Proceedings of the 4th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS'18)
Madrid, Spain – August 21 – 23, 2018
Paper No. MVML 104
DOI: 10.11159/mvml18.104

A Model Based on Clustering and Association Rules for Detection of Fraud in Banking Transactions

Mehrdad Kargari, Abdollah Eshghi
Tarbiat Modares University
Jalal AleAhmad Nasr, Tehran, Iran
M_kargari@modares.ac.ir; a.eshghi@modares.ac.ir

Abstract **–** In recent years, fraud in banking transactions has turned into a serious problem for which different supervised and unsupervised algorithms have been suggested. In this paper, a semi-supervised combined model based on clustering algorithms and association rule mining is devised in order to detect frauds and suspicious behaviors in banking transactions. To this end original and non-fraud transaction data of the customers is collected for the analysis. Next, repetitive patterns of customer behaviors are extracted through association rules and used as normal rules so that any new transaction must conform to at least one of these rules. In behavior analysis component, a fuzzy clustering algorithm is employed to extract the normal behavior patterns of customers. Abnormal transactions belong to none of these clusters and will be recognized as high risk. The final understanding of a transaction will be gained through combining the results of association rules and clustering patterns. Findings suggest that the employment of both rule-based and clustering-based components leads to the detection of more frauds while fewer alarms will go off.

Keywords**:** Fraud Detection, Clustering, Association Rules, Apriori, Cmeans.

1. Introduction

The application of electronic banking services has grown significantly over the last three years [1]. These services are prone to frauds to the extent that an approximation of 60 percent of the frauds have been done through electronic channels like mobile and internet banking [2].

Fraud, as an unpleasant event, has a negative impact in the progress of every industry and community and the impact is especially more damaging in the banking industry. Statistics reveal that other than a remarkable loss of profit due to fraud, such a criminal deception is also an important deterrent for customers using banking electronic services in countries in which electronic commerce is in its initial state [3]. Although a lot of effort has been put into fighting against fraud in recent years and several machine learning and data mining methods have been used, the ever-changing behaviors of customers and fraudsters make it impossible to completely model the normal behaviors of customers or eradicate the frauds [4]. Moreover, applying static machine learning methods without the ability to reconcile with new strategies is not really useful for fraud detection either [5].

Some of the challenges and drawbacks of fraud detection which are mentioned in the literature [6]–[8] include lack of valid and supervised datasets for research, imbalanced data, costs of using fraud detection systems, and the ever-changing behaviors of customers.

Another main challenge of improving fraud detection algorithms is that the knowledge sharing in this field is not very impressive and while there exist a lot of studies and published papers on fraud detection, very few of them have claimed to be used in practice [9].

Lack of supervised datasets for extracting fraud detection models has rendered the role and importance of unsupervised methods more obvious. In this paper, a combined semi-supervised method using clustering and association rules is introduced and suspicious transactions which may not be done by the card owners are identified based on the customers' previous transactions and the historical normal trend of their behaviors.

2. Literature Review

There are three main categories of algorithms for fraud detection, namely supervised, unsupervised and semi-supervised methods [10]. Lack of supervised datasets on the one hand, and the shortcomings of supervised algorithms on the other hand have led researchers to pay more attention to unsupervised and semi-supervised methods in recent years [10].

