Egyptian Computer Science Journal (ISSN: 1110 – 2586)
Volume 40 – Issue 03, September 2016

Enhanced Fraud Miner: Credit Card Fraud Detection using Clustering Data Mining Techniques

Mohamed Hegazy[1], Ahmed Madian[2, 3], Mohamed Ragaie[1]
[1]Information System department the Arab Academy for Science Technology and Maritime Transport
[2] Nano-Electronics Integrated Systems Center, Nile University, Cairo, Egypt
[3] National Center for Radiation Research, Egyptian Atomic Energy Authority, Egypt
hegazy_prog@yahoo.com, ah_madian@hotmail.com, ragaie2@mcit.gov.eg

Abstract

This paper aimed to build unified pattern per customer not only represent normal behavior but also Fraud pattern that's represented previously and confirmed as fraud transactions that's facilitate studding fraudsters behavior. An enhancement for the proposed algorithm of Fraud Miner has been proposed. This enhancement involves introducing LINGO clustering Data mining algorithm by replacing Apriori algorithm used in Fraud Miner for Frequently Pattern creation and facilitate summarize customer previous behavior either within his Legal or Fraud transactions. Using this algorithm provide more chance for easily fraud detection as the fraudsters always behaving same as customer behaviors instead of study fraudster behavior the customer frequent behavior will be identified from his legal or previously confirmed transactions being fraud. A performance comparison with other algorithms has been carried out.

Keywords*: Apiriori, Clustering data mining, Fraud detection, Lingo, PCI-DSS.*

1. Introduction

The innovation of new technologies, communication techniques and from the fact of Fraud is ubiquitous; it does not discriminate in its occurrence. Anti-fraud controls can effectively reduce the likelihood and potential impact of fraud, actually no entity is immune to this threat. Unfortunately, however, many organizations still suffer from an "it can't happen here" mindset [1].

Valuable knowledge and interesting patterns are hidden in this data. There are huge potential for banks to apply data mining in their decision making processes in areas like marketing, credit risk management, and detection of money laundering, liquidity management, investment banking and detection of fraud transactions in time Failures in these areas can lead to unpleasant outcomes for the bank such as losing customers to competition, financial loss, reputational loss and hefty fines from the regulators. According to Review provided in [2] about data mining we have two common used approaches for fraud detection in banking industries as shown in Figure 1.

Users go through reports generated by banking information system and use it in their decision making process. Manual analysis has limitations because volumes of data that can

be manually analyzed are limited and hence the decisions may not be as accurate as intended.

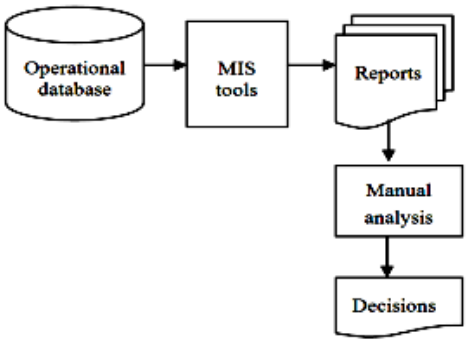

Figure 1. Conventional decision making process

It is assumed that valuable information are hidden in this volume of operational and historic data that can be used for critical decision making process if they are discovered and put to use by capable tools [3]. For example, a decision support system based on data mining techniques can be employed to improve the quality of lending process in a bank [4]. The second recommended approach that showed in Figure 2 how data mining can improve decision making process.

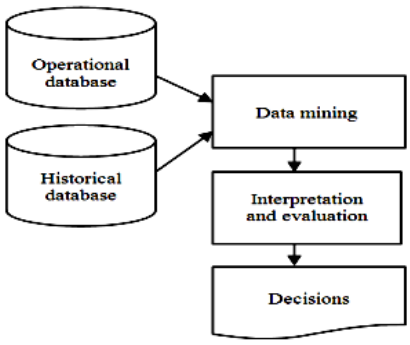

Figure 2. Decision making with data mining

According to ACFE (Association of Certified Fraud Examiners) report of 2014 that contains analysis of 1483 cases that caused in excess of $ 3 billion in losses due to fraud as the median loss caused by a single case of occupational fraud is $ 145,000 that record that financial statement fraud schema type represent high median losses that mainly appeared on Banking and financial services industry that we have the main focused here specially Credit card fraud is practiced most frequently amongst the varied financial frauds due to its acceptance and widespread usage as it offers more convenience to its users.

