

Detection of Fraud Patterns in Accounting Accounts Using Data Mining Techniques

Alexander Báez Hernández¹, Debrayan Bravo Hidalgo²

¹Administrative Sciences Department, Universidad Central del Ecuador, Quito, Ecuador

²Department of Electrical Engineering and Electronics, Universidad San Francisco de Quito (USFQ), Quito, Ecuador

Email: albarez@uce.edu.ec, dbrayanbh@gmail.com

How to cite this paper: Hernández, A. B., & Hidalgo, D. B. (2020). Detection of Fraud Patterns in Accounting Accounts Using Data Mining Techniques. *Open Journal of Business and Management*, 8, 1609-1618. <https://doi.org/10.4236/ojbm.2020.84102>

Received: June 4, 2020

Accepted: July 17, 2020

Published: July 20, 2020

Copyright © 2020 by author(s) and Scientific Research Publishing Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Accounting databases with fraudulent transactions inside was used to detect fraud patterns by data mining tool. The object was accomplished by the following method: first, inside data, fraudulent transactions according to three fraud patterns were settled, over it the algorithms, Euclidian distance and local outlier factors were run using Rapidminer program. As a result the fraud patterns were shown in different ways according to the specific graphics contributed by the program. In conclusion, clusters grouping by Euclidian distance with k Means algorithm ($k = 4$) allowed an adequate visualization of the values' distribution, as consequence was detected the first and third fraud patterns. The application of the outlier's detection algorithm (LOF) detected the three fraud patterns in a clear way as a consequence of the insolate outliers in different graphics, shown by Rapidminer program, with different variables correlated.

Keywords

Recognition, Financial Fraud, Tool, Data Mining

1. Introduction

Accounting data analysis has become a necessary process for fraud detecting. It is a currency problem that has increased the interest in accounting researching (Seo, Choi, Choi, Lee, & Lee, 2009). Fraud Detection and control have raised relevance in the audit process (Westhausen, 2017) where the informatics program is taking more presence as a tool capable to perceive the deception in accounting data (Gottschalk & Solli-Saether, 2010).

Data mining is one of the most important current models of modern intelligent business analysis and decision support tools. This relevance is accepted by

the transcendental professional accounting organizations. The American Institute of Certified Public Accountants (AICPA) has identified data mining as one of the top ten technologies going forward, and the Institute of Internal Auditors (IIA) has listed data mining as one of the top four research priorities (Koh & Low, 2004). Additionally, Chartered Global Management Accountants (CGMA) has reported that N 50% of corporate leaders rank big data and data mining among the top ten corporate priorities that are central to the data-driven business era (CGMA, 2013). Data mining has been defined as the process of identifying valid, potentially novel, and ultimately understandable patterns in the data (Pujari, 2001). It is also known as the process of extracting knowledge from massive amounts of data (Han et al., 2006) to increase the efficiency of decisions in a particular discipline. The key approach to data mining, therefore, is to leverage an organization's data assets for financial or non-financial benefits. Therefore, data mining has been applied to almost all commercial and non-commercial disciplines, including accounting.

Data mining has given the possibility to analyze big data where some elements have been considered in untypical way or behavior in its input or manage as a support of traditional making decisions (Mraović, 2008). It has turned in strong tool to analyze the accounting data (Debreceeny & Gray, 2010; Joudaki et al., 2015; Qin, 2014; Wu, Ou, Lin, Chang, & Yen, 2012). Business knowledge, data interpretation and logical reason to apply analytic method to build the data are embraced with data mining analysis (Jackson, 2002). A lot of algorithms for different analyses are possible in data mining (Wu et al., 2008). A broad scope of application in accounting science was resumed by Amani & Fadlalla (2017) in the following statements:

- 1) Data classification and mapping to predefine qualitative attributes in itself
- 2) Data clustering in clusters with specific significance
- 3) Focus in predictive approach to find future numerical values or a classification of no numerical values
- 4) Outliers detection, indeed values that are out of range
- 5) Optimization, the best solution choice for a means set
- 6) Visualization, adequate image for the best data comprehension
- 7) Regression, estimation of a dependent variable from a set of independent variables.

Outliers' detection was done for two different algorithms:

- 1) Outliers detection considered the distance it locates higher atypical values in a data set no labeled to get subset of it. This subset is denominated outliers solution set (Angiulli, Basta, & Pizzuti, 2006).

