Journal of Engineering Science and Technology Review
Special Issue on Telecommunications, Informatics, Energy and Management 2019

JOURNAL OF
Engineering Science and Technology Review

www.jestr.org

Conference Article

Knowledge Mining from Accounting Data as Imechanism for Decision Support

Evaggelia Tsapani*, Elpida Tenidou, Dimosthenis Pappas and Stavros Valsamidis

International Hellenic University, Campus of Kavala, Agios Loukas, 65404 Kavala, Greece

Received 25 September 2019; Accepted 28 February 2020

Abstract

The rise of automated Accounting Information Systems (AIS) in enterprises caused the growing size of accounting-relevant data. Thus, the data analysis for accounting purpose is heavily increasing in complexity. Although, the basic data analysis can be performed using a range of tools, including spreadsheets, database query and reporting systems, the use of specific tools devoted to this purpose is preferable. Accountants need to know basic statistics and information technology in order to explain the accounting events and forecast the future financial situation of their businesses. Data mining is an interdisciplinary science which stands between statistics and information technology and has already been used in AIS. Although free and open source software such as R and Weka, ideal for the implementation of data mining techniques are used for academic reasons, the accounting firms are unwilling to use them for professional reasons.
This study provides an analysis of accounting data using the most popular data mining methods: clustering, classification and association rule mining. The algorithms SimpleKmeans, OneR, and Apriori respectively were used for the aforementioned methods. The data pre-processing and visualization are also used in the study. A case study of an existing business is used for the implementation of the aforementioned data mining techniques with the use of the open source software package WEKA. The five variables which are used for data analysis, Day, Document, Customer, Value and Category may support the decision making process regarding the customers and their transactions.

Keywords: Accounting Information Systems, Support Systems, Data mining techniques, WEKA.

1. Introduction

The paradigm of accounting data has tremendous impacts on both IT departments and accounting [1]. Financial statements are produced in automated *Accounting Information Systems (AIS)* [2] and the accountant is faced with risen complexity and risks due to an increasing processing of ever-growing data [3].

The basic data analysis can be performed using a range of tools, including spreadsheets and database query and reporting systems [4]. There are certainly risks from using spreadsheets, apparent to any accountant because of the difficulty of ensuring data integrity. General purpose analysis tools also have their own limitations [5]. It is clear that the analytics process must be managed in order to be relied upon by accounting, 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 [6].

The objective is to make the use of data analytics a sustainable, efficient, and repeatable process [7]. 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 [8].

While the free and open source software such as R and Weka are used widely for educational purposes, the accounting firms are hesitant to use them because they are not validated [9]. They also think that free and open source software is less user friendly than proprietary software. In the same study, they support that other, more specialized functions can be contracted to other experts. On the other hand, accountants need to know basic statistics and information technology in order to explain the accounting events and forecast the future financial situation of their businesses.

There are many issues related to Accounting Information Systems and *Decision Support Systems* [10] [11]. Data mining techniques have already been applied for accounting information systems [12]. There are several data mining (DM) techniques but the most popular are clustering, categorization and finding association rule [13]. There various definitions for data mining. Data mining is the process of finding patterns and get useful information from the data [14]. Data mining is a multidisciplinary methodology for extracting knowledge from data [15]. Data mining is an iterative process of creating predictive and descriptive models, by uncovering previously unknown trends and patterns in vast amounts of data [16].

Data mining techniques are used in forensic accounting to detect fraud in large data volumes and complexities of financial data [17]. The authors propose a framework for accounting fraud detection based on data mining techniques. Wang presents automated accounting fraud detection by categorizing, comparing, and summarizing a data set of published technical and review articles in accounting fraud detection [18]. Data mining techniques facilitate the auditors

*E-mail address: e.tsapani@gmail.com, svalsam@teiemt.gr
ISSN: 1791-2377

to perform fraud detection [19]. Data mining techniques in accounting are also explored in a framework for these types of applications [20]. The area of accounting that benefited the most from data mining is assurance and compliance. The two well-known accounting reporting perspectives, retrospection and prospection, are combined with the three well-accepted goals of data mining: description, prediction, and prescription.