When it comes to the unsupervised algorithms, clustering is of considerable importance. K-means is a clustering algorithm and has been used in many other studies [11]–[13]. It is also a method for vector quantization and is particularly deployed in case of detecting abnormal and fraudulent behaviors [14]. In [15]–[17], the hierarchical clustering methods were applied for fraud detection. DBSCAN which is a density-based clustering method [18] was used by Panigrahi et al. [19] for fraud detection. They used a set of attributes like the amount of transaction, billing and shipping addresses as well as an inter-transaction time gap for generating a set of normal clusters from the customer's credit card transactions. If a recently-arrived transaction did not belong to any of the generated clusters, then it was regarded as an outlier transaction. The extent of deviation of an incoming transaction from the normal clusters was used for determining the degree of outlierness. In [20], an improved k-means clustering method was employed and outlier detection was used as a method for detecting frauds. According to SAS report [20], in addition to the rule-based methods, an analytical component is also needed in any fraud detection system. This analytical component is based on unsupervised or semi-supervised algorithms. While known frauds can be detected by rule-based components which are supervised methods, new and unknown frauds are detectable by unsupervised or semi-supervised methods. In [21], a semi-supervised method was introduced through non-fraud transactions. In this method, some rules were extracted by Apriori algorithm and association rules. Then, by applying the rules to non-fraud transactions, those rules which identified non-fraud transactions as fraud transactions were removed. The remaining rules were deployed for monitoring the system and replicated rules were generated by making some changes to them. This method was used to take action for internal fraud detection. Normal profiles of customers are used for outlier detection in [22]. In [23], a semi-supervised neural network method was put forth. Its neural network consisted of a hidden layer and the same number of input and output neurons. Outlier detection, classification and clustering methods were discussed in [24].

3. Clustering and Association Rules Methods

Clustering is an unsupervised method for learning in which samples are divided into meaningful classes called clusters. These clusters are always homogeneous. The samples inside a cluster bear a strong resemblance to each other whereas those from different clusters are markedly different. The process of clustering is composed of several steps which are depicted in **Error! Reference source not found.**. As the figure clearly demonstrates, the feedback cycle improves the clustering results incrementally.

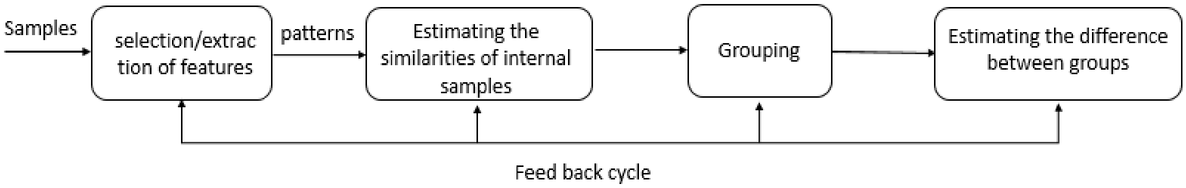
Samples
selection/extrac tion of features
patterns
Estimating the similarities of internal samples
Grouping
Estimating the difference between groups
Feed back cycle

Fig. 1: Main steps in clustering [25].

Clustering algorithms are divided into two main categories: partitional algorithms and hierarchical algorithms. Hierarchical algorithms are based on a single or complete link and partitional algorithms are divided into square error-based algorithms, graph theoretic-based algorithms, mixture resolving-based algorithms or mode seeking-based algorithms [25]. The k-means algorithm, which is used in this paper, is a square error-based algorithm.

Association analysis or affinity analysis is one of the data mining methods which studies the attributes of characteristics that "go together" [26]. This algorithm attempts to uncover rules for quantifying the relationship between two or more attributes. Association rules take the form of "IF antecedent, THEN consequent", along with a measure of the support and confidence [26]. For example, in fraud detection, a rule can be defined as: "IF a male person younger than age 1 is engaged in a transaction in time t1 through a channel like ch1, THEN the transaction amount must be less than amnt1". This rule means that the set of attributes like "attributes= {gender, age, time, channel, amount}" always come together with values like "values= {male, age1, t1, ch1, amnt1}". If the values of a transaction are in accordance with this rule, then the transaction can be regarded as normal and vice versa. There are different algorithms for association rules including Apriori, Eclat, and FP-growth. In the present paper, Apriori algorithm is employed.

4. Research Methodology

The proposed model in this paper is composed of two main components: rule-based component and trend analysis-based component. The overall structure of the proposed model is depicted in Fig. 1.

