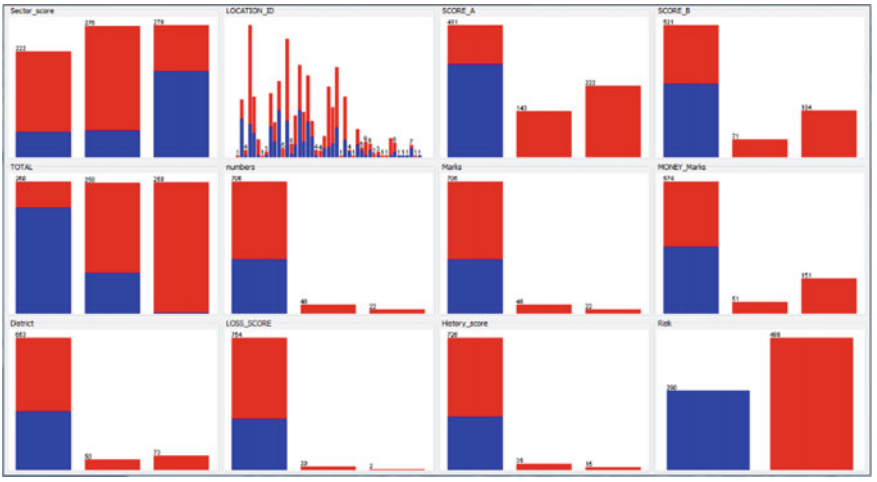

Fig. 9 Visualization of the attributes with class variable “Risk”

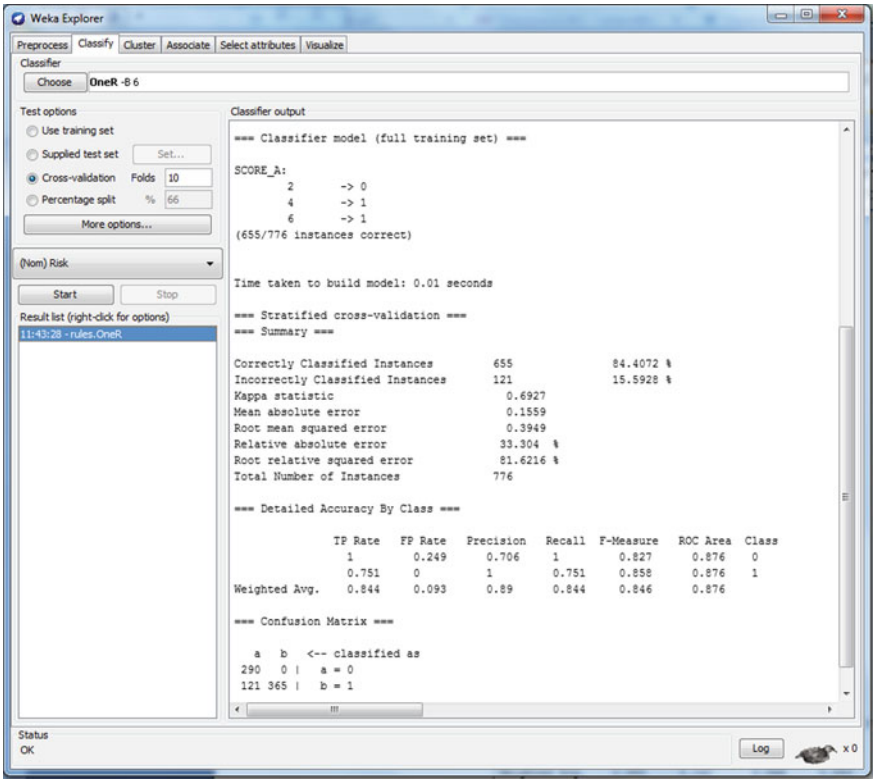

Fig. 10 Classification results using variable "Risk" as class

4.4 Association Rule Mining

The Apriori algorithm (Agrawal et al. 1993) was used for finding association rules for our dataset. The WEKA produced a list of 15 rules (Table 4) with the support of the antecedent and the consequent (total number of items) at 0.1 minimum, and the confidence of the rule at 0.9 minimum (percentage of items in a 0 to 1 scale). The application of the Apriori algorithm for association provided useful insights into the audit data. Table 4 shows how a large number of association rules can be discovered.