This study presents the implementation of classification clustering and association rule mining in the accounting data of an existing business, with the use of the WEKA software package. The outcomes may support the decision making process regarding the customers and their transactions.

2. Material and method

The material is the dataset and the method is the DM techniques used.

2.1 Dataset

The dataset was collected from a company located in Alexandroupolis. The data were collected during 2017 and involve 1021 financial transactions (instances). The data are originally in .csv (comma separated values) form. Each record is described by 5 variables. The five variables which are used are Day, Document, Customer, Value and Category. Table 1 presents each variable in detail.

Table 1. The variables used in our analysis

Variable Name	Description	Type
Day	The day which took place the invoicing process	Nominal
Document	The commercial invoice	Nominal
Customer	The name of the customer	Nominal
Value	The value of products before taxis	Numeric
Category	If the customer is a company or not (A for companies, B for persons)	Nominal

2.2 Tool

The WEKA computer package is used in order to apply the data mining techniques to the dataset. Weka is not just a simple software package but an integrated environment full of old and new machine learning algorithms, data preprocessing and visualization tools.

Apart of its embedded algorithms, the user can plug in his/her own data mining algorithms. Some common functions of Weka are the preparation of the input data, the statistical evaluation of learning schemes and the visualization both to the input data and the result of learning by focusing in specific changes. WEKA is developed and maintained by the University of Waikato in New Zealand [21]. It can run on all the well known operating systems such as Linux, Windows, and Macintosh without limitations. The basic Graphical User Interface (GUI) of WEKA is depicted in Fig. 1 [21].

2.3 Method

Classification is the method where given data are assigned to predefined classes [22]. Classes are also called categories or labels or targets. Thus, the main goal is to predict the class for a given unlabeled item. The classification algorithms are used in data mining tasks because it is necessary in many cases the data to be assigned to a finite set of predefined classes [23]. The most popular techniques for classification problems are: classification trees, logistic regression, discriminant analysis, neural networks, boosted trees, random forests, deep learning methods, nearest neighbors, support vector machines. In machine learning, the techniques which are used are: Naive Bayes Classifier, Logistic Regression, Nearest Neighbor, Decision Trees, Boosted Trees, Support Vector Machines, Neural Networks and Random Forest. All the aforementioned techniques are also implemented in WEKA with various algorithms. The OneR algorithm uses the minimum-error attribute for prediction, by discretizing numeric attributes. The attribute/s which best describe(s) the classification will be discovered [24].

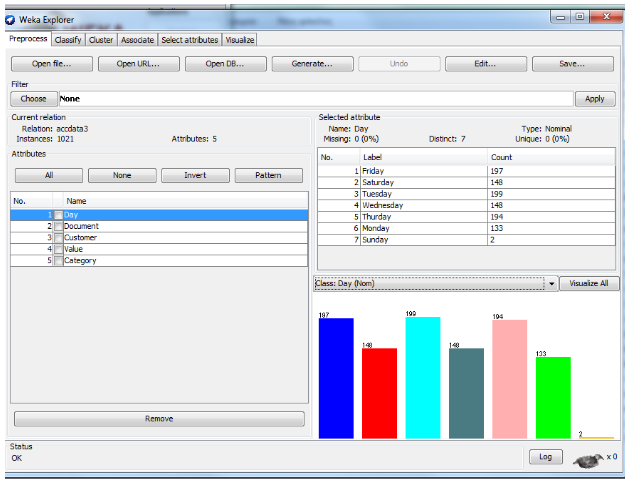

Fig. 1. WEKA environment

Clustering is one exploratory data analysis method used for the perception of the data. It is the task of identifying clusters in the data such that data in the same cluster are very similar while data in different clusters are very different [25]. Which similarity measures will be used is dependent on the application. Another goal of clustering is to reduce the amount of data by grouping similar data together. This is similar to the way people organize data in order to categorize them into groups. Unlike classification, clustering is an unsupervised learning method because we cannot compare the output of the clustering algorithm to the predefined classes and evaluate its performance [26]. We just explore the structure of the data by grouping the data points into different groups. There are two basic types of clustering: hierarchical and partitional. Hierarchical clustering is either the gradual split of larger clusters into smaller or the merge of smaller clusters into larger ones. Partitional clustering decomposes the dataset into a set of disjoint clusters. The SimpleKmeans algorithm which is a variation of K-means and belongs to the latter type of clustering [27].