- 2) Detecting outlier in a data set considered a Local Outlier Factor (LOF). El LOF is focused on local density where the location is expressed for more close K neighbors whose distance is used to estimate the density. When it compares the local density from object with the local density of their neighbors it possible to identify areas with similar density and point with minor density that its neigh-

bors. It is outliers (Breunig, Kriegel, Ng, & Sander, 2000). First data classification used a common clustering suggested by Abonyi & Feil (2007) and K middle algorithmic considered of frequency utilization for this purpose, according to Likas, Vlassis, & Verbeek (2003). Detecting of fraud pattern in accounting data with outlier data mining algorithmic was the objective.

2. Methods

Data base was made with real assent of a month from enterprise accounting bank. Inside of it was settling transactions from three-fraud pattern.

- 1) A little Withdraws (credit) of 0.05 cents many times.
- 2) Withdraws (credit), a constant amount with regular period of time.
- 3) A high quantitative withdraw (credit), more than average account entries.

RapidMiner program was used for Data Analysis according to the argument from (Amer & Goldstein, 2012; Hofmann & Klinkenberg, 2013; Jungermann, 2009). In the process, RapidMiner program reads first the database from Excel and selects (filter) the real values to avoid null values in attributes analyzed in the algorithmics as first step. Continued was applied the algorithmic process, Euclidian distance and local outlier factors as is showed in Figure 1 and Figure 2.

The outliers were determined by operator using LOF algorithmic and local density in it (Breunig et al., 2000).

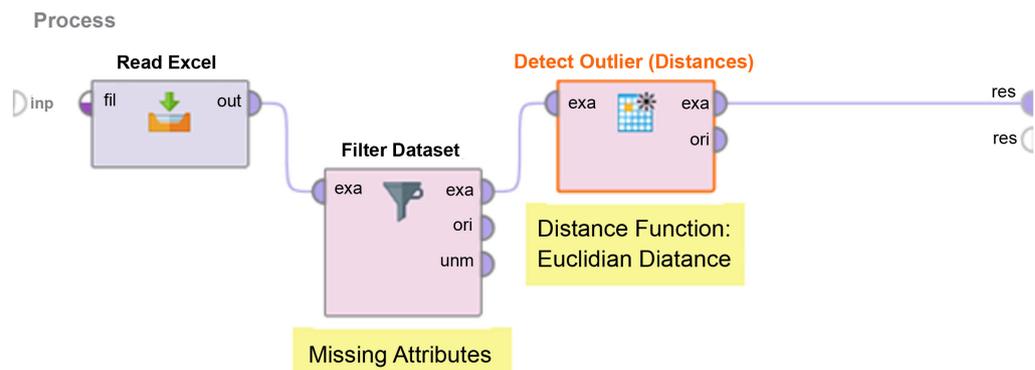


Figure 1. Clustering by distance function: Euclidian distance.

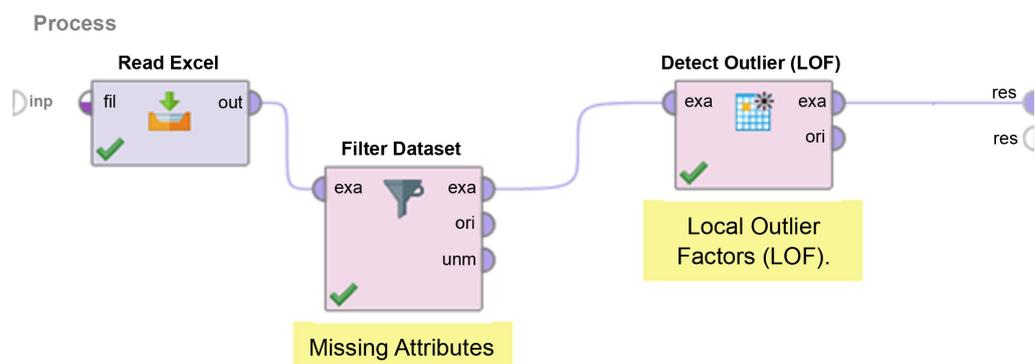


Figure 2. Local outlier factors (LOF) detection.