Due to the nature of huge amount of transactions that need to be manually analyzed which are limited that negatively impacts the decisions accuracy , Data Mining techniques as One domain data mining can excel at, suspicious transaction monitoring, has emerged for the first time as the most effective fraud detection method in 2011 according to Survey [5]. Out of the available data mining techniques, clustering has proven itself a constant applied solution for detecting fraud. Anomaly detection in the past couple of years achieved good results in knowing customers' pattern that’s detects any Out-Of-Norm activities.

This paper proposes Credit card fraud detection model that's handle imbalanced dataset and facilitate knowing of customers' patterns by splitting data into legal (confirmed True transactions) and fraud (Confirmed Fraudster behaviours) patterns to eliminate the problem of imbalanced dataset.

Paper organized as follows Section 2 provides the innovation of enhanced fraud Miner and introduces the usage of LINGO clustering Data mining algorithm.

Section 3 explains the methodology and techniques approached for implementation. Section 4 involves testing of proposed model and results discussion specially, when compared with the original algorithm of fraud Miner. Section 5 provides conclusion and Future work.

2. Background and Related Work

Many security controls implemented by VISA and MasterCard on transactions level that reflected either from PCI-DSS (Payment Card Industry Data Security Standards) set of rules and policies or from the force of using the new chip cards that uses EMV (Euro pay MasterCard VISA) technology [6] that solves many security vulnerabilities appeared on old magnetic strip cards some of these vulnerabilities related to card skimming that consider type of card fraud which is involves the theft of credit card information used in an otherwise legitimate transaction as EMV achieved great efforts concerning contactless operations [7] whereas most of Europe's cards are Contact-based as EMV facilitate offline authorization With ISO 14443 becoming a payment standard and merchants are able to accept contactless payments from different card companies, including Visa and MasterCard that reflect on many convenience to merchants.

Comparative study for different biometric authenticators technologies that might be used in online banking [8],study proved that fingerprint, iris and face are the most appropriate for inclusion in biometric authentication systems of online banking specially when two-factor authentication are required.

Credit card transactions trained using Baum-Welch algorithm in [9] by modeling sequence of operation using Hidden Markov Model (HMM) and dividing transactions into three groups high, medium and low according to transaction amount so that spending profile of cardholder created more easier.

Hybrid algorithm proposed in [10] for credit card fraud detection based on combination of Naïve Bayes algorithm with Hidden Markov model and offering OTP (One Time Password) for newly transactions for more security about newly behaviours.

Principle component analysis proposed in [11] aimed to represent each sample of transaction with few number of values so that attributes could be reduced by determine attributes contains major information and facilitate faster fraud detection for credit card transactions.

Novel web clustering Data Mining invented in [12] which considered as strong emphasis is placed on the high quality of group descriptions named by LINGO algorithm based on Latent Semantic Indexing and Singular Value Decomposition that's firstly identify good cluster labels with meaningful meaning then assigning the contents to each label. LINGO produces reasonably described and meaningful clusters when implemented into the Carrot2 framework that significantly influence the quality of clustering.

LINGO algorithm implemented in [13] and proposed this algorithm as an approach for clustering could be used for outlier detection as the proposed algorithm idea is to first find meaningful descriptions of clusters (Description come First "DCF") then assign documents to the produced labels using (LSI) Latent Semantic Indexing and regarding documents assigning Vector Space Model used to determine cluster content.

Many studies and efforts spend in credit card fraud detection, K-Mean proposed in [14] as clustering data mining algorithm to indicate legal/ fraud transactions however real data seems not available but proposed algorithm showed significant results in fraud detection.