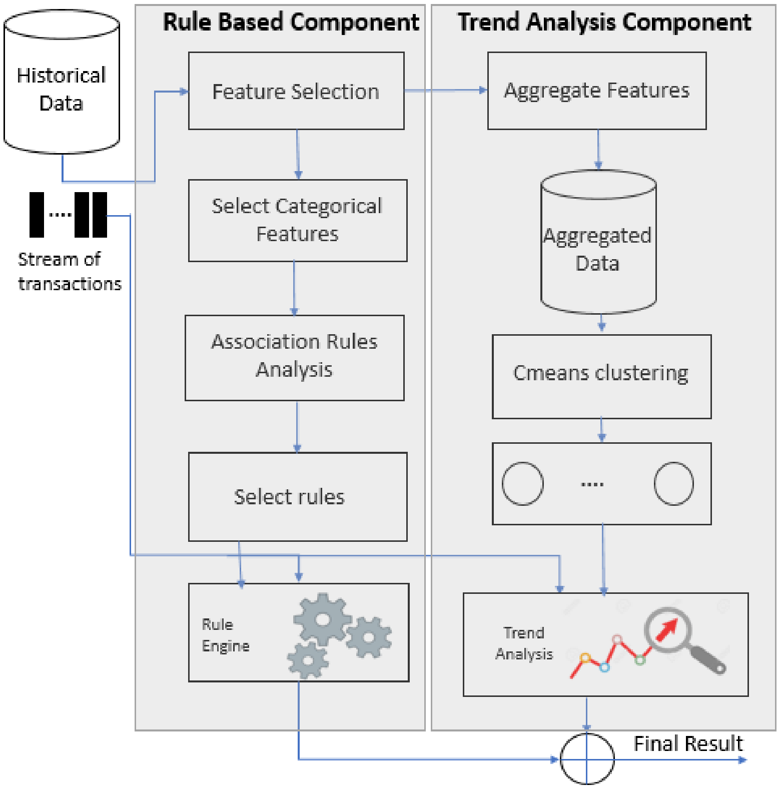

Fig. 1: The overall structure of the proposed model.

Each transaction will pass through both components and the result of each component is calculated independently. The final result will be obtained by combining the results of these two components. The result of each component is either "low risk" or "high risk" and the final result is either "low risk", "medium risk" or "high risk" (**Error! Reference source not found.**).

Table 1: Risk of each component and the combined (Final) risk.

Rule Based Component	Trend Analysis Component	Final Result
Low Risk	Low Risk	Low Risk
Low Risk	High Risk	Medium Risk
High Risk	Low Risk	Medium Risk
High Risk	High Risk	High Risk

One of the first and also main steps in fraud detection is selecting the most efficient attributes [2]. Selected attributes are either the main attributes which are available in transactions or derived attributes which are extracted by aggregation or statistical methods. The main attributes used in this paper are listed in Table 1.

Table 1: Main attributes.

	Attribute name	Description
1	Transaction ID	The unique ID of a transaction
2	Time	Time of a transaction
3	Account number	The account number to which a transaction belongs
4	Card number	The number of the card to which a transaction belongs
5	Transaction type	Type of a transaction (ATM, POS, Internet , …)
6	Entry mode	Card-present mode or card-not-present mode
7	Amount	Amount of a transaction
8	Merchant code	Code of a merchant
9	Merchant group	Guild of the merchant
10	Gender	Gender of the card owner (male or female)
11	Age	Age of the card owner
12	Bank	The bank to which the card belongs

Derived attributes are not available in a transaction directly and must be calculated. These attributes can be simple or aggregated. Aggregated attributes play significant roles in analyzing behavior trends and also in fraud detection algorithms [2], [27]. Aggregated attributes are numerical attributes revealing the trend of a transaction's main attributes in a period of time (hourly, daily, monthly, yearly, etc.). In this paper, the aggregated amount and number together with the maximum amount and number of transactions in 1 hour, 3 hours, 12 hours, 24 hours, 1 week and one month in different channels (ATM, POS …) are calculated and used for analysis. The extraction of aggregated features in through different channels in different periods of time is shown in Fig. 2.