3. Result and Discussion

3.1. Bibliometrics Analysis of Detection of Fraud Pattern in Accounting Accounts Using Data Mining Techniques

Using as search criteria “detection of fraud and data mining” in the title, abstract, and the keywords of the scientific contributions in the Scopus academic directory. 921 results were detected between Conference Paper, Article and Review. **Figure 3** shows the quantitative evolution that this line of research has undergone in recent years. The trend of the curve is increasing, in this figure. It is evident that the international scientific community considers fraud detection techniques through data mining as an effective and practical method. This figure was made using the bibliometric analysis tools provided by Scopus.

The bibliographic information of each of the 921 detected documents was imported from the Scopus platform, in (.ris) format. This process allowed this data to be processed in the VOSviewer science bibliometric analysis software. This computational tool allows building and visualizing bibliometric networks of scientific activity in journals, researchers or publications. As well as generating maps of scientific activity based on citations, bibliographic coupling, co-citations or authorship relationships, among others. The thermos research limitations density map shows the concentration of the most recent terms in a given research area. **Figure 4** shows a term density map that shows which are the most frequent terms and their relationship in research in this area of knowledge. This type of figures allows a very clear idea of current research trends and trends in the subject in question.

3.2. Detection of Fraud Pattern in Accounting Accounts Using Data Mining Techniques

- Clustering

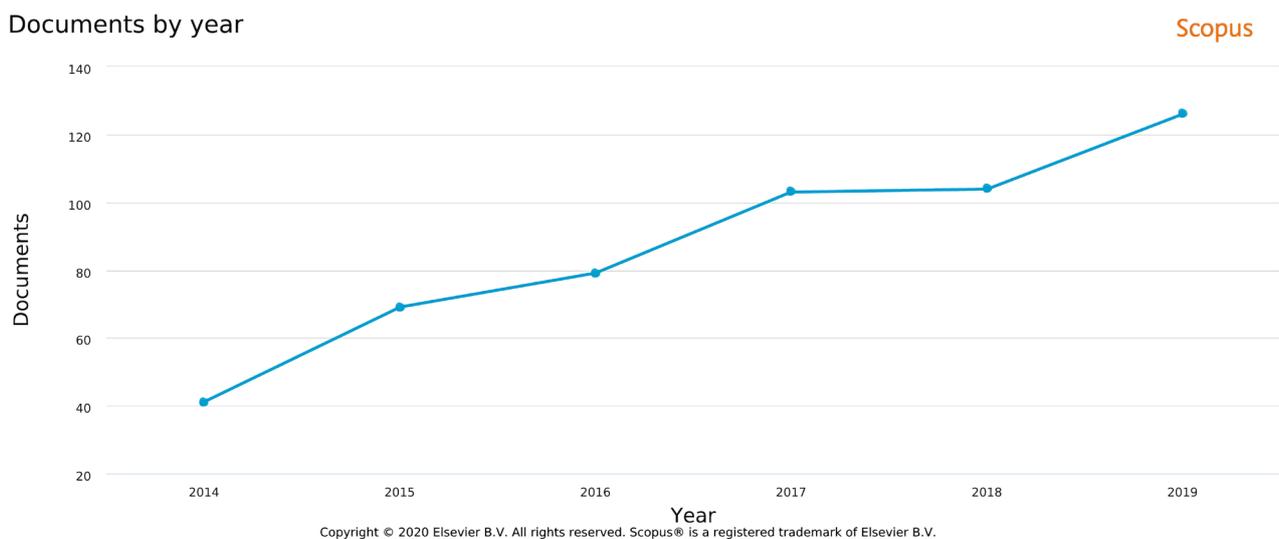


Figure 3. Number of scientific contributions per year in the subject area of fraud detection and data mining.