Relationship mining is a technique which discovers relationships between variables, in a dataset with a large number of variables [28]. Although there are four types of relationship mining (association rule mining, correlation mining, sequential pattern mining and causal data mining), association rule mining is the most well known and studied data mining task [29]. Association rule mining is a method which observes frequently occurring patterns, correlations, or associations from datasets found in various databases and other forms of repositories [30]. Association rule mining include If.. Then.. statements which help discover relationships in data. The part if is called antecedent and the part Then is called consequent. The antecedent is an item found within the dataset while the consequent is an item found in combination with the antecedent. The association rules are discovered by searching data for frequent if-then patterns and using specific metrics to identify the most important relationships [30]. Although there are several metrics, the strength of the rules is usually measured by the values of support and confidence. Support indicates how frequently the if/then relationship appears in the database. It is the percentage of transactions in the database that contains both the antecedent X and the consequent Y in all transactions in the database. Confidence indicates the number of times these relationships have been found to be true It is the percentage of transactions with antecedent X in the database that also contain the consequent Y. Sometimes, another metric called lift can be used to compare confidence with expected confidence. Nevertheless, it is not used in this approach. The most popular algorithms for generating association rules are Apriori, AIS, Eclat, SETM, and FP-Growth. Apriori algorithm is the most popular and effective algorithm for finding association rules over the discretized accounting data table [31]. It uses 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 ways for categorization of association rule mining. All include the subjectiveness because the findings of the rules are compared with users' previous knowledge. One categorization is to Expected and previously known, Unexpected and Unknown rules [32]. Another categorization is based on the Unexpectedness and Actionability of rules [33] [34]. A rule is expected and previously known if it just confirms user beliefs. Alhough these rules are already known, they verify and validate the user expectations. A rule is categorized as unexpected if the rule contradicts to user beliefs. An unknown rule does not belong to any category and should be categorized only by domain specific experts. Unexpected are the rules which are revealed to be either unknown or to contradict the user's knowledge. A rule offers actionability to the user if s/he can react according to the findings of the rule.

3. Results

As it is depicted in figure 1, the dataset contains 1021 instances. There are no missing values for the attributes Day, Document, Customer, Value and Category.

3.1 Pre-processing

The first step is to preprocess the data. The filter NumericalToNominal was applied to the numeric variable Value in order to convert the numeric data value to nominal. Since the other variables are nominal, they will not be converted. As fig. 2 depicts, the filter Discretize is selected.

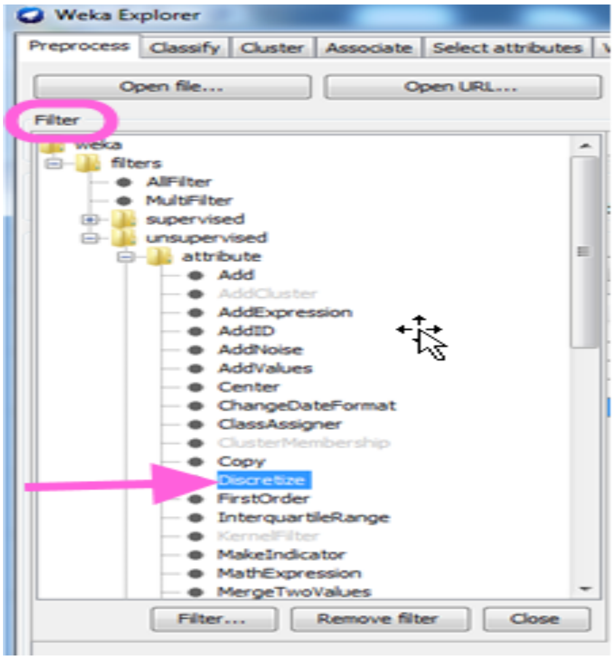

Fig. 2. Filter Discretize

Since we define the attribute we want to include in the pre-process (4 → Value), we activate the filter by pressing the Apply button (figures 3 and 4).

