Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 148 (2019) 45–54

Procedia Computer Science

www.elsevier.com/locate/procedia

Second International Conference on Intelligent Computing in Data Sciences (ICDS 2018)

Performance of machine learning techniques in the detection of financial frauds

I.SADGALI[a], N.SAEL[a], F.BENABBOU[a]

[a]*Laboratory of Modeling and Information Technology*
Faculty of sciences Ben M'SIK, University Hassan II, Casablanca, Morocco
sadgali.imane@gmail.com, saelnawal@hotmail.com, faouzia.benabbou@univh2c.ma
00212661822838

Abstract

Financial fraud presents more and more threat that has serious consequences in the financial sector. As a result, financial institutions are forced to continually im-prove their fraud detection systems. In recent years, several studies have used machine learning and data mining techniques to provide solutions to this problem. In this paper, we propose a state of art on various fraud techniques, as well as detection and prevention techniques proposed in the literature such as classification, clustering, and regression. The aim of this study is to identify the techniques and methods that give the best results that have been perfected so far.

Keywords: Fraud detection, financial fraud, machine-learning, performance;

1.Introduction

Financial fraud affects tremendously both the financial industry and everyday life. Fraud can reduce confidence in the industry, destabilize savings and affect the cost of living. Financial institutions use a variety of fraud prevention models to address this problem. However, fraudsters are adaptive, and over time, they conceive several ways of intruding such protective models. Despite the best effort of financial institutions, law enforcement and government, financial fraud continues to grow. Fraudsters today can be a very inventive, intelligent and fast fraternity.
This paper, seeks to carry out comparative analysis of financial fraud detection techniques, like machine-learning techniques, who plays an important role in fraud detection, as it is often applied to extract and uncover the hidden truths behind very large quantities of data. Also, many modern techniques for detecting fraud are continually involves and applied to many areas due to the remarkable lift of fraud which effects on financial field in each year. Our objective is to point out their strength and weaknesses and also aim to identify the open issues of fraud analysis.

10.1016/j.procs.2019.01.007

The rest of this paper is organized as follows. Section 2 contains a definition and types of financial fraud, Section 3 present techniques implemented for financial fraud detection, Section 4 present review of related work. The comparative study of various techniques is detailed in Section 5. Finally, we discuss the results and Future work in Section.

2.Financial fraud

2.1 Definition

Fraud definition, according to the Association of Certified Fraud Examiners (ACFE) "ACFE Association of Fraud Examiners Certificates", fraud includes any intentional or deliberate act of depriving another of property or money by cunning, deception or other unfair acts [12].

2.2 Types of financial fraud

There are several types of financial fraud; we present here a brief description of some of the main types of fraud.

Insurance fraud can occur at many points in the insurance process (e.g., application, eligibility, rating, billing, and claims), and can be committed by consumers, agents and brokers, insurance company employees, healthcare providers, and others [1, 2]

Securities and commodities fraud, the FBI [3] provides brief descriptions of some of the most prevalent securities and commodities frauds encountered today, for example, "Market Manipulation, High Yield Investment Fraud, The Ponzi Scheme, The Pyramid Scheme, Prime Bank Scheme, Advance Fee Fraud, Hedge Fund Fraud, Commodities Fraud, Foreign Exchange Fraud, Broker Embezzlement and Late-Day Trading." According to another definition by CULS [4], securities frauds include theft from manipulation of the market, theft from securities accounts, and wire fraud.

Money Laundering is the process by which criminals conceal or disguise the proceeds of their crimes or convert those proceeds into goods and services. It allows criminals to inject their illegal money into the stream of commerce, thus corrupting financial institutions and the money supply and giving criminals unwarranted economic power [5]. Gao and Ye [6] similarly define money laundering as the process by which criminals "wash dirty money" to disguise its illicit origin and make it appear legitimate and "clean."

Financial statement fraud (corporate fraud), financial statements are a company's basic documents to reflect its financial status [10]. It had an objective as.

- Fraud these statements to make the business more profitable
- Improvement of the performance of the actions
- Reduction of tax obligations
- Attempt to exaggerate performance due to managerial pressure

Credit card fraud is essentially of two types; application and behavioural fraud [26]. Application fraud is where fraudsters obtains new cards from issuing companies using false information or other people's information. Behavioural fraud can be of four types: mail theft, stolen/lost card, counterfeit card and 'card holder not present' fraud [25]

Mortgage Fraud is a specific form of financial fraud that refers to the manipulation of a property or mortgage documents. It is often committed to distort the value of a property for the purpose of influencing a lender to finance a loan for it [5].