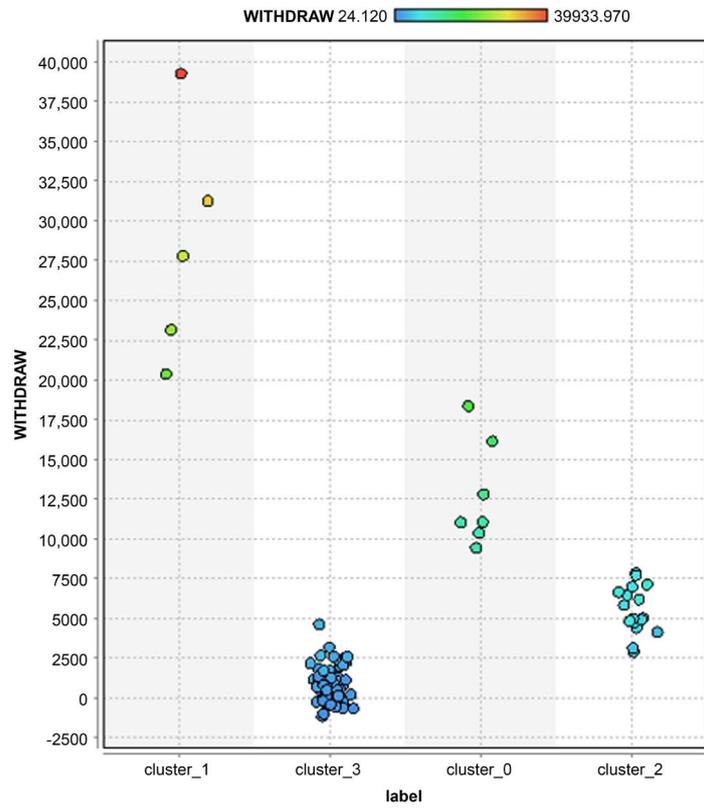


Figure 5. Clustering withdraw.

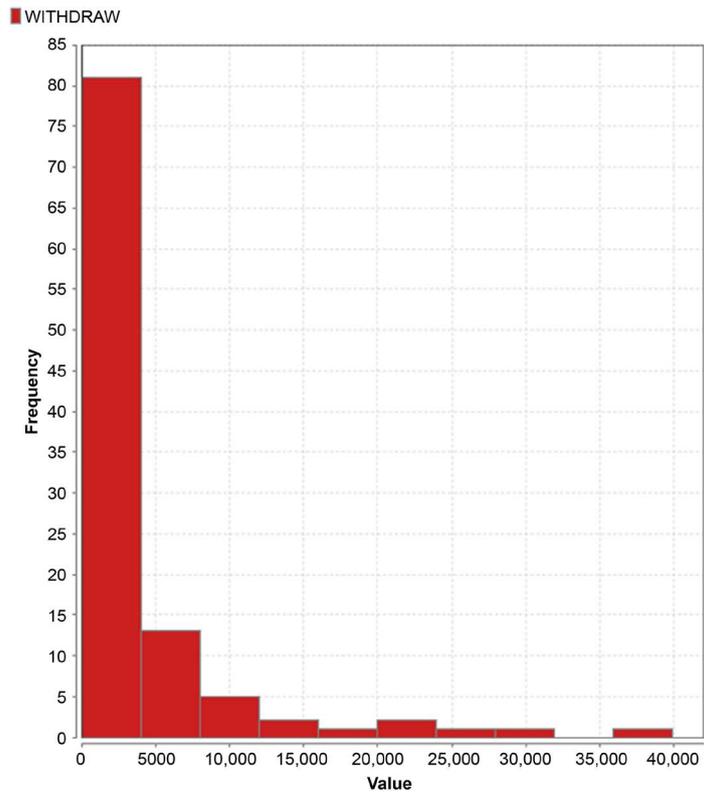


Figure 6. Withdraw frequency distribution.