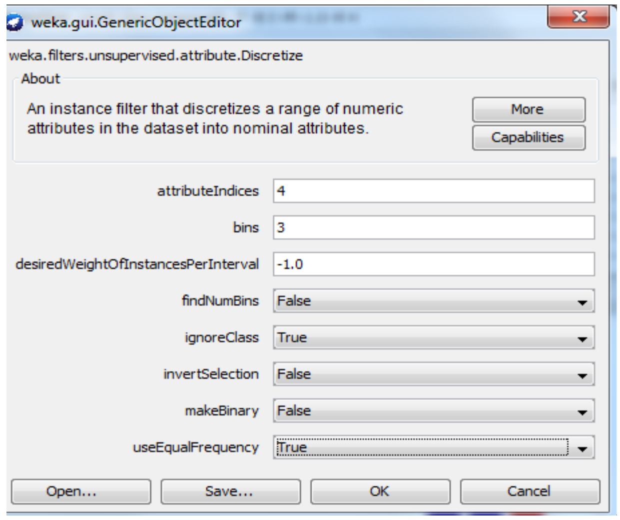

Fig. 3. Discretization Options

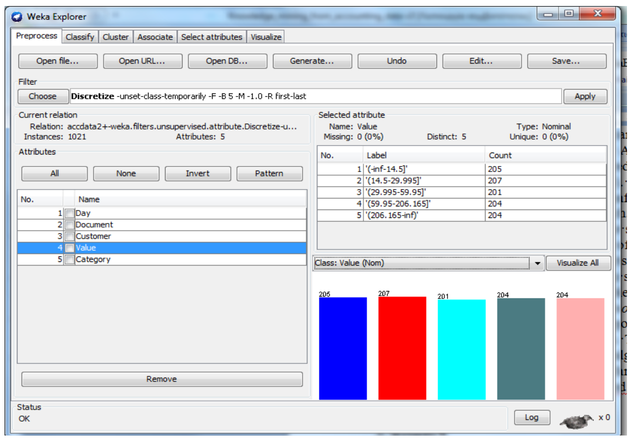

Fig. 4. Discretization results

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

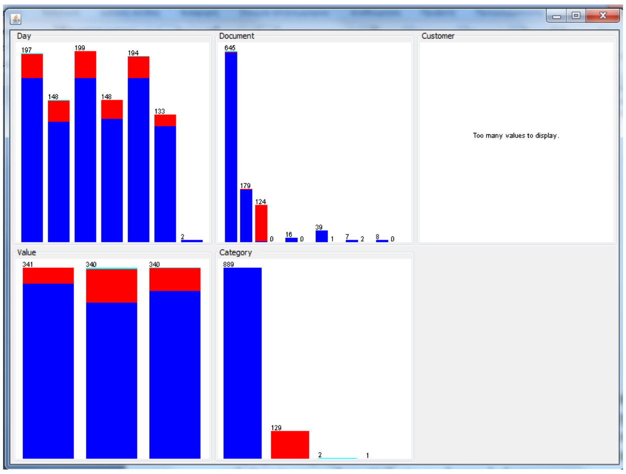

Fig. 5. Visualizing results

It is worth mentioning that the attribute customer cannot be visualized because there are many different values that cannot be grouped. Nevertheless, in the dataset it is possible to display the graphical representations of each attribute in relation to any other attribute as portrayed in fig. 6.