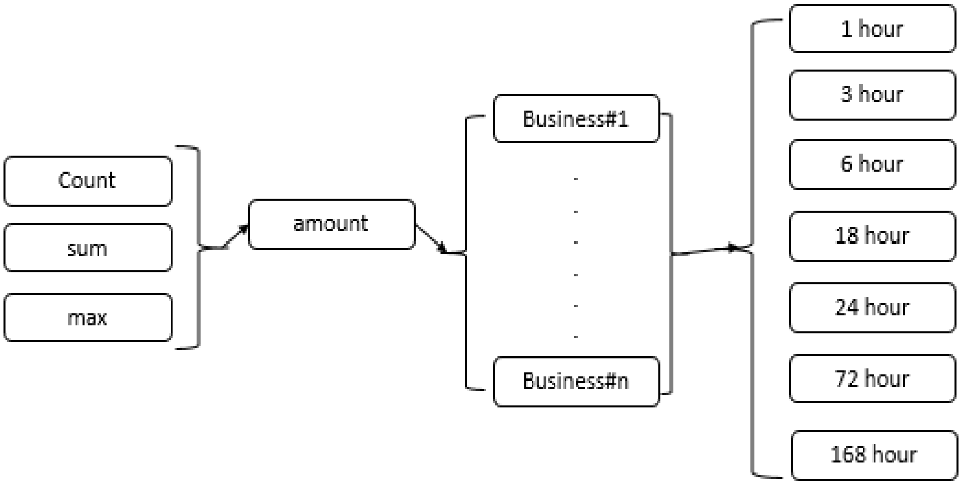

Fig. 2: The extraction of aggregated features in through different channels in different periods of time.

4.1. Rule Based Component

As mentioned earlier, association rules are used for the rule-based component. The first step in this component is preparing the data. The dataset must be changed in a way that can be used by association rules algorithms. For discretization of the attributes, each attribute is divided to its possible categories. For example, the attribute of gender is discretized to "male" and "female". A code is assigned to each category. In the discretization of attributes is shown.

Table 3: discretization of attributes for association rules.

attribute	Discretized attributes	Code		
Gender	Female	0	1	0
	Male	1	0	1
Age	Teen	0	0	0
	Young	0	0	1
	Middle-aged	1	0	0
	Old	0	1	0
Time	Between 0 and 7	0	1	0
	Between 7 and 13	1	0	0
	Between 13 and 21	0	0	1
	Between 21 and 24	0	0	0
Channel	Mobile channel	0	1	0
	POS channel	1	0	0
	Internet channel	0	0	1
	ATM channel	0	0	0
Amount	Low amount	0	0	0
	Average amount	0	0	1
	High amount	1	0	0
	Very high amount	0	1	0

The Apriori algorithm is applied on the prepared data. After applying the algorithm, a subset of attributes with most repetitions will emerge. In fact, the aim of the Apriori algorithm is to find the dependencies between different attributes in a dataset. A sample rule extracted by the Apriori algorithm is provided below:

*"**IF** (a teen man at the time period between 9 p.m. and 10 p.m. is engaged in a transaction on the internet channel and through a terminal with type ter1), **THEN** (the transaction amount is low)".*

The rules were optimized by applying them on the test dataset and removing those rules which were not in accordance with any data. The remaining rules were used for analyzing the newly-arrived transactions. When a new transaction arrives, all the rules are applied to it and if no rule fires for that transaction, it will be regraded as a risky transaction. This result then will be combined with the result of trend analysis component.

4.2. Trend Analysis Component

In this component, the c-mean algorithm which is a fuzzy version of k-means is applied to the normal transactions of the dataset. The numerical main attributes and aggregated attributes are used for this algorithm. As shown in Fig. 2, in addition to the attributes like amount and account balance, there are 86 newly-derived aggregated numeric attributes, which means a total of 88 attributes are employed for clustering. In order to avoid the curse of dimensionality, the process of dimension reduction has been implemented on the data. Multiple data reduction methods are discussed in [28]. Since all the attributes used for clustering are numerical, here the PCA (Principal Component Analysis) method is used. In Fig. 3, the result of applying PCA to the dataset is demonstrated.