This work aimed to enhance current Fraud Miner algorithm provided by [15] that proposed new fraud detection technique to handle imbalance class by identifying fraud patterns for each customer instead of finding a common pattern for fraudulent behaviour.

Using data Pre-classification that determine legal and fraud transactions per customer to facilitate speed up the process of fraud detection as both legal and fraud behaviour doesn't have frequently changes, it's changes over longer period of time.

Applying Apriori algorithm to generate legal and fraud patterns then using customized matching algorithm transaction fraud detection process become more easier and enables real time detection ,it's recorded highest fraud detection rate and showed good performance in handling class imbalance when compared to NB(Naive Bayes) , SVM (Support Vector Machine) , RF(Random Forest), KNN (K-Nearest Neighbour) classifiers.

It's noticed that the majority of clustering mining algorithms made the discovery process first then based on contents labels inducted that some time result in some groups' description meaningless as this problem solved in [12] by introducing Lingo as radically different approach to finding and describing groups that aimed to firstly find meaningful cluster description then assigns snippers to them and introduce DCF(Description Comes First) method using Singular Value decomposition(SVD) as reduction technique [13] as indicated below summarized steps:

(a) Data Pre-processing by apply stemming, text filtering and remove stop words.
(b) Feature Extraction that aimed to discover frequent items and phrases.
(c) Cluster Label induction that's find best matching phrase.
(d) Cluster Content Discovery that's assign contents to the resulted clusters.
(e) Final Cluster formation that's Calculate cluster scores and apply cluster merging.

Lingo algorithm implemented as a component of Carrot2 framework that achieved good results, Apriori algorithm have the lowest memory usage when compared with different algorithms of association rules [16], Apriori requires many database scan till having a refined pattern that's negatively affect the performance due to consuming time [17] which reduce the performance of fraud, This paper implements Lingo algorithm for valid/fraud pattern creation instead of Apriori algorithm that proposed in fraud Miner[15].

3. Methodology

The proposed enhancement for Fraud Miner will involve Phases as follows:

3.1 Data Preparation

Due to sensitivity and confidentiality of needed card holder data required for test and Banks limitation to provide this data for test so before starting Data preparation phase it's required to have transactions simulator that responsible for simulate transactions and prepare

appropriate imbalanced dataset.

Data Pre-processing would be required after dataset formulation as follows:

(a) Refine data by Remove the transactions corresponding to those customers who have only one transaction in dataset.

(b) Segregate transactions into legal and fraud transactions.

The refined imbalanced data represented in Table 1.

Table 1. Imbalanced data number of transactions in training set

Number of customers	Legal	Fraud	Total
200	*40225*	*12236*	*52461*
400	*60957*	*14075*	*75032*
600	*101600*	*16050*	*117650*
801	*133526*	*20522*	*154048*
1000	*165841*	*24372*	*190213*
1200	*209097*	*26817*	*235914*
1400	*241657*	*31126*	*272783*

3.2 Algorithms Implementation and Patterns Creation

Prepare an Implementation for Apiriori and Lingo algorithm according to the nature of simulated test data we have below lingo attributes should be set within Lingo algorithm:

(a) Set desired Cluster Count Base to 25 as this attribute refer to desired cluster count base as a Base factor used to calculate the number of clusters based on the number of documents on input. The larger the value, the more clusters will be created. The number of clusters created by the algorithm will be proportional to the cluster count base, but not in a linear way.

(b) Set cluster Merging Threshold value to be 0.9 as this attribute refer to Cluster merging threshold. The percentage overlap between two cluster's documents required for the clusters to be merged into one clusters. Low values will result in more aggressive merging, which may lead to irrelevant documents in clusters. High values will result in fewer clusters being merged, which may lead to very similar or duplicated clusters.

(c) Set Stop Word Label Filter. enabled false instead of disabled as this attribute intend to Remove stop labels. Removes labels that are declared as stop labels in the stop labels. <lang> files. Please note that adding a long list of regular expressions to the stop labels file may result in a noticeable performance penalty.

(d) Set Document Assigner.min Cluster Size 1 instead of default 2 that’s determines the minimum number of documents in each cluster.