There is couple of uninteresting rules regarding the aim of the research, like the similar rules 1 and 2 which show expected or conformed relationships. If Marks = 2 then numbers is between 0 and 5.25 and vice versa. These are also symmetrical rules since the antecedent element and the consequent element are interchanged.

There are some similar rules, rules with the same element in antecedent and consequent but interchanged (3 and 4, and 5 and 6). The variables Marks and numbers appear in antecedent and consequent elements but they are interchanged. There is also a symmetric triad of rules (10, 11 and 12) where Marks and numbers appear also in antecedent and consequent elements interchanged.

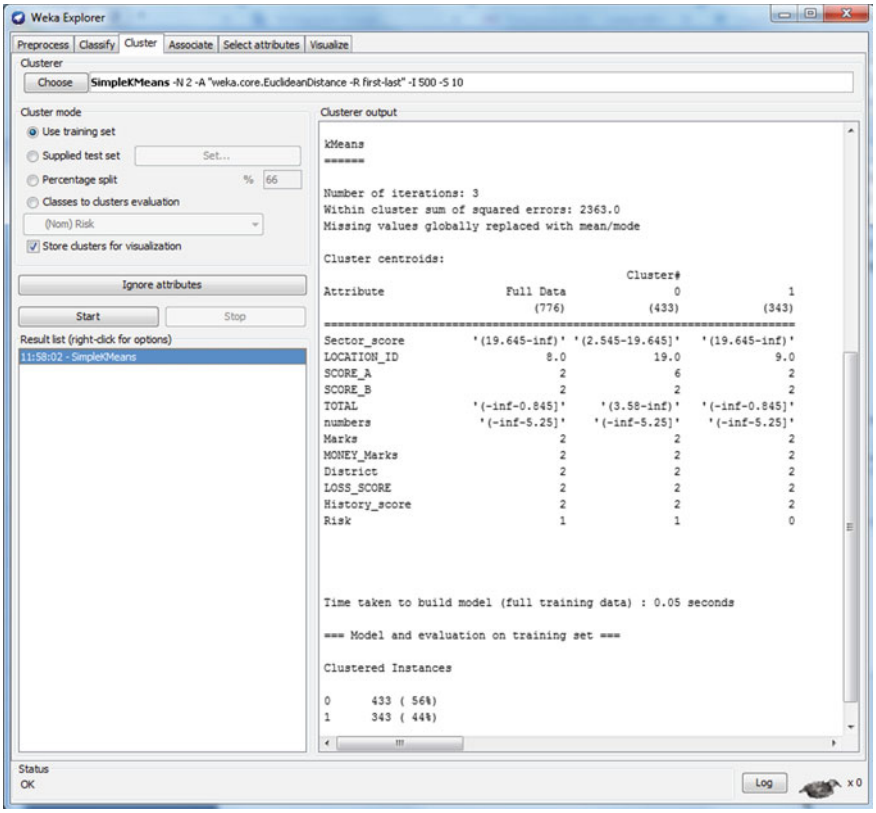

Fig. 11 Clustering results. Variable "Risk" is used for assessing the clustering

There are is also an uninteresting or redundant rule (rules with a generalization of relationships of other rules, like rule 15 with rules 13 and 14).

But there are also interesting rules such as 7, 8 and 9 which offer actionability for an auditor. These three rules are useful for an auditor, since s/he can pay more attention to the companies with History_score = 2, numbers between 0 and 5.25 and Marks = 2.