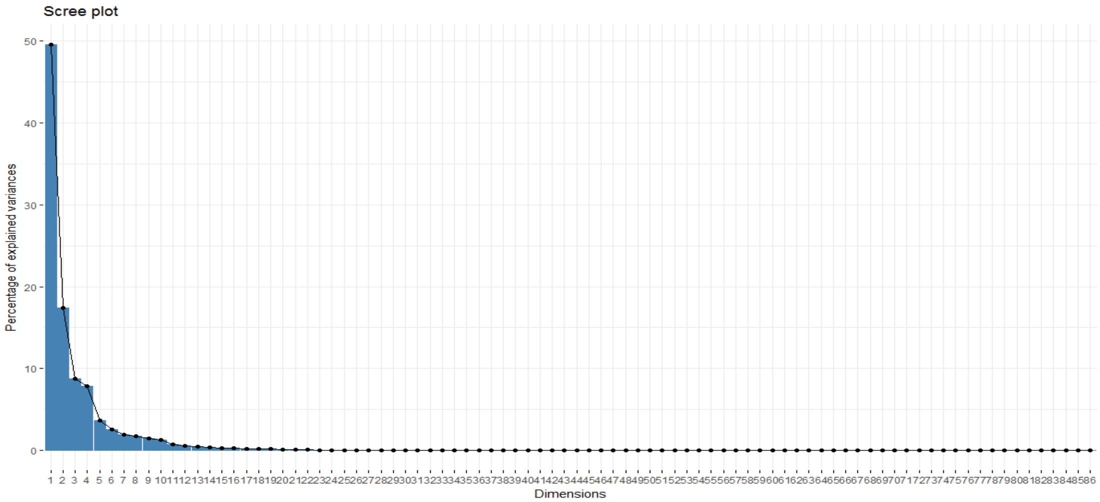

Fig. 3: The result of PCA on clustering dataset.

Eigenvalue is a key element for determining the PCs. If the eigenvalue for a PC is greater than 1, then it has a greater variance compared to the variables in the original data. Hence the PC can be held. There are 12 meaningful and usable PCs in the dataset; therefore, the number of dimensions is reduced to 12.

The c-means algorithm, categorizes data to c different clusters. Like k-means the number of clusters (c), must be determined at first. Since the clustering algorithm is applied to the data of any card holder separately, the number of clusters for each card holder data may vary. Various methods are discussed in [28] for determining the number of clusters and here the SSE (Sum of Square Error) method is used. By depicting the SSE chart vs the number of clusters and finding the heel of the diagram, the optimized number of clusters can be find. Fig. 4 shows how the number of clusters (c) can be found for a card holder. Here the number of optimized cluster is 8.

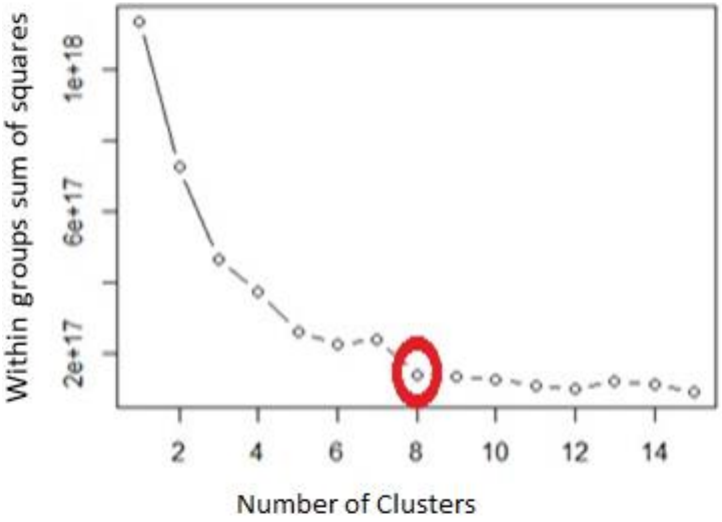

Fig. 4: Calculating C.