3. Financial Fraud detection techniques

As stated above, we investigated techniques used for financial fraud detection in anterior works, this bellow our suggested classification for this techniques and a small description for each one of them.

3.1 Descriptive or Unsupervised Techniques

This paragraph, describes descriptive models, that is, the unsupervised learning functions. These functions do not predict a target value, but focus more on the intrinsic structure, relations, interconnectedness, etc.

Self-Organizing Maps [30] a self-organizing map (SOM) is a neural network technique but used unsupervised learning. SOM allows users to visualize data from high dimensional to low dimensional.

Group method of data handling (GMDH) is an inductive learning algorithm for modeling complex systems. It is a self-organizing approach that tests increasingly complicated models and evaluates them using some external criterion on separate parts of the data sample [11]

Outlier detection methods (OD) is very different from the traditional observation method. Outlier method is used to detect unusual behavior of a system using a different mechanism. [38].

Association rule analysis (AR) are defined on transaction sets. Given that it is more common to work with tuples rather than transactions in a database, various solutions to this problem has been proposed. When working with relational databases, it is usual is to consider each item to be a pair (attribute, value) and each transaction to be a tuple in a table [22].

Density based spatial clustering of applications with noise (DBSCAN) is a density based clustering algorithm which can be used to filter out outliers and discover clusters of arbitrary shapes [23].

3.2 Predictive Techniques

In predictive analysis, the purpose is to build an analytic model that predicts target objects of interest.

Logistic Regression (LR) logistic regression is a type of generalized linear model. Using simple linear regression is inappropriate when the variable to be predicted is binary; due to normality assumptions. [7].

Decision Trees (DT) is a tree structure, where each node represents a test on an attribute and each branch represents an outcome of the test. In this way, the tree attempts to divide observations into mutually exclusive subgroups [14].

Classification and Regression Tree (CART) is a computerized, non-parametric technique different from traditional statistical methods. CART applies the binary Recursive Partitioning Algorithm (RPA) to best classify samples into a number of non-overlapping regions, each of which corresponds to a terminal node in the tree [20].

*Decision Trees C4.5 g*ives algorithm and solutions to a set of problems that have arisen over the years among decision tree researchers like handling various problems such as missing attribute values [19].

*Cost-sensitive decision tree (CSDT) a*n induction algorithm developed to identify fraudulent credit card transactions are given. In the well-known decision tree algorithms, the splitting criteria are either insensitive to costs and class distributions or the cost is fixed to a constant ratio [29].

Neural Networks (NN) is a mature technology with an established theory and recognized application areas. A NN consists of a number of neurons, i.e., interconnected processing units. Associated with each connection is a numerical value, called "weight"[14].

Probabilistic neural network (PNN) is a feed-forward NN involving a one pass training algorithm used for classification and mapping of data. It is a pattern classification network, based on the classical Bayes classifier, which is statistically an optimal classifier that seeks to minimize the risk of misclassification [11].

Support Vector Machines (SVM) use a linear model to implement nonlinear class boundaries by mapping input vectors nonlinearly into a high-dimensional feature space. In the new space, an optimal separating hyperplane is constructed [11].

Naïve Bayes (NB) a classification tool simply uses Bayes conditional probability rule. Each attribute and class label are considered random variable, and assuming that the attributes are independent, the naïve Bayes finds a class to the new observation that maximizes its probability given the values of the attributes.[9].

Bayesian belief network (BBN) allow for the representation of dependencies among subsets of attributes. A BBN is a directed acyclic graph, where each node represents an attribute and each arrow represents a probabilistic dependence [14].

Bayesian skewed logit model (BSL) this model incorporates the possibility of using asymmetric links in order to measure the probability of yi = 0 and yi = 1 in non-balanced samples [9]

K-nearest neighbor (KNN) is used largely in detection systems. It is also proved that KNN works extremely well in credit card fraud detection systems using supervised learning techniques. [38].

Bivariate Probit Model (BP) is typically used where a dichotomous indicator is the outcome of interest and the determinants of the probable outcome includes qualitative information in the form of a dummy variable where, even after controlling for a set of covariates, the possibility that the dummy explanatory variable is endogenous cannot be ruled out a priori [40].