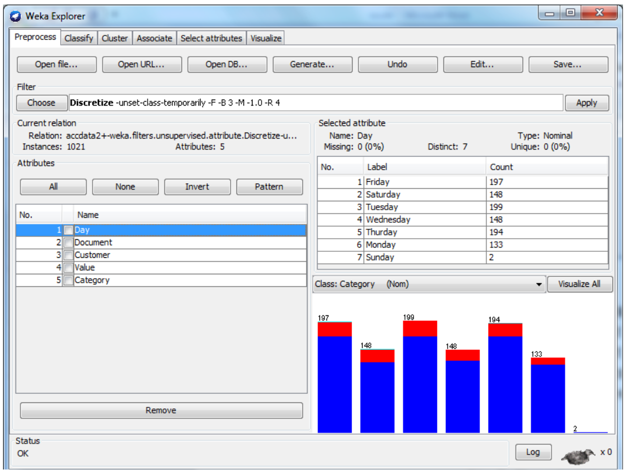

Fig 6. The attribute day in relation to Category (class)

3.2 Classification

The algorithm OneR is applied in the classification step. The attribute Category is used as class for classification. The overall results are presented in figure 7. The 1013 out of 1021 instances are correct with precision 99.2165%. The confusion matrix is portrayed in table 2.

3.3 Clustering

Kmeans algorithm is an iterative algorithm that divides a dataset into K pre-defined distinct non-overlapping subgroups named clusters. Each data point can belong to one group only. The data points are grouped to a cluster such that the sum of the squared distance between the data points and the cluster's centroid is at the minimum. The data points which belong to the same cluster have the less variation. Some implementations of K-means only allow numerical values for attributes [26]. Weka has an implementation of K-means which is called SimpleKmeans. SimpleKMeans algorithm can handle both categorical and numerical attributes. Furthermore, the numerical attributes are automatically normalized in distance computations. Thus, the use of filters is not prerequisite in the preprocessing tasks for clustering in WEKA. The default Euclidean distance is replaced by the Manhattan distance, so that the centroids will be computed as the component-wise median rather than mean [25]. The number of clusters is proposed to be 3. After clustering with SimpleKMeans, the clustered instances are partitioned in the three clusters as 490 (48%), 235 (23%), 296 (29%), items and percentages respectively.

```
=== Classifier model (full training set) ===

Document:
        TDA     -> A
        ALP     -> A
        APY     -> B
        PT-DP   -> A
        TPY     -> A
        ALP-EP  -> A
        ALP EP  -> A
        APY EP  -> B
        ALP X   -> A
        APY     -> A
        PT DP   -> A
        PTDP    -> A
(1014/1021 instances correct)


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1013               99.2165 %
Incorrectly Classified Instances         8                0.7835 %
Kappa statistic                          0.964
Mean absolute error                      0.0039
Root mean squared error                  0.0626
Relative absolute error                  3.5127 %
Root relative squared error             26.642  %
Total Number of Instances             1021

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.998     0.047     0.993      0.998    0.996      0.976    A
                 0.953     0.002     0.984      0.953    0.969      0.976    B
                 0         0         0          0        0          ?        A
                 0         0         0          0        0          ?        A
Weighted Avg.    0.992     0.041     0.992      0.992    0.992      0.976
```

Fig 7. Classification results with Algorithm OneR

Table 2. The confusion matrix

a	b	<--classified as
890	2	a = A
6	123	b = B

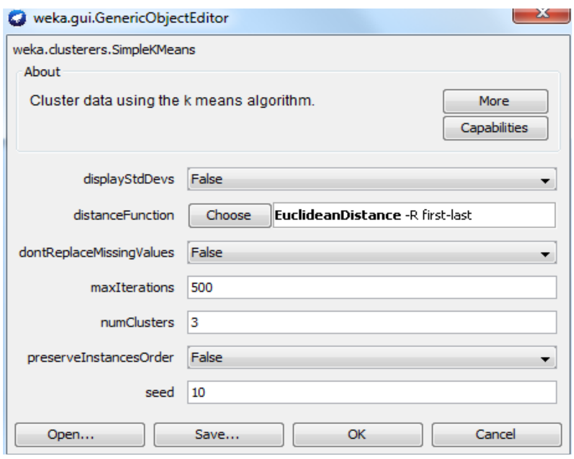

Fig. 8. Clustering Options

There are differences among the three clusters. We remark that for the attributes in the first cluster (0), the day is "Thursday", the Document is "TDA", the Value less than 22€, the Customer is "RETAIL CUSTOMERS". For the attributes in the second cluster (1), the day is "Saturday", the Document is "ALP", the Value greater than 80€ and the