Run Apiriori Pattern Generation application to generate customers' related Fraud and Legal patterns using Apiriori algorithm then run LINGO pattern Generation application to generate customers' related Fraud and Legal patterns using LINGO algorithm.

4. Proposed Algorithm Efficiency

This section aimed to test proposed algorithm efficiency when compared to old fraud miner and discuss output results.

4.1 Results

Due to the imbalanced nature of data we have 4 classification metrics relevant to credit card fraud detection measures [18] fraud detection rate, false alarm rate, balanced classification rate, and Matthews's correlation coefficient.

Here, fraud is considered as positive class and legal as negative class and hence the meaning of the terms P, N, TP, TN, FP, and FN are defined as follows:

Positives (P): number of fraud transactions;
Negatives (N): number of legal transactions;
True positives (TP): number of fraud transactions predicted as fraud;
False negatives (FN): number of fraud transactions predicted as legal.
True negatives (TN): number of legal transactions predicted as legal;
False positives (FP): number of legal transactions predicted as fraud;

Before starting the comparison we need to prepare data by develop program that's get fraud and legal transactions per table count and pass every single transaction as incoming transaction to verify and by implementing matching algorithm provided in [15] to detect whether the incoming transaction fraud or legal and put results in summary as represented in Table 2 sample of legal transactions test.

Table 2. Sample of Legal Transactions

Working algorithm	Customers count	Fraud count	Legal count
Apiriori	200	51	149
Apiriori	400	79	321
Apiriori	600	118	482
Apiriori	801	123	678
Apiriori	1000	123	877
Apiriori	1200	123	1077
Apiriori	1400	123	1277
Lingo	200	44	156
Lingo	400	67	333
Lingo	600	100	500
Lingo	801	102	699
Lingo	1000	102	898
Lingo	1200	112	1088
Lingo	1400	106	1294

The performance of the enhanced algorithm evaluated with the original fraud miner proposed in [15] as from the Fraud miner performance evaluation previously held with other four credit card fraud detection algorithms support vector machine (SVM), K-Nearest

neighbour classifier, naıve Bayes classifier, and random forest, Fraud Miner were having highest fraud detection rate than other classifiers with very less false alarm rate.

Fraud Detection Rate is the percentage of correct positive fraud transactions from actual Fraud transactions.

Figure 3 shows the performance of Enhanced fraud Miner that's represented the same performance of the original Fraud Miner.

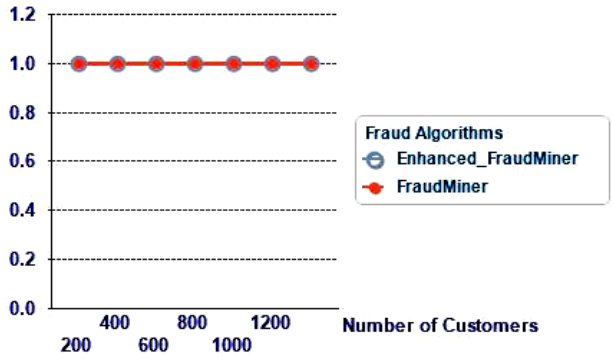

Figure 3. Sensitivity Rate

False Alarm Rate. Which represent number of actual negatives transactions predicted as positives.

Figure 4 shows the performance of enhanced Fraud Miner on False alarm rate that's represented more reducing in false alarm rate when compared with the original Fraud Miner.

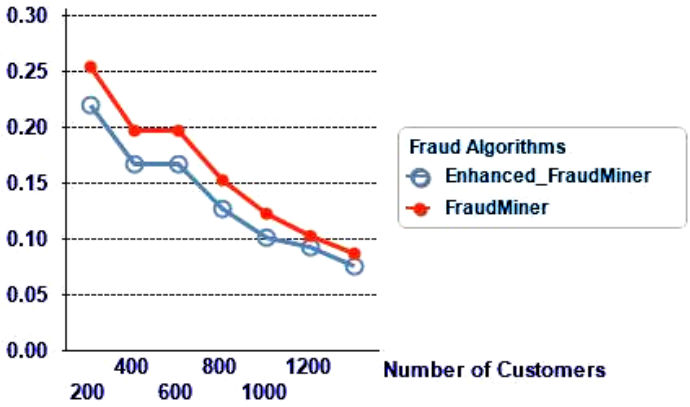

Figure 4. False Alarm Rate

Balanced Classification Rate (BCR). Which represent the average of sensitivity and specificity as Figure 5 represented small enhancement in this rate for the enhanced fraud Miner.