3.3 Artificial & Computational Intelligence Techniques

This part, describes artificial and computational intelligence models, which is, a set of nature-inspired computational methodologies and approaches to address complex real-world problems to which mathematical or traditional modelling.

Genetic Algorithm (GA) in Genetic Algorithm i.e. inspired from natural evolution, randomly generated rules are considered as an initial population[15].

Genetic programming (GP) is an extension of genetic algorithms (GA). It is a search methodology belonging to the family of evolutionary computation. GP randomly generates an initial population of solutions. Then, the initial population is manipulated using various genetic operators to produce new populations [11].

Scatter Search (SS) is an evolutionary algorithm, which shares some common characteristics with the GA. It operates on a set of solutions, the reference set, by combining these solutions to create new ones [27].

Hidden Markov Model (HMM) it differs from the normal statistical Markov model by having invisible states, but each state randomly generates one of the visible states. A hidden Markov model can be presented as the simplest dynamic Bayesian network [36].

Iterative Dichotomiser 3 (ID3) for dealing with symbolic data by expressing the knowledge as a decision tree [39].

Artificial Immune System (AIS) the human biological immune system has a number of fundamental characteristics that can be adapted as design principles for AIS applications in various problem domains [28].

Artificial Immune Recognition System (AIRS) both self/non-self cells and detector cells are represented as feature vectors. In order to reduce redundancy, ARB (Artificial Recognition Ball) is used which is representative of similar memory cells [31].

Artificial neural network (ANN) artificial neural networks were first created with the purpose to imitate the behavior of the human brain. A neural network is the connection of elementary objects called the simple neuron [32].

Multilayer Perception Algorithm (MPL) is an artificial neural network and is a nonparametric estimator that can be using for classifying and detecting intrusions [32].

Parenclitic Network (PN) a network reconstruction technique that allows highlighting the differences between one instance and a set of standard [33].

Multi-layer feed forward neural network (MLFF-NN) is one of the most common NN structures, as they are simple and effective, and have found home in a wide assortment of machine learning applications [11].

3.4 Other concepts

This paragraph, present some concepts, which associated with techniques above, for hybrids models.

Fuzzy logic for representing the cognitive uncertainties, measuring the intensity of the truth values for unquantifiable measures or probabilistic measures within the range of 0 and 1 [39]. *Fuzzy association rules (FAR)* can be found in the literature such as a generalization of association rules when initial data are fuzzy or if they have been previously processed to provide them with imprecision [22].

Dempster Shafer Theory (DST) or evidence theory is a general framework for reasoning with uncertainty, the role of DST is to combine evidences from the rules R1and R2 and compute an overall belief value for each transaction. [23].

Computational fraud detection model (CFDM) using SAS® Enterprise Miner™ (EM) as an automation tool to develop the model. The process retains maximal information and uses essentially all of it in processing the documents. [18].

Locally Weighted Learning (LWL) represent nonlinear functions, yet has simple training rules with a single global optimum for building a local model in response to a query. This allows complex nonlinear models to be identified (trained) quickly [41].

4 Related Work

For each type of fraud, several techniques has been used each of which has advantages and shortcomings.

4.1 Insurance fraud

In 2007, StijnViaene found that with claim amount information available at screening time detection rules can be accommodated to increase expected profits, he used logistic regression model [7]. Jean Pinquet presented a statistical approach that counteracts selection bias without using a random auditing strategy [8]. In 2008, the use of an asymmetric link notably improves the percentage of cases that are correctly classified after the model estimation [9], showed by Bermudez.

4.2 card fraud

In 2008, Quah and Sriganesh. proposed, for real-time fraud detection, a new and innovative approach; it makes use of self-organization map, Neural Networks and rules induction [21].

In 2009, a novel methodology has been applied, on credit card fraud, using AR and FAR, was proposed by Sanchez [22]. Suvasini investigated a fusion approach using Dempster–Shafer theory and Bayesian learning [23]. Whitrow developed a framework for transaction aggregation, using a variety of classification methods and a realistic cost-based performance measure [24].

In 2011, Bhattacharyya evaluated two advanced data mining approaches, support vector machines and random forests, together with the well-known logistic regression [25]. Duman suggested a novel combination of the two well-known meta-heuristic approaches, namely the genetic algorithms and the scatter search. The method was applied to real data and very successful results were obtained [27].