The result of the clustering for the data of a sample card is illustrated in Table 4. Each transaction of the card belongs to one of the 8 clusters with a degree of certainty. The main cluster for a transaction has the maximum degree of belonging. For example in this table, the main cluster for the transaction with a code of 202, is the cluster with the degree of belonging equal to 0.5721.

Table 4: The result of clustering for a sample card data.

Transaction code	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Custer 5	Cluster 6	Cluster 7	Cluster 8
202	0.07565	0.07531	0.57214	0.07238	0.07542	0.06665	0.05931	0.00311
551	0.01166	0.01169	0.01169	0.01185	0.01169	0.01092	0.01072	0.91710
286	0.17097	0.16761	0.16761	0.14874	0.16863	0.16793	0.11864	0.00309
615	0.12767	0.12173	0.12173	0.09552	0.12348	0.30507	0.10339	0.00134
654	0.07867	0.07445	0.07445	0.05685	0.07568	0.43822	0.25260	0.00097
32	0.06750	0.06525	0.06525	0.05511	0.06592	0.22636	0.48238	0.00193

For an arriving transaction, if the degree of belonging for all clusters is less than a predefined threshold, then that transaction will be regarded as an outlier or risky transaction.

5. Results and Discussion

The dataset deployed in this research is provided by an Iranian bank and consists of cards data during February 2015 to January 2016. Each transaction has 12 raw attributes as shown in Table 1, plus 186 aggregated attributes as described in Fig. 2. There are 1934 fraudulent transactions belonging to 580 cards in the dataset which are called victim cards here. The number of all the transactions made by the victim cards is 683100. The existing dataset has label only for the transactions of the victims. These fraudulent transactions are reported by the card owners and receive their label as fraudulent after being assessed by auditors. The other transactions in our dataset have no label, so it is not known whether they are fraudulent or not. The output of applying the c-mean clustering algorithm on the data for a sample card is shown in Fig. 5.

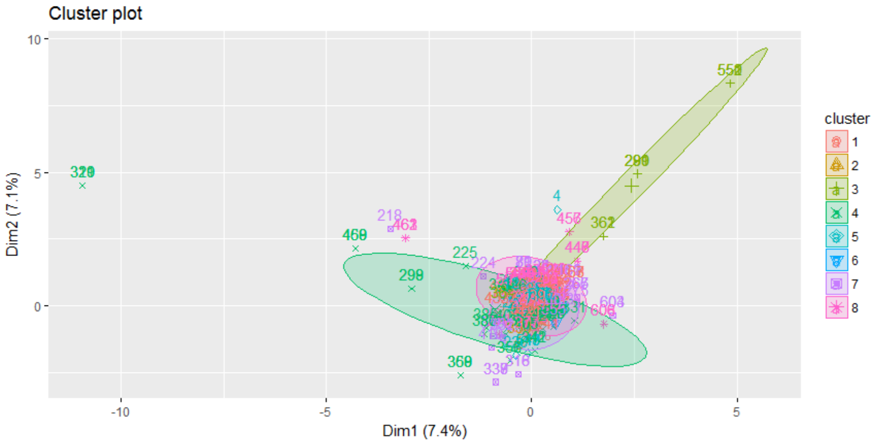

Fig. 5: The output of applying c-means on transactions of a sample card.

The confusion matrix is used for the evaluation of the proposed model. Detection rate and false alarm rates as two main factors for the evaluation of fraud detection algorithms is used in this paper. A good fraud detection system has the maximum TP (detected frauds) and minimum FP and FN (false alarm rates).