Customer is "RETAIL CUSTOMERS". Finally, for the attributes in the third cluster (2), the day is "Friday", the Document is "TDA", the Value between 22€ and 80€ and the Customer "MIXIOYRHS ATHANASIOS K SIA OE". All the clusters concede in Category "A".

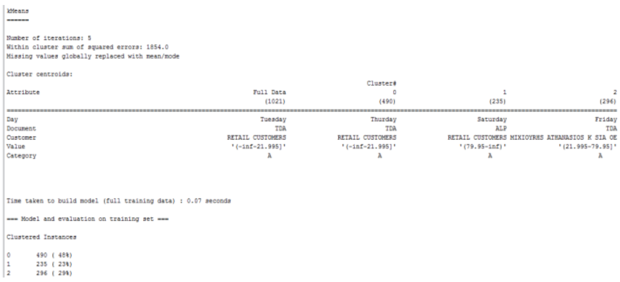

Fig. 9. Clustering with SimpleKMeans

The clusters of the data allow us to better assess the information contained in the dataset. Visualization of the results using the attributes day and value is portrayed in fig. 10. The grey points indicate low values whereas points in black show high values.

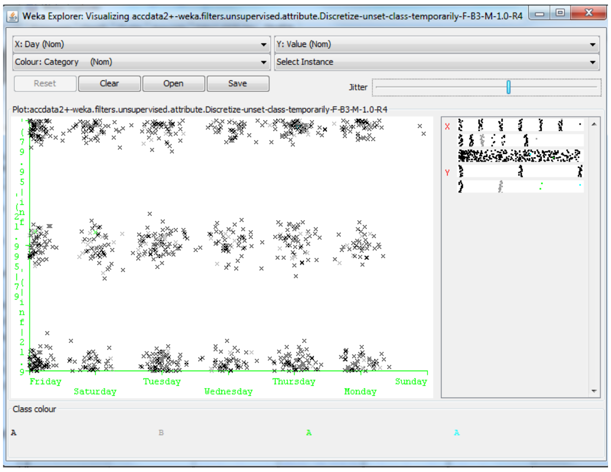
Fig. 10. Visualization of clusters based on day and value

Visualization of the results using the attributes category and value is also portrayed in fig. 11. The black points indicate category A while grey points indicate category B.

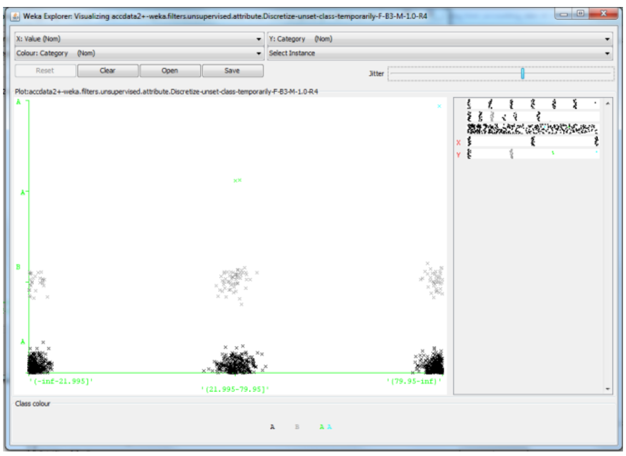
Fig. 11. Visualization of clusters based on day and category