In 2012, Wong proposed the use of AIS on one aspect of security management; the detection of credit card fraud [28]. In 2013, Sahin developed a methodology for fraud detection using decision tree and showed that this cost-sensitive decision tree algorithm outperforms the existing methods [29].

In 2014, Olszewski proposed a method of the detection threshold setting based on SOM [30]. Halvaiee developed a novel model for credit card fraud detection using AIS and introduced a new model called AIS-based Fraud Detection Model (AFDM), increase the accuracy up to 25%, reduce the cost up to 85%, and decrease system response time up to 40% comparedto the base algorithm [31].

In 2015, L.Dhanabal analysed NSL-KDD data set, and used it to study the effectiveness of the various classification algorithms in detecting the anomalies [35], the analysis was done using classification algorithms available in the data-mining tool WEKA.

In 2016, Dai developed a hybrid framework with Big Data technologies [36]; that implement it with latest big data technologies like Hadoop, Spark, Storm, HBase, which showed great potentials of achieving the goals. Adewumi focused on recent Machine Learning based and Nature Inspired based credit card fraud detection techniques proposed in literature [37].

In 2017, Mubalaik proposed an ANN-MPL multilayer perception [32], based on implementation of well-known machines learning techniques, it helps to anticipate and quickly detect fraud. Zanin proposed hybrid data mining / complex network, classification algorithm, able to detect illegal instances in a real card transaction data set [33]. Malini implemented KNN algorithm and outlier detection methods to optimize the best solution for the fraud detection problem [38] These approaches are proved to minimize the false alarm rates and increase the fraud detection rate. Askari proposed fraud detection algorithm based on Fuzzy-ID3 [39]. Experimental result exhibits that the technique is efficient one in detecting frauds.

4.3 Financial statement fraud

In 2007, Kirkos investigated the usefulness of Decision Trees, Neural Networks and Bayesian Belief Networks in the identification of fraudulent financial statements [14].Genetic algorithm approach was proposed by HOOGS the patterns are capable of identifying potentially fraudulent behavior despite occasional missing values, and provide low false positive rates [15].

In 2008, BAI proposed in [20] Classification and Regression Tree (CART), to identify and predict the impacts of False Financial Statements (FFS).

In 2011, Cecchini developed a methodology for automating ontology creation using WordNet [16]. Humpherys proposed a parsimonious model with Naïve Bayes and C4.5 achieved the highest classification accuracy [17] and Glancy proposed, for detecting fraud in financial reporting, a computational fraud detection model, using a quantitative

approach on textual data [18]. Also, Ravisankar gave a comparison of data mining techniques; Multilayer Feed Forward Neural Network (MLFF), Support Vector Machines (SVM), Genetic Programming (GP), Group Method of Data Handling (GMDH), Logistic Regression (LR), and Probabilistic Neural Network (PNN) [11] in the same year.

5 A comparative study

In this section, we will analyze the contribution of each technique and its effectiveness, in order to find a promising combination for future work.

5.1 Criteria

In our comparative tables, we regroup and synthesize most used criteria in anterior works in order to have most complete comparison:

- Real time: parameter show if the technique is able to run in real time (R) or not (NR).
- Accuracy: a validation parameter of precision (TP+TN)/ (TP+FP+TN+FN)*
- Sensitivity (or recall): is the measure of the proportion of the number of fraudulent cases predicted correctly as fraudulent by a particular model, gives the accuracy on the fraud cases TP/ (TP+FP)
- Dataset: size, type (particular (P), standard (S) or generic (G))
- Observations: limitations (-) and contributions (+) of the technique

5.2 Comparative tables

In this paragraph, we present three tables, for each type of fraud, we give a summary of techniques in past work and observations belong this study.

Table 1. Contributions and limitations of ML techniques applied to Insurance frauds

Technique	Real time	Validation			Data set		Observations	Reference
		Accuracy (%)	TP (%)	Sensitivity (%)	Size	Type		
LR	NA	99.42	66,67	NA	Claims during 2000	P	- The expected cost can be unprofitable for the company	[7]
BP	NA	NA	NA	NA	Claims during 2000	P	+ The expected overestimation of fraud risk derived was corrected.	[8]
BSL	NA	99,5	98,46*	NA	10 000 automobile claims	P	- Present a lack of fit due to the incorrect classification of zero cases - Unable to signal the significance of the parameter associated to the variable proxim	[9]

* TP: true positive / TN: true negative / FP: False positive / FN: False negative