A comparison among the results of clustering, association rules and the proposed model is depicted in a ROC diagram in Fig. 6.

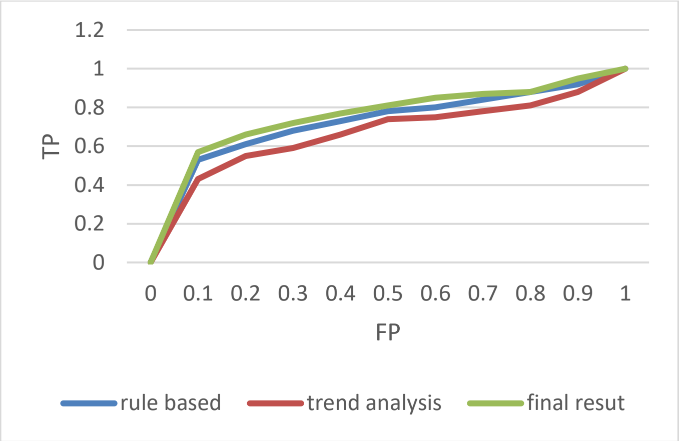

Fig. 6: The ROC diagram for comparing the results of clustering, association rules and the proposed model.

It is obvious from the picture that the result of the proposed model is more exact than than each of clustering and association rules alone. Several entities and events such as human wisdom, analytic tools and business systems are involved in the occurrence of a fraud. Effective and instant detection of sophisticated frauds requires more accurate analytic tools and algorithms. In this paper, the study and practice in the real world are presented and a model consisting of two main components, namely rule-based component, trend analysis-based component is introduced. Transaction aggregation and derived attributes are also found to be useful in fraud detection. Here, in addition to the association rules for extracting rules, trend analysis was carried out as a semi-supervised method and it was discovered that despite the fact that semi-supervised methods have lower detection rates in comparison to association rules (Fig. 6), combining them leads to improved results.

References

[1] T. B. Joewono, B. A. Effendi, H. S. A. Gultom, and R. P. Rajagukguk, "Influence of Personal Banking Behaviour on the Usage of the Electronic Card for Toll Road Payment," *Transp. Res. Procedia*, vol. 25, pp. 4454–4471, Jan. 2017.

[2] A. Correa Bahnsen, D. Aouada, A. Stojanovic, and B. Ottersten, "Feature engineering strategies for credit card fraud detection," *Expert Syst. Appl.*, vol. 51, pp. 134–142, Jun. 2016.

[3] H. Hoehle, E. Scornavacca, and S. Huff, "Three decades of research on consumer adoption and utilization of electronic banking channels: A literature analysis," *Decis. Support Syst.*, vol. 54, no. 1, pp. 122–132, Dec. 2012.

[4] V. Van Vlasselaer et al., "APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions," *Decis. Support Syst.*, vol. 75, pp. 38–48, Jul. 2015.

[5] A. Dal Pozzolo, O. Caelen, Y.-A. Le Borgne, S. Waterschoot, and G. Bontempi, "Learned lessons in credit card fraud detection from a practitioner perspective," *Expert Syst. Appl.*, vol. 41, no. 10, pp. 4915–4928, Aug. 2014.

[6] M. F. A. Gadi, X. Wang, and A. P. do Lago, "Credit Card Fraud Detection with Artificial Immune System," in Artificial Immune Systems, 2008, pp. 119–131.

[7] R. J. Bolton, D. J. Hand, and D. J. H, "Unsupervised Profiling Methods for Fraud Detection," in *Proc. Credit Scoring and Credit Control VII*, pp. 5–7, 2001.

[8] D. J. Hand, C. Whitrow, N. M. Adams, P. Juszczak, and D. Weston, "Performance criteria for plastic card fraud detection tools," *J. Oper. Res. Soc.*, vol. 59, no. 7, pp. 956–962, Jul. 2008.