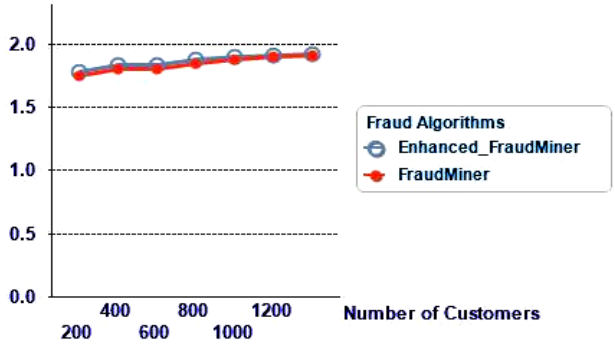

Figure 5. Balanced Classifier Rate

Matthews correlation coefficient (MCC). Which is used as a measure of the quality of binary classifications as according to nature of simulated test data and from Figure 6 it's found that enhanced Fraud Miner have some enhancement in quality of binary classifier.

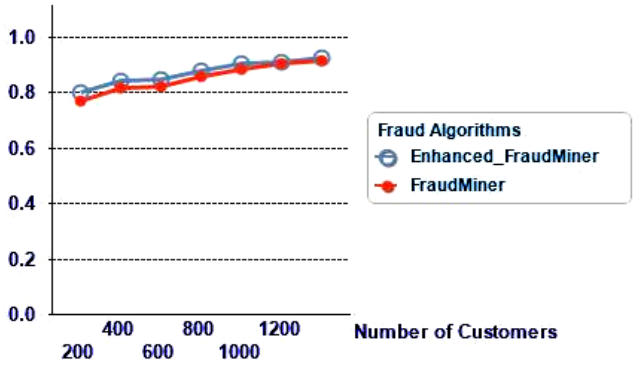

Figure 6. Mathews Correlation Coefficient

4.2 Discussion

Measuring Fraud detection algorithm require to pay more attention for sensitivity and False alarm rate, Fraud Miner recorded highest fraud detection and lowest false alarm rates when compared to other classifier. In proposed enhanced algorithm (Figures 3 and 4) we have not only output the same sensitivity but also decreasing false alarm rate and improving customer satisfaction.

Quality of algorithms that handled imbalanced data measures using balanced metrics of BCR and MCC, A coefficient of +1 means highest quality, 0 means algorithm act like random prediction system, and -1 means very low quality and algorithm failed in prediction and observation. Proposed enhanced algorithm (Figures 5 and 6) shows the same quality achieved in fraud Miner with some coefficient enhancements that's increases prediction.

Using the same matching algorithm in the proposed enhanced fraud miner resulted in keeping limitation within fraud detection, especially in case of identical transactions exist both in Legal and Fraud patterns (overlapping) that's leads to unable to recognize fraud transactions.

5. Conclusion and Future Work

In this paper, a Survey on different Data Mining techniques used for Credit card fraud detection has been introduced. Fraud/Legal Pattern creation for each customer facilitate customer profile detection not only normal behaviour but also fraudster behaviors on his account and this made fraud detection easier. Also, the LINGO algorithm proved to be used to create Legal/Fraud patterns per each customer instead of Apiriori Algorithm and work almost in the similar efficient. By using simulated test transactions it's found that LINGO generate meaningful summarized patterns more than output patterns from Apriori Algorithm. Fraud/Legal Pattern creation facilitate fast of fraud detection process and could be used to verify transaction near real time transactions.

According to comparison results of proposed model for enhancing Fraud Miner algorithm it's achieved good enhancements especially with the high important measure of False alarm rate that decreased more in the proposed model also it's noticed an improvements of algorithm quality that represented from Matthews correlation coefficient.