Table 2. Contributions and limitations of ML techniques applied to credit card frauds[†]

Technique	Real time	Validation			Data set		Observations	Reference
		Accuracy (%)	TP (%)	Sensitivity (%)	Size	Type		
SOM +NN + RI	R	NA	NA	NA	over 200 million customers	P	+ Clustering helps in identifying new hidden patterns in input data +the filtering of transactions for further review reduces the overall cost as well as processing time.	[21]
AR + FAR	R	NA	NA	NA	12,107 transactions	P	+The applied methodology overcomes the difficulties of minimum support and confidence, optimizes the execution times, reduces the excessive generation of rules, and helps make the results more intuitive, thereby facilitating the work of fraud analysts	[22]
DST + NB	R	NA	98	NA	NA	G	+ architecture has been kept flexible so that new rules using any other effective technique can also be included at a later stage	[23]
RF + AG	NR	NA	NA	NA	175million transactions	P	+ The aggregation period has a major impact upon the performance of classifiers for fraud detection.	[24]
SVM + AG	NR	NA	NA	NA				
LR + AG	NR	NA	NA	NA	1.1million transactions			
KNN+ AG	NR	NA	NA	NA				
SVM	NR	95,30	NA	72,70	2420 fraudulent transactions	G	+ While sensitivity, and accuracy decreased with lower proportions of fraud in the training data, precision showed an opposite trend	[25]
RF	NR	90,80	NA	52,40				
LR	NR	94,20	NA	65,40				
GA + SS	NR	NA	NA	NA	100,000 fraudulent	P	+ Bank Management should increase the monitoring capacity if they want to face less losses due to fraud	[27]
AIS	NR	80	88*	NA	640 361 total transactions	P	+ three mechanisms (new transaction representation and variable width r-contiguous bit matching algorithm, vaccination process and memory cell evolution process) significantly improve the performance of the AIS	[28]
SVM	NR	NA	90.0	NA	978 fraudulent records 22 million normal transactions	P	+We cannot use misclassification cost without incorporating the class distribution or an impurity measure in cost calculations.	[29]
CART	NR	NA	83,1	NA				
CSDT	NR	NA	92.1	NA				
SOM	NR	100	NA	NA	10,000 accounts of selected credit card (1.01.2005 - 1.03.2005)	P	+ high-dimensional data projected onto a 2-dimensional space can be easily analyzed and interpreted even by a non-expert.	[30]
AIRS + CC[‡]	NR	NA	83*	NA	3.74% fraudulent transactions	P	+improving memory cell generation improves the detection rate. +Changing distance function, performs better regarding FP.	[31]
DT	NR	91,03	NA	NA	NSL-KDD dataset [34]	S	+ Reduce the complexity of host based analysis engines. - Tend to rely on the innate logging and monitoring capabilities of the server.	[32]
NB	NR	99,02	NA	NA				
ANN+	NR	99,47	NA	NA				

[†] NA: Not Addressed
Proxim: Accident occurred between the policy issue date and the effective starting date, 1; otherwise 0.
*: Calculated

[‡] CC: cloud computing

MPL								
PN+ ANN + MPL	NR	NA	NA	NA	Transactions 01-2011 → 12- 2012.	P	+the addition of parenclitic features to the raw data set enhance the obtained results, with the error dropping from a 19:2 to a 12:23%.	[33]
SVM	NR	98,8	NA	NA	NSL-KDD	S	+NSL-KDD dataset show that it is a best candidate data set to simulate and test the performance of IDS. +Correlation, based Feature Selection method for dimensionality reduction, reduces the detection time and increase the accuracy rate.	[35]
NB	NR	74,9	NA	NA				
DBSCAN + HMM + LR	R	NA	NA	NA	5 dataset, 10,000,.. 1,000,000 within one year.	G	-should allocate ore computing resources to streaming Detection Layer and Batch Training Layer to improve the overall performance during real system deployment	[36]
KNN+ OD	NR	NA	NA	NA	NA	P	+OD works fast and well on online large datasets +KNN can suit for detecting fraud with the limitation of memory	[38]
FL+ ID3	NR	89	NA	NA	NA	P	+reduce the irrelevant processing so that the detection can be done in optimal time	[39]

Table 3. Contributions and limitations of ML techniques applied to financial statement frauds[§]