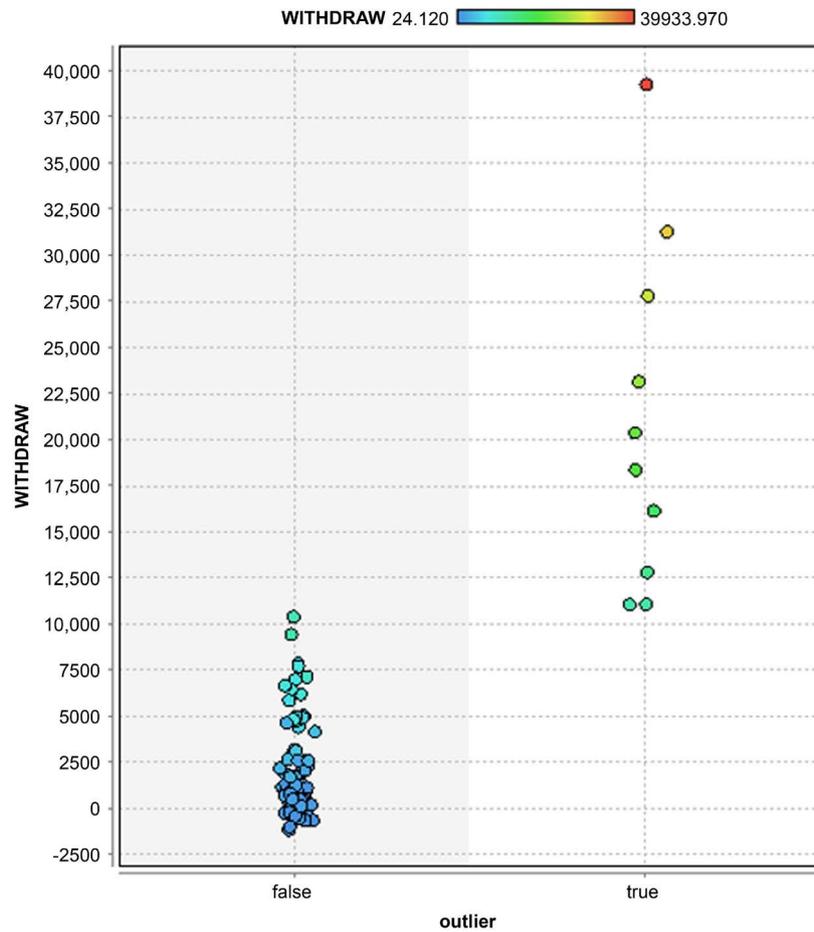


Figure 7. Detection using LOF algorithmic.

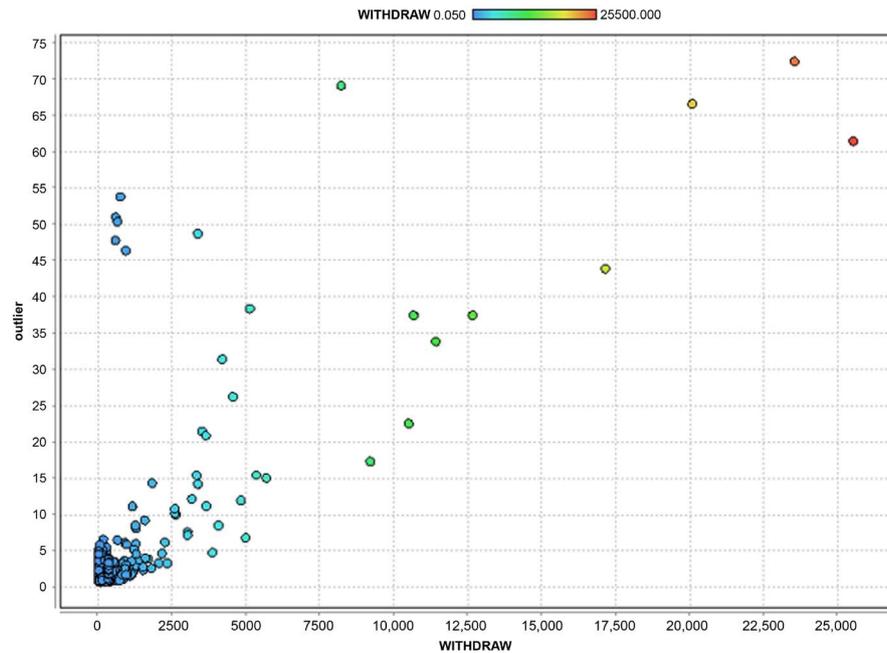


Figure 8. Outlier value distribution using LOF.

irregular entries settle in accounts. The matter is that available information per day gives a temporality sequence in the way that the fraud happened to building architecture used in its committed. It is visible in **Figure 9** clusters of transactions in dark color blue concentrate in specific days. This indicated a suspicious regular performance in the accounting operations associated with time. It's a way that detected the second pattern as a paradox: an irregular proceeding committed as shady regular accounting record correlated with time period.

Research limitations:

- Bibliographic references you will consider are from the Scopus database.
- Only manuscripts published in the English language are considered.
- Database was made only with real assent of a month from enterprise accounting bank.