3.4 Association rule mining

The Apriori algorithm [30] was used for finding association rules for our dataset. The WEKA produced a list of 15 rules (Table 3) 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). Table 3 presents a great number of association rules. They are ranked in descending order of the metric confidence. There are some uninteresting rules regarding the aim of the research, like rules 2, 6 and 11 which confirm existing "beliefs". For example, if the document is ALP then the Customer is RETAIL CUSTOMERS. There are some similar rules, rules with the same element (Customer=RETAIL CUSTOMERS) to appear in antecedent and consequent, such as rules 12 and 13. Another couple of similar rules is rules 9 and 10. The same element (Customer=RETAIL CUSTOMERS) is appeared in antecedent and consequent. There are rules which are subsets of other rules like rule 5 and 4, and 2 and 3, i.e. 4 includes 5 and 2 includes 3. The rules 7 and 15 reveal interesting information [30]. The rules 8 and 14 are symmetrical because the antecedent and the consequent elements are interchanged.

Table 3. The 15 best rules found with Apriori algorithm

Best rules found:
1. Document=TDA Value='(-inf-21.995]' 245 ==> Category=A 245 conf:(1)
2. Document=ALP 180 ==> Customer=RETAIL CUSTOMERS 180 conf:(1)
3. Document=ALP Category=A 178 ==> Customer=RETAIL CUSTOMERS 178 conf:(1)
4. Document=ALP Value='(79.95-inf)' 172 ==> Customer=RETAIL CUSTOMERS 172 conf:(1)
5. Document=ALP Value='(79.95-inf)' Category=A 171 ==> Customer=RETAIL CUSTOMERS 171 conf:(1)
6. Document=TDA 645 ==> Category=A 643 conf:(1)
7. Document=TDA Value='(21.995-79.95]' 273 ==> Category=A 272 conf:(1)
8. Document=ALP Value='(79.95-inf)' 172 ==> Category=A 172 conf:(1)
14. Category=A 172 ==> Document=ALP Value='(79.95-inf)' 172 conf:(1)
9. Document=ALP Customer=RETAIL CUSTOMERS Value='(79.95-inf)' 172 ==> Category=A 171 conf:(0.99)
10. Document=ALP Value='(79.95-inf)' 172 ==> Customer=RETAIL CUSTOMERS Category=A 171 conf:(0.99)
11. Document=ALP 180 ==> Category=A 178 conf:(0.99)
12. Document=ALP Customer=RETAIL CUSTOMERS 180 ==> Category=A 178 conf:(0.99)
13. Document=ALP 180 ==> Customer=RETAIL CUSTOMERS Category=A 178 conf:(0.99)
15. Value='(21.995-79.95]' Category=A 280 ==> Document=TDA 272 conf:(0.97)

4. Discussion and conclusions

This study provided an overview of issues related to data analysis of accounting data using the most popular data mining techniques such as clustering, classification and association rule mining. A case study of an existing business was used for the implementation of the aforementioned data mining techniques with the use of the open source software package WEKA.

In the classification step, the algorithm OneR was applied with the use of the attribute Category as class for classification. The attribute Document proved to be dependent on the Category. In the clustering step, three clusters of data were revealed with their specific characteristics. It is worth mentioning that customers with valued greater than 80€ were observed on Saturday and the

Document was "ALP". In association rule mining, there were revealed 15 rules. Rules 7 and 15 show interesting information. If the Document is "TDA" and the Value is between 22€ and 80€ the Category is "A".

The overall outcome may support the decision making process regarding the customers and their transactions. The findings confirm previous knowledge though offer a lot of actionability for a middle size company. The same techniques could be also applied to other companies of this size. Nevertheless, it is hardly possible to conduct a study that does not contain weaknesses or an element of bias. Therefore, there are some limitations in this study as well. More specifically, the sample size is relatively small (1021 transactions) and may not be adequate to generalize the results. This research, despite its limitations, can contribute, to the existing knowledge. Moreover, it is important to repeat this study to other companies in the future.