Technique	Real time	Validation			Data set		Observations	Reference
		Accuracy (%)	TP (%)	Sensitivity (%)	Size	Type		
DT	NA	73.6	75	NA	76 Greek firms	P	+ associated falsification with financial distress, since it used Z score as a first level splitter + revealed dependencies between falsification and the ratios debt to equity, net profit to total assets, sales to total assets, working capital to total assets and Z score.	[14]
NN	NA	80	82.5	NA				
BBN	NA	90.3	88.9	NA				
GA	NA	95	63	NA	AAERs published by the SEC between 2002-2004	S	+ a successful technique for detecting discriminatory patterns in challenging domains characterized by high dimensionality + Patterns are easily translated to domain appropriate language → easily understood by external stakeholders.	[15]
Ontology+ WN**	NA	83.87	81.97	NA	MDAs for 78 companies between 1994 to 1999	S	+The methodology can be applied to available text for any financial problem where the goal is to create a dictionary (ontology) of discriminating concepts.	[16]
LR	NA	63.4	62,9	NA	101 companies	S	+The four-variable model has the predictive accuracy equal to the 10-variable model and performs better than the 24-variable model. +The best performance came from the Naïve Bayes classifier and the C4.5 decision tree classifiers using the 10-variable model.	[17]
NB	NA	67.3	66.7	NA				
SVM	NA	65.8	64.3	NA				
C4.5	NA	67.3	68.0	NA				
LWL	NA	60.4	60.6	NA				
CFDM	NA	90,9*	80*	NA	AAERs published by the SEC 2006-2008	S	+ has the potential to serve as a filtering tool for regulators to focus their resources and subsequently increase the detection of financial reporting fraud	[18]
MLFF-NN	NA	78.36	NA	80.21				
SVM	NA	70.41	NA	55.43	202 companies:		+ Hybrid data mining techniques that combine two or more	

[§] AAER: Accounting and Auditing Enforcement Release SEC: Securities and Exchange Commission
MDAs: Management's Discussion and Analysis
AG: Aggregation
** WN: WordNet

GP	NA	94.14	NA	95.09	101 fraudulent 101 non-fraudulent	S	classifiers can be used on the same dataset. + Results are much superior to an earlier study on the same dataset.	[11]
GMDH	NA	93.00	NA	91.46				
LR	NA	66.86	NA	63.32				
PNN	NA	98.09	NA	98.09				
CART	NA	72.38	NA	72.40				
CART	NA	NA	98,39	NA	24 false, 124 non-false: financial reports	S	+ CART produce more accurate classification on the fraud cases	[20]
LR	NA	NA	95,97	NA				

5.3 Synthesis and discussion

It could be observed that almost, all implemented algorithms, do not work in real time.

As can be seen, the detection of credit card fraud uses several ML techniques, especially those of artificial intelligences and combines them with optimization techniques such as aggregation, for the detection of frauds of financial statements it is based mainly on text processing techniques.

For fraud insurance, the non-necessity of the real-time processing, makes the detection of the fraud easier, nevertheless the difficulty resides in the fact that these deceptions are human and can be well masked. Comparing the logistic regression and the Bayesian results, we see that the Bayes logistic model gives posterior estimations for the true positive rate.

In financial statement fraud, results based on the accuracy indicated that the PNN was the best performing (98.09%) following by Genetic algorithm (95%) who gave marginally lower accuracies in most cases.

Naives bays and SVM gives good results with NSL-KDD dataset (99,02%, 98,8%) for credit card fraud. Also we found that:

- The aggregation period has a major impact on the performance for fraud detection. Aggregating a product improves the prediction rate for all techniques except for CART.
- SOM Clustering helps to identify new patterns hidden in the input data, which otherwise cannot be identified by traditional statistical methods, transaction filtering for further examination reduces overall cost as well as processing time.
- Cost-sensitive decision tree approaches is used in credit card fraud detection and show that it outperforms the models built using the traditional data mining methods such as decision trees, ANN and SVM
- Logistic Regression works well with linear data for credit card fraud detection.
- Support vector machine method is capable of detecting the fraudulent activity at the time of transaction.
- Complex networks can be used as a way to improve data mining models; they may be integrated as complementary tools in a synergistic manner in order to improve the classification rates obtained by classical data mining algorithms.
- KNN method can suit for detecting fraud with the limitation of memory. By the meantime, outlier detection mechanism helps to detect the credit card fraud using less memory and computation requirements.
- Outlier detection works fast and well on online large datasets.