4. Conclusion

The application of Data Mining using the detection of Outliers (both Distance Algorithm and Local Outliers Factor) made it possible to detect (fraudulent) patterns injected into an accounting database that simulated three possible types of fraud:

Table 1. Centroids calculation for withdrawing.

Attribute	Cluster 0	Cluster 1	Cluster 2	Cluster 3
WITHDRAW	12420.83	28537.636	5679.66	882.63

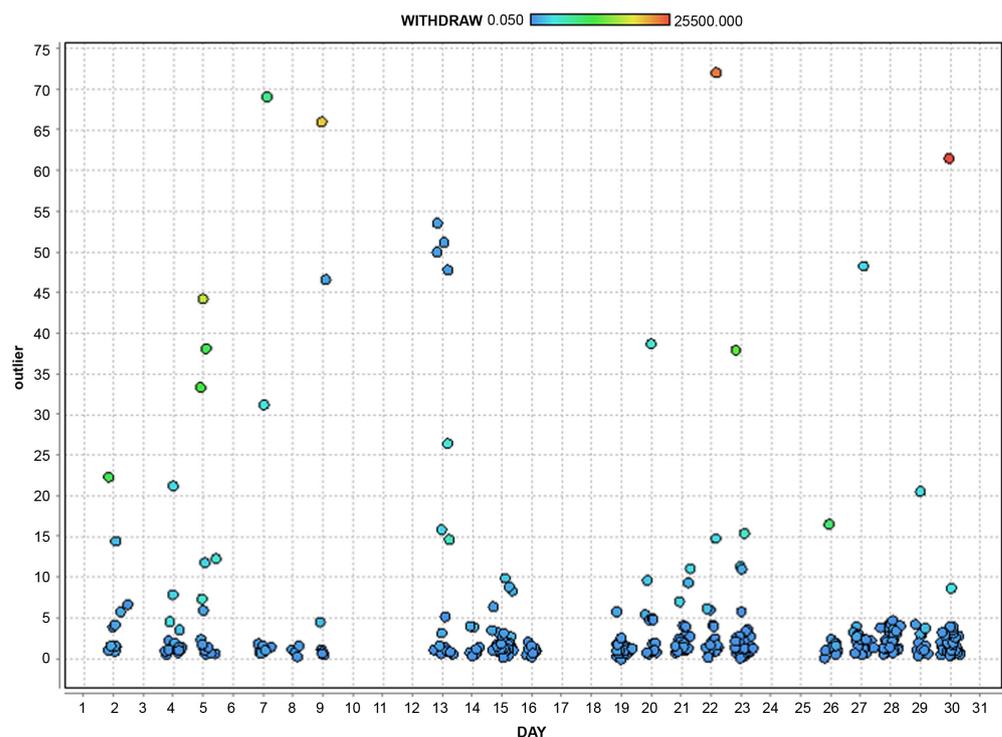


Figure 9. Daily outliers distribution.

- 1) A little Withdraws (credit) of 0.05 cents many times.
- 2) Withdraws (credit), a constant amount with a regular period of time.
- 3) A high quantitative withdraw (credit), more than average account entries.

The detection of Outliers offered the following possibilities:

The clusters grouping indicated, in a simple way, the possible movements' distribution for auditors. The use of the k Means algorithm ($k = 4$) allowed an adequate visualization of the values' distribution. It has detected the pattern third in cluster one, and first pattern in cluster three (Figure 5. Clustering Withdraw). The application of the outlier's detection algorithm (LOF) detected, Figure 6 and Figure 7, the first and third pattern in a clear way as a consequence of the insolate outliers (true and false) or transactions density (high and low) Daily Outliers distribution, Figure 5, detected all the patterns. It's the only that remarked the second pattern as an expression of a paradoxid result: a regular transaction correlated with time variable is an irregular procedure in accounting data. It's the core idea used to detect the second pattern.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