Summarizing the results from the classification, the clustering and the association rule mining methods, it can be concluded that:

1. The attribute which best describes the classification is the variable SCORE_A. The attribute "Risk" (Fraud/Non fraud) is used as a class.
2. Using "Risk" as class attribute in clustering, the results show that companies which belong to the second cluster have better values in the parameters regarding the Risk.
3. For companies with History_score = 2, numbers between 0 and 5.25 and Marks = 2, an auditor must pay more attention.

Table 4 Best rules found with Apriori algorithm based on confidence metric

Best rules found:
1. Marks = 2 706 ==> numbers = '(-inf-5.25]' 706 conf:(1)
2. numbers = '(-inf-5.25]' 706 ==> Marks = 2 706 conf:(1)
3. Marks = 2 LOSS_SCORE = 2 688 ==> numbers = '(-inf-5.25]' 688 conf:(1)
4. numbers = '(-inf-5.25]' LOSS_SCORE = 2 688 ==> Marks = 2 688 conf:(1)
5. Marks = 2 History_score = 2 673 ==> numbers = '(-inf-5.25]' 673 conf:(1)
6. numbers = '(-inf-5.25]' History_score = 2 673 ==> Marks = 2 673 conf:(1)
7. History_score = 2 726 ==> LOSS_SCORE = 2 710 conf:(0.98)
8. numbers = '(-inf-5.25]' 706 ==> LOSS_SCORE = 2 688 conf:(0.97)
9. Marks = 2 706 ==> LOSS_SCORE = 2 688 conf:(0.97)
10. numbers = '(-inf-5.25]' Marks = 2 706 ==> LOSS_SCORE = 2 688 conf:(0.97)
11. Marks = 2 706 ==> numbers = '(-inf-5.25]' LOSS_SCORE = 2 688 conf:(0.97)
12. numbers = '(-inf-5.25]' 706 ==> Marks = 2 LOSS_SCORE = 2 688 conf:(0.97)
13. numbers = '(-inf-5.25]' 706 ==> History_score = 2 673 conf:(0.95)
14. Marks = 2 706 ==> History_score = 2 673 conf:(0.95)
15. numbers = '(-inf-5.25]' Marks = 2 706 ==> History_score = 2 673 conf:(0.95)

5 Discussion and Conclusions

In this paper, a framework is proposed for audit, accounting, financial, and risk management executives. It identifies the management of audit alarms and the prevention of the alarm floods as critical tasks in the implementation process. The developed framework solves these problems by using the data mining techniques. The audit data originated from an existing audit organization stored in a well known data repository and the used software package was WEKA. With this pilot application of audit data, an audit process is carried out and the proposed decision support framework is able to assist an auditor to decide on the size of work required for a particular company or organization, or even omit to visit low-risk companies. Predicting fraud in a company is an important step in the preliminary planning stage of the audit, as high-risk companies are targeted to maximize audit research.

Since, the implementation of auditing is a recognized challenge among researchers and practitioners, and traditional audit tools and techniques neglect the potential of data analytics, the development of an appropriate audit framework based on data

mining tools and techniques is imperative need. We analyzed established audit data considering the dimensions of the data paradigm in this paper. This led us to a proposal of a conceptual architecture for an integrated audit approach. The proposed framework is independent of the particular dataset and can be applied to other similar datasets by using the same data mining techniques. The outcomes support the decision-making process regarding the companies it audits. The training and testing of a risk detection and management model contributes to cover an existing research gap. With the increasing number of financial fraud cases, the application of data mining techniques could play a big part in improving the quality of conducting audit in the future.

The question of whether the proposed framework can be applied to other financial and administrative applications can only be answered satisfactorily once it will be tested to them as well. The use of the method requires users with specific capabilities and knowledge. That is to know to use in depth audit and data mining techniques.

References

Adamyk, O., Adamyk, B., Khorunzhak, N. (2018). Auditing of the software of computer accounting system. In *ICTERI Workshops* (pp. 251–262).