[9] N. Carneiro, G. Figueira, and M. Costa, "A data mining based system for credit-card fraud detection in e-tail," *Decis. Support Syst.*, vol. 95, pp. 91–101, Mar. 2017.

[10] S. Wang, "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, 2010.

[11] N. A. L. Khac and M. T. Kechadi, "Application of Data Mining for Anti-money Laundering Detection: A Case Study," in *2010 IEEE International Conference on Data Mining Workshops*, pp. 577–584, 2010.

[12] J. Wu, H. Xiong, and J. Chen, "COG: local decomposition for rare class analysis," *Data Min. Knowl. Discov.*, vol. 20, no. 2, pp. 191–220, Mar. 2010.

[13] R. Liu, X. l Qian, S. Mao, and S. z Zhu, “Research on anti-money laundering based on core decision tree algorithm,” in *2011 Chinese Control and Decision Conference (CCDC)*, pp. 4322–4325, 2011.

[14] W. H. Chang and J. S. Chang, “Using clustering techniques to analyze fraudulent behavior changes in online auctions,” in *2010 International Conference on Networking and Information Technology*, pp. 34–38, 2010.

[15] L. Torgo and C. Soares, “Resource-bounded Outlier Detection Using Clustering Methods,” in *Proceedings of the 2010 Conference on Data Mining for Business Applications*, Amsterdam, The Netherlands, The Netherlands, pp. 84–98, 2010.

[16] L. Torgo and E. Lopes, “Utility-based Fraud Detection,” in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, Barcelona, Catalonia, Spain, vol. 2, pp. 1517–1522, 2011.

[17] F. H. Glancy and S. B. Yadav, “A computational model for financial reporting fraud detection,” *Decis. Support Syst.*, vol. 50, no. 3, pp. 595–601, Feb. 2011.

[18] M. Ester, H. Peter Kriegel, J. Sander, and X. Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,” 1996. [Online]. Available: http://www.aaai.org/Library/KDD/1996/kdd96-037.php [March 14, 2018].

[19] S. Panigrahi, A. Kundu, S. Sural, and A. K. Majumdar, “Credit card fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning,” *Inf. Fusion*, vol. 10, no. 4, pp. 354–363, Oct. 2009.

[20] M. Singh and S. Raheja, “Credit Card Fraud Detection by Improving K-Means,” vol. 2, no. 5, 2014.

[21] J. Kim, A. Ong, and R. E. Overill, “Design of an artificial immune system as a novel anomaly detector for combating financial fraud in the retail sector,” in *The 2003 Congress on Evolutionary Computation, CEC ’03*, vol. 1, pp. 405–412, 2003.

[22] U. Murad and G. Pinkas, “Unsupervised Profiling for Identifying Superimposed Fraud,” in *Principles of Data Mining and Knowledge Discovery*, pp. 251–261, 1999.

[23] E. Aleskerov, B. Freisleben, and B. Rao, “CARDWATCH: a neural network based database mining system for credit card fraud detection,” in *Proceedings of the IEEE/IAFE 1997 Computational Intelligence for Financial Engineering (CIFEr)*, pp. 220–226, 1997.

[24] H. Issa and M. A. Vasarhelyi, “Application of Anomaly Detection Techniques to Identify Fraudulent Refunds,” *Social Science Research Network*, Rochester, NY, SSRN Scholarly Paper ID 1910468, Aug. 2011.

[25] A. Sorin, “Survey of Clustering based Financial Fraud Detection Research,” *Informatica Economică*, vol. 16, 2012.

[26] D. T. Larose and C. D. Larose, “Association Rules,” in Discovering Knowledge in Data, John Wiley & Sons, Inc., 2014, pp. 247–265.

[27] S. Jha, M. Guillen, and J. Christopher Westland, “Employing transaction aggregation strategy to detect credit card fraud,” *Expert Syst. Appl.*, vol. 39, no. 16, pp. 12650–12657, Nov. 2012.

[28] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Haryana, India; Burlington, MA: Morgan Kaufmann, 2011.