- Abonyi, J., & Feil, B. (2007). *Cluster Analysis for Data Mining and System Identification*. Basel: Birkhäuser.
- 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. <https://doi.org/10.1016/j.accinf.2016.12.004>
- Amer, M., & Goldstein, M. (2012). Nearest-Neighbor and Clustering Based Anomaly Detection Algorithms for RapidMiner. *Proc. of the 3rd RapidMiner Community Meeting and Conference (RCOMM 2012)* (pp. 1-12).
- Angiulli, F., Basta, S., & Pizzuti, C. (2006). Distance-Based Detection and Prediction of Outliers. *IEEE Transactions on Knowledge and Data Engineering*, 18, 145-160. <https://doi.org/10.1109/TKDE.2006.29>
- Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: Identifying Density-Based Local Outliers. *ACM SIGMOD Record*, 29, 93-104.
- Chartered Global Management Accountant (CGMA) (2013). *Report from Insight to Impact: Unlocking Opportunities in Big Data*. http://www.cgma.org/Resources/Reports/DownloadableDocuments/From_insight_to_impact-unlocking_the_opportunities_in_big_data.pdf
- Debreceny, R. S., & Gray, G. L. (2010). Data Mining Journal Entries for Fraud Detection: An Exploratory Study. *International Journal of Accounting Information Systems*, 11, 157-181. <https://doi.org/10.1016/j.accinf.2010.08.001>
- Gottschalk, P., & Solli-Saether, H. (2010). Computer Information Systems in Financial Crime Investigations. *Journal of Computer Information Systems*, 50, 41-49.
- Han, J., Kamber, M., & Pei, J. (2006). *Data Mining: Concepts and Techniques*. Burlington, MA: Morgan Kaufmann.

- Hofmann, M., & Klinkenberg, R. (2013). *RapidMiner: Data Mining Use Cases and Business Analytics Applications*. Boca Raton, FL: CRC Press.
- Jackson, J. (2002). Data Mining; A Conceptual Overview. *Communications of the Association for Information Systems*, 8, 19. <https://doi.org/10.17705/1CAIS.00819>
- Joudaki, H., Rashidian, A., Minaei-Bidgoli, B., Mahmoodi, M., Geraili, B., Nasiri, M., & Arab, M. (2015). Using Data Mining to Detect Health Care Fraud and Abuse: A Review of Literature. *Global Journal of Health Science*, 7, 194. <https://doi.org/10.5539/gjhs.v7n1p194>
- Jungermann, F. (2009). Information Extraction with Rapidminer. *The Proceedings of the GSCL Symposium Sprachtechnologie und eHumanities, TU Dortmund*.
- Koh, H. C., & Low, C. K. (2004). Going Concern Prediction Using Data Mining Techniques. *Managerial Auditing Journal*, 19, 462-476. <https://doi.org/10.1108/02686900410524436>
- Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The Global k-Means Clustering Algorithm. *Pattern Recognition*, 36, 451-461. [https://doi.org/10.1016/S0031-3203\(02\)00060-2](https://doi.org/10.1016/S0031-3203(02)00060-2)
- Mraović, B. (2008). Relevance of Data Mining for Accounting: Social Implications. *Social Responsibility Journal*, 4, 439-455. <https://doi.org/10.1108/17471110810909858>
- Pujari, A. K. (2001). *Data Mining Techniques*. Bangalore: Universities Press.
- Qin, B. Z. (2014). Study on the Financial Audit Risk Prevention Based on Data Mining Technology. *Applied Mechanics and Materials*, 608-609, 351-354. <https://doi.org/10.4028/www.scientific.net/AMM.608-609.351>
- Seo, K., Choi, J., Choi, Y., Lee, D., & Lee, S. (2009). Research about Extracting and Analyzing Accounting Data of Company to Detect Financial Fraud. *IEEE International Conference on Intelligence and Security Informatics* (pp. 200-202). <https://doi.org/10.1109/ISI.2009.5137302>
- Westhausen, H.-U. (2017). The Escalating Relevance of Internal Auditing as Anti-Fraud Control. *Journal of Financial Crime*, 24, 322-328. <https://doi.org/10.1108/JFC-06-2016-0041>
- Wu, R. S., Ou, C. S., Lin, H. Y., Chang, S. I., & Yen, D. C. (2012). Using Data Mining Technique to Enhance Tax Evasion Detection Performance. *Expert Systems with Applications*, 39, 8769-8777. <https://doi.org/10.1016/j.eswa.2012.01.204>
- Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., Philip, S. Y. et al. (2008). Top 10 Algorithms in Data Mining. *Knowledge and Information Systems*, 14, 1-37. <https://doi.org/10.1007/s10115-007-0114-2>