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data* (pp. 207–216).

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In *Proceeding 20th International Conference Very Large Data Base* (Vol. 1215, pp. 487–499). VLDB.

Alles, M., Brennan, G., Kogan, A., & Vasarhelyi, M. A. (2018). Continuous monitoring of business process controls: A pilot implementation of a continuous auditing system at Siemens. In *Continuous Auditing: Theory and Application* (pp. 219–246). Bingley, West Yorkshire, England; Emerald Publishing Limited.

Amani, F. A., & Fadlalla, A. M. (2017). Data mining applications in accounting: A review of the literature and organizing framework. *International Journal of Accounting Information Systems, 24,* 32–58.

Antipova, T., & Rocha, Á. (Eds.). (2018). *Information technology science*. Cham: Springer.

Appelbaum, D. (2017). Introduction to data analysis for auditors and accountants. *The CPA Journal, 7*.

Appelbaum, D., Kogan, A., Vasarhelyi, M. A. (2016). *Analytics in external auditing: A literature review*. Rutgers University CARLab Newark, NJ, USA Working Paper.

Bellino, C., Wells, J., & Hunt, S. (2007). *Global technology sudit guide (GTAG) 8: Auditing application controls.*

Cao, M., Chychyla, R., & Stewart, T. (2015). Big data analytics in financial statement audits. *Accounting Horizons, 29*(2), 423–429.

Cosserat, G. W., & Rodda, N. (2004). *Modern Auditing*. Hoboken, New Jersey: Wiley.

Cosserat, G. (2009). Accepting the engagement and planning the audit. *Modern Auditing*, 734–736.

Gelinas. U. J., Dull, R. B., Wheeler, P. & Hill, M. C. (2017). *Accounting information systems.* Cengage Learning.

Gallegos, F. S., Manson, D. P., & Gonzales, C. (2004). *Information technology control and audit*. Auerbach Publications.

Gangolly, J. S. (2016). American institute of certified public accountants, INC., audit analytics and continuous audit: Looking towards the future. *Journal of Emerging Technologies in Accounting, 13*(1):187–188.

Ghasemi, M., Shafeiepour, V., Aslani, M., & Barvayeh, E. (2011). The impact of information technology (IT) on modern accounting systems. *Procedia-Social and Behavioral Sciences, 28,* 112–116.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter, 11*(1), 10–18.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: Concepts and techniques.* Burlington, Massachusetts: Morgan Kaufmann.

Henry, E., & Robinson, T. R. (2009). *Financial statement analysis: An introduction. International financial statement analysis.* Hoboken, New Jersey: Wiley.

Hipp, J., Güntzer, U., & Nakhaeizadeh, G. (2000). Algorithms for association rule mining—A general survey and comparison. *ACM SIGKDD Explorations Newsletter, 2*(1), 58–64.

Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning, 11*(1), 63–90.

Hooda, N., Bawa, S., & Rana, P. S. (2018). Fraudulent firm classification: A case study of an external audit. *Applied Artificial Intelligence, 32*(1), 48–64.

Kantardzic, M. (2003). *Data mining: Concepts, models, methods, and algorithms.* New York, NY: Wiley.

Kaufmann, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis.* New York: Wiley.

Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications, 32*(4), 995–1003.

Kotsiantis, S., Koumanakos, E., Tzelepis, D., & Tampakas, V. (2006). Forecasting fraudulent financial statements using data mining. *International Journal of Computational Intelligence, 3*(2), 104–110.

Kuhn Jr, J. R., & Sutton, S. G. (2010). Continuous auditing in ERP system environments: The current state and future directions. *Journal of Information Systems, 24*(1), 91–112.

Lenz, R., & Hahn, U. (2015). A synthesis of empirical internal audit effectiveness literature pointing to new research opportunities. *Managerial Auditing Journal, 30*(1), 5–33.