6 Conclusion and future work

In this study, it was found that hybrid fraud detection techniques are the most used, as they combine the strengths of several traditional detection methods. In addition, we discover that the studies do not smother all types of fraud, and each type of fraud has constraints specific to it; response required in real time, text analysis…

In our future work we will focus on continuing the study of credit card fraud to improve current algorithms, we wish to include a hybrid model that is both able to handle imbalanced dataset and the real-time problem, to have a response during the financial transaction runtime, with an improved accuracy.

References

Coalition against Insurance Fraud, "Learn about fraud", http: //www.insurancefraud.org/learn_about_fraud.htm.
2. J.L. Kaminski, Insurance Fraud, OLR Research Report, http: //www.cga.ct.gov/2005/ rpt/2005-R-0025.htm. 2004
3. FBI, Federal Bureau of Investigation, Financial Crimes Report to the Public Fiscal Year, Department of Justice, United States, http: //www.fbi.gov/publications/financial/fcs_report2007/financial_crime_2007.htm.(2007)
4. CULS, Cornell University Law School, White-Collar Crime: an overview, http: // topics.law.cornell.edu/wex/White-collar_crime (2009)
5. Ngai E, Hu Y, Wong Y, Chen Y, and Sun X The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature Decision Support Systems Volume 50, Issue 3, p559-569 (2011)
6. Z.Gao, M.Ye, A framework for data mining-based anti-money laundering research, Journal of Money Laundering Control 10 (2), p170–179