The future work suggested to apply LINGO3G as enhanced LINGO algorithm that have many improvements like achieving very fast clustering over huge snippets records, improving cluster label quality , Hierarchical Clustering , promoting specific words or phrases in the output cluster labels and defining groups of words or phrases to be treated as synonymous with advanced Results tuning.

Credit card transactions could be segregated according to transaction source so that group ATM, POS, Merchant Draft Entry (MDE) and other types of transactions to facilitate more speed in fraud detection process.

References

[1]. Singleton T. W.: Fraud Auditing and Forensic Accounting 4th edition, Ed. John Wiley and Sons, (2010)

[2]. Pulakkazhy, S., &Balan, R. V. S.: Data Mining in Banking and Its Applications-a Review. Journal of Computer Science, 9(10), pp. 1252–1259. doi:10.3844/jcssp.2013.1252.1259, (2013)

[3]. Kazi, I.M. and. Q.B. Ahmed: Use of data mining in banking. Int. J. Eng. Res. Appli., pp. 738-742 (2012)

[4]. Ionita, I. and L. Ionita: A decision support based on data mining in e-banking. IEEE Preccedings of the 10th Reodunet International Conference (RoEduNet), Jun. 23-25, IEEE Xplore Press, Iasi, pp. 1-5. DOI: 10.1109/RoEduNet.2011.5993710 (2011)

[5]. Sorin, A.: Survey of Clustering based Financial Fraud Detection Research. InformaticaEconomica, 16(1), pp. 110–123, (2012)

[6]. EMV-Book3 Card, I. C. vol. 3, (June 2008)

[7]. Greenemeier, L., : Visa expands contactless card efforts, InformationWeek, March 27, http://tinyurl.com/ykzo4t (2006)

[8]. Parusheva, S. "A comparative study on the application of biometric technologies for authentication in online banking", Egyptian Computer Science Journal (ECS), Vol.39, No.4, ISSN-1110-2586, September 2015, pp.115–126.

[9]. Matheswaran, P., Me, E. S. S., & Rajesh, R. (2015). Fraud Detection in Credit Card Using DataMining Techniques, II(I), 11–18

[10]. Gupta, A., &Raikwal, J.: Fraud Detection in credit Card Transaction Using Hybrid Model, 3(1), pp. 3730–3735, (2014)

[11]. D. Pawar, A., N. Kalavadekar, P., & N. Tambe, S.: A Survey on Outlier Detection Techniques for Credit Card Fraud Detection. IOSR Journal of Computer Engineering, 16(2), pp. 44–48. doi:10.9790/0661-16264448 , (2014)

[12]. Osi'nski, S.: An algorithm for clustering WEB SEARCH RESULTs. Journal of Mathematical Psychology, 12(3), pp. 328–383 ,(2003)

[13]. Fafat, P. C., &Sikchi, P. S. S. :Lingo an approach for Clustering, 1(3), pp. 1–3 , (2012)

[14]. Journal, I., Applications, C., & Design, V. :Fraud Detection in Credit Card by Clustering Approach, 98(3), pp. 29–32 (2014)

[15]. Seeja, K. R., &Zareapoor, M. FraudMiner: A Novel Credit Card Fraud Detection Model Based on Frequent Itemset Mining. TheScientificWorldJournal 2014,252797. doi:10.1155/2014/252797, (2014)

[16]. Alhamzi, A., Nasr, M., &Salama, S. "A Comparative Study of Association Rules Algorithms on Large Databases", Egyptian Computer Science Journal (ECS), Vol.38, No.3, ISSN-1110-2586, September 2014, pp.51–62.

[17]. Mohammed Al-Maolegi, B. A.: An improved apriori algorithm for association rules of mining. International Journal on Natural Language Computing, 3(1), pp. 942–946. doi:10.1109/ITIME.2009.5236211, (2014)

[18]. François, D.: Binary classification performances measure cheat sheet. Journal of Machine Learning Research, 7, pp. 1–30, (2006).