Lientz, B., & Larssen, L. (2012). *Manage IT as a Business.* London: Routledge.

Linoff, G. S., & Berry, M. J. (2011). *Data mining techniques: for marketing, sales, and customer relationship management*. Indianapolis, Indiana: Wiley.

Liu, B., & Hsu, W. (1996). Post-analysis of learned rules. In *AAAI/IAAI* (Vol. 1, pp. 828–834).

Liu, B., Hsu, W., Chen, S., & Ma, Y. (2000). Analyzing the subjective interestingness of association rules. *IEEE Intelligent Systems, 15*(5), 47–55.

Liu, B., Hsu, W., & Ma, Y. (1998). Integrating classification and association rule mining. In *KDD* (Vol. 98, pp. 80–86).

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (pp. 281–297). California, USA.

Minaei-Bidgoli, B., Tan, P. N., & Punch, W. F. (2004). Mining interesting contrast rules for a web-based educational system. In *Proceedings of International Conference on Machine Learning Applications, Louisville* (pp. 320– 327). USA.

Moffitt, K. C., & Vasarhelyi, M. A. (2013). AIS in an age of big data. *Journal of Information Systems, 27*(2), 1–19.

Murthy, U. S., & Groomer, S. M. (2004). A continuous auditing web services model for XML-based accounting systems. *International Journal of Accounting Information Systems, 5*(2), 139–163.

PwC. (2013). Maximising internal audit value: 2013 state of the internal audit profession survey—Russia supplement. PwC Russia: Moscow. https://www.pwc.ru/ru/riskassurance/assets/russian-ia-survey-2013-en.pdf. Accessed October 20, 2019.

Schaltegger, S., & Burritt, R. (2017). *Contemporary environmental accounting: Issues, concepts and practice*. London: Routledge.

Sharma, A., & Panigrahi, P. K. (2013). A review of financial accounting fraud detection based on data mining techniques. arXiv preprint arXiv:1309.3944.

Singleton, T. (2006). Generalized audit software: Effective and efficient tool for today's IT audits, IS ACA.

Singleton, T., & Singleton, A. J. (2005). Auditing headaches? Relieve them with CAR. *Journal of Corporate Accounting & Finance, 16*(4), 17–27.

Socea, A. D. (2012). Managerial decision-making and financial accounting information. *Procedia-Social and Behavioral Sciences, 58,* 47–55.

The Institute of Internal Auditors Research Foundation. (2015). Staying a step ahead internal audit's use of technology.

Tysiac, K. (2015). Data analytics helps auditors gain deep insight. *Journal of Accountancy, 219*(4), 52.

UCI1. (2018). https://archive.ics.uci.edu/ml/index.php. Accessed November 14, 2019.

UCI2. (2018). https://archive.ics.uci.edu/ml/datasets/Audit+Data#. Accessed November 14, 2019.

Vasarhelyi, M., Kogan, A., & Tuttle, B. M. (2015). Big data in accounting: An overview. *Accounting Horizons, 29*(2), 381–396.

Vasarhelyi, M., Romero, S., Kuenkaikaew, S., & Littley, J. (2012). Adopting continuous audit/continuous monitoring in internal audit. *ISACA Journal, 3,* 31.

Wang, S. (2010). A comprehensive survey of data mining-based accounting-fraud detection research. In *2010 International Conference on Intelligent Computation Technology and Automation* (Vol. 1, pp. 50–53). IEEE.

Weka. (2018). https://www.cs.waikato.ac.nz/ml/weka/. Accessed November 20, 2019.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Burlington, Massachusetts: Morgan Kaufmann.

Zhang, J., Yang, X., & Appelbaum, D. (2015). Toward effective Big Data analysis in continuous auditing. *Accounting Horizons, 29*(2), 469–476.

Zhao, N., Yen, D., & Chang, I. (2004). Auditing in the e-commerce era. *Information Management & Computer Security, 12*(5), 389–400.