(2007)
7. StijnViaene a , Mercedes Ayuso b , Montserrat Guillen b,*, Dirk Van Gheel c , Guido Dedene, Strategies for detecting fraudulent claims in the automobile insurance industry, Elsevier, European Journal of Operational Research Volume 176, Issue 1, p565-583 (2007)
8. Jean Pinquet Mercedes Ayuso Montserrat Guill´en, Selection bias and auditing policies for insurance claims, The Journal of Risk and Insurance, Vol. 74, No. 2, p425-440(2007)
9. Ll. Bermudez, J.M. Perez, M. Ayusoc , E. Gomez , F.J. Vazquez, A Bayesian dichotomous model with asymmetric link for fraud in insurance, Elsevier, p779- 786(2007)
10. W.H. Beaver, Financial ratios as predictors of failure, Journal of Accounting Research 4 p71−111. (1966)
11. Ravisankar P, Ravi V, Raghava Rao G, and Bose, Detection of financial statement fraud and feature selection using data mining techniques, Elsevier, Decision Support Systems Volume 50, Issue 2, p491-500 (2011)
12. (Date last accessed 15-July-2014). Online Available: http://www.acfe.com/uploadedfiles/acfewebsite/content/documents/rttn-2010.pdf.
13. D.Zhang, L.Zhou, Discovering Golden Nuggets: Data Mining in Financial Application, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) (Volume: 34, Issue: 4), p513-522 (2004)
14. E.Kirkos, C.Spathis, Y.Manolopoulos, Data Mining techniques for the detection of fraudulent financial statements, Elsevier, Expert Systems with Applications Volume 32, Issue 4, p995- 1003(2007)
15. B.Hoogs, T.Kiehl, C.Lacomb, D.Senturk, A genetic algorithm approach to detecting temporal patterns indicative of financial statement, InterScience, Intelligent systems in accounting, finance and mamagement, Volume15, Issue1-2, p41-56 (2007)
16. M.Cecchini, H.Aytug, G.Koehler, P.Pathak, Making words work: Using financial text as a predictor of financial events, Elsevier, Decision Support Systems Volume 50, Issue 1, p164-175 (2010)
17. L. Humpherys, C. Moffitt, B. Burns, K. Burgoon, F. Felix, Identification of fraudulent financial statements using linguistic credibility analysis, Elsevier, Decision Support Systems Volume 50, Issue3, p585-594.(2010)
18. H. Glancy, B. Yadav, A computational model for financial reporting fraud detection, Elesevier, p596-601(2010)
19. R. Quinlan, C4.5: programs for machine learning, Morgan Kaufmann Publishers, Machine Learning, Volume 16, Issue 3, p235-240 (1994)
20. B.Bai, J.Yen, X.Yang, False financial statements: characteristics of china' s listed companies and cart detection approach, International Journal of Information Technology & Decision Making Vol. 7, No. 2, p 339-359 (2008)
21. J.Quah, M.Sriganesh, Real-time credit card fraud detection using computational intelligence, Elsevier, Expert Systems with Applications, Volume 35, Issue 4, p1721-1732 (2008)
22. D.Sa´nchez, M.A. Vila, L. Cerda, J.M. Serrano, Association rules applied to credit card fraud detection, Elsevier, Expert Systems with Applications, Volume 36, Issue 2, Part 2, p3630-3640(2009)
23. S.Panigrahi,A.Kundu, S.Sural, A.K.Majumdar, Credit card fraud detection: A fusion approach using Dempster−Shafer theory and Bayesian learning, Elsevier, Information Fusion, Volume 10, Issue 4, p354-363 (2009)
24. C. Whitrow, D. J. Hand, P. Juszczak, D. Weston, N. M. Adams, Transaction aggregation as a strategy for credit card fraud detection, Springer, Data Mining and Knowledge Discovery, Volume 18, Issue 1, p30-55 (2009)
25. S.Bhattacharyya, S.Jha, K.Tharakunnel, J.C. Westland, Data mining for credit card fraud: A comparative study, Elsevier, Decision Support Systems, Volume 50, Issue 3, p602-613(2011)
26. Y. Jin, R.M. Rejesus, B.B. Little, Binary choice models for rare events data: a crop insurance fraud application, Applied Economics Volume 37, Issue 7, p841−848. (2005)
27. E.Duman, H.Ozcelik, Detecting credit card fraud by genetic algorithm and scatter search, Elsevier, Expert Systems with Applications, Volume 38, Issue 10, p13057-13063(2011)
28. N.Wong, P.Ray, G.Stephens, L.Lewis, Artificial immune systems for the detection of credit card fraud: an architecture, prototype and preliminary results, Information Systems Journal, Volume22, Issue1, p53-76(2012)
29. Y.Sahin, S.Bulkan, E.Duman, A cost-sensitive decision tree approach for fraud detection, Elsevier, Expert Systems with Applications, Volume 40, Issue 15, p5916-5924 (2013)
30. D.Olszewski, Fraud detection using self-organizing map visualizing the user profiles, Elsevier, Knowledge-Based Systems, Volume 70, p324-333 (2014)
31. N.S.Halvaiee, M.K.Akbari, A novel model for credit card fraud detection usingArtificial Immune Systems, Elsevier, Applied Soft Computing, Volume 24, p40-49 (2014)
32. A.Mubalik (Mubarek) , E.Adali, Multilayer Perception Neural network technique for fraud detection, IEEE, Computer Science and Engineering (UBMK), International Conference, p383-387 (2017)
33. M.Zanin, M.Romance, S.Moral, R.Criado, Credit card fraud detection through parenclitic network analysis, arXiv; 1706.01953v1, p1-8 (2017)
34. M.Tavallaee, E.Bagheri, W.Lu, A. Ghorbani, A Detailed Analysis of the KDD CUP 99Data Set, IEEE, Computational Intelligence in Security and Defense Applications (CISDA 2009).
35. L.Dhanabal, Dr. S.P. Shantharajah, A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 6, p446-452 (2015)
36. Y.Dai, J.Yan, X.Tang, H.Zhao, M.Guo, Online Credit Card Fraud Detection: A Hybrid Framework with Big Data Technologies, IEEE, Trustcom/BigDataSE/ISPA, p1644-1652 (2016)
37. O. Adewumi, A. Akinyelu, A survey of machine-learning and nature-inspired based credit card fraud detection techniques, Springer, International Journal of System Assurance Engineering and Management, Volume 8, Supplement 2, p937−953 (2016)
38. N.Malini, M.Pushpa, Analysis on Credit Card Fraud Identification Techniques based on KNN and Outlier Detection, IEEE, Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), Third International Conference (2017)
39. S.Askari, A.Hussain, Credit Card Fraud Detection Using Fuzzy ID3, IEEE, Computing, Communication and Automation (ICCCA), p446-452 (2017)
40. C.Li, D.S. Poskitt, X.Zhao, The Bivariate Probit Model, Maximum Likelihood Estimation, Pseudo True Parameters and Partial Identification, http://business.monash.edu/econometrics-and-business-statistics/research/publications, p2-34 (2016)
41. C.G. Atkeson, A.W. Moore, S. Schaal, Locally weighted learning for control, Artificial Intelligence Review 11, p1−5 (1997)