References

1. 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.
2. Adamyk, O., Adamyk, B., & Khorunzhak, N. (2018). Auditing of the Software of Computer Accounting System.
3. Kiesow, A., Zarvic, N., & Thomas, O. (2014). Continuous Auditing in Big Data Computing Environments: Towards an Integrated Audit Approach by Using CAATTs. In *GI-Jahrestagung* (pp. 901-912).
4. Antipova, T., & Rocha, Á. (Eds.). (2018). *Information Technology Science*. Springer International Publishing.
5. Henry, E., & Robinson, T. R. (2009). Financial Statement Analysis: An Introduction. International Financial Statement Analysis.
6. Bellino, C., Wells, J., & Hunt, S. (2007). Global Technology Audit Guide (GTAG) 8: Auditing Application Controls.
7. Zhang, J., Yang, X., & Appelbaum, D. (2015). Toward effective Big Data analysis in continuous auditing. *Accounting Horizons*, 29(2), 469-476.
8. Lientz, B., & Larssen, L. (2012). *Manage IT as a Business*. Routledge.
9. Appelbaum, D., Kogan, A., & Vasarhelyi, M. A. (2017). An introduction to data analysis for auditors and accountants. *CPA Journal*, *87*(2), 32-37.
10. Schaltegger, S., & Burritt, R. (2017). *Contemporary environmental accounting: issues, concepts and practice*. Routledge.
11. Socea, A. D. (2012). Managerial decision-making and financial accounting information. *Procedia-Social and Behavioral Sciences*, *58*, 47-55.
12. Gelinas, U. J., Dull, R. B., Wheeler, P., & Hill, M. C. (2017). *Accounting Information Systems*. Cengage Learning.
13. Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.
14. Chen, M. S., Han, J., & Yu, P. S. (1996). Data mining: an overview from a database perspective. *IEEE Transactions on Knowledge and data Engineering*, *8*(6), 866-883.
15. Zhao, Z. A., & Liu, H. (2011). *Spectral Feature Selection for Data Mining (Open Access)*. Chapman and Hall/CRC.
16. Kantardzic, M. (2003) *Data Mining: Concepts, Models, Methods, and Algorithms*. New York, NY: John Wiley & Sons.
17. Sharma, A., & Panigrahi, P. K. (2013). *A review of financial accounting fraud detection based on data mining techniques*. arXiv preprint arXiv:1309.3944.
18. 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.
19. 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.
20. 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.
21. 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.
22. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.
23. Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine learning, 11*(1), 63-90.
24. Theodosiou, T., Valsamidis, S., & Hatziliadis, G. (2011). Application of Data Mining Techniques to Olea europaea var. media oblonga production from Thassos Island.
25. Valsamidis, S., Kontogiannis, S., Kazanidis, I., & Karakos, A. (2011). E-learning platform usage analysis. *Interdisciplinary Journal of E-Learning and Learning Objects*, *7*(1), 185-204.
26. Kaufmann, L. and Rousseeuw, P.J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*, New York, John Wiley & Sons.
27. 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.
28. Linoff, G. S., & Berry, M. J. (2011). *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons.
29. Ma, B. L. W. H. Y., & Liu, B. (1998). Integrating classification and association rule mining. In *Proceedings of the fourth international conference on knowledge discovery and data mining*.
30. Agrawal R., Imielinski, T. and Swami, A.N. (1993). Mining Association Rules between Sets of Items in Large Databases. *In Proc. of SIGMOD*, 207-216.
31. Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. *Proceedings of 20th International Conference on Very Large Data Bases* (pp. 487-499).
32. Minaei-Bidgoli, B., Tan, P-N. & Punch, W.F. (2004). Mining Interesting Contrast Rules for a Web-based Educational System. *Proceedings of Int. Conf. on Machine Learning Applications*, Louisville, USA 2004 (pp. 320- 327).
33. Liu, B. and Hsu, W. (1996). Post-Analysis of Learned Rules. *Proceedings of National Conference on Artificial Intelligence*. Portland, Oregon, USA, (pp. 828–834).
34. Liu, B., Hsu, W., Chen, S. and Ma, Y. (2000). Analyzing the Subjective Interestingness of Association Rules. *IEEE Intelligent Systems*, 15(5), 47–55.