Evaluating Case-Based Reasoning Knowledge Discovery in Fraud Detection

Adeyinka Adedoyin, Stelios Kapetanakis, Miltos Petridis, and Emmanouil Panaousis

University of Brighton
{a.adedoyin,s.kapetanakis,m.petridis,e.panaousis}@brighton.ac.uk

Abstract. The volume of banking transaction has increased considerably in the recent years with advancement in financial transactions payment methods. Consequently, the number of fraud cases has also increased, causing billion of dollar losses each year worldwide, although from Literature, there has been substantial work in the domain of fraud detection by both the industry and academia's. Despite the substantial work, there are few researches in applying case-based reasoning (CBR) approach in the context of detecting Financial Fraud. In this paper we aim at evaluating the performance of CBR in Identifying fraudulent patterns among financial transaction by comparing it with logistic regression (LR) and neural network (NN) which are often used in many related work. To evaluate our approach simulated data, based on a sample of real anonymous transaction provided by a bank was used and the result shows that LR outperformed NN and CBR model, with a steady increase in precision, sensitivity and specificity as the percentage ratio for the training and test data were varied. This was due to the linearity, fuzziness and presence of uncertainty in the sampling dataset. Therefore, we can reach a conclusion that part of the possible reasons why there are few research in applying CBR to the context of detecting financial fraud patterns may be due to incomplete information, fuzziness and uncertainty in the available data sets used for experimentation.

Keywords: Fraud Detection, Case-Based Reasoning and Multi-intelligent system.

1 Introduction

Financial Fraud exists for a long time and it can take an unlimited variety of forms. Consequently, with the fast growing channels of banking which have made it easier for us to communicate and carry out financial transaction with convenience, the number of fraud cases has also increased, causing billion of dollar losses each year worldwide [1]. Over the past few decades, financial institutions, government and international organization have made corresponding laws, regulations and used advanced methods to prevent and monitor such fraudulent

activities. However, most of them seems to be faint as these channels of banking and fraud detection methodology evolve; perpetrators have become more sophisticated in tandem with these improvements [2].

Furthermore, with technological advancement in the channels of banking, the financial transaction dataset production, collection and storage has dramatically increased in dimension. These dataset(s) are increasing in dimension in three ways: (i) the number of records in the database, (ii) the number of fields or attributes associated with a record, (iii) the complexity of the data itself. However, extracting pertinent knowledge from such complex databases in a search for fraudulent activities calls for or requires more than mere novelty of statistical model, to the use of fast and efficient Artificial Intelligence Techniques [3]. Traditionally in the past, statistical tools such as univariate statistical models, Multiple Discriminant analysis, Linear Probability Models, Logistic Regression and Probit analysis have been applied to financial fraud detection for years. These methods have been proven to be effective with small sample sizes and when theory or experience indicates an underlying relationship between dependent and predictor variables [4]. In addition, these statistical models do require a few assumptions and are usually constrained by their demand for data linearity, which makes it fairly difficult to process massive and complicated data. However, to get rid of some of these cumbersome requirements of statistical methods professionals both in the industry and academics in related field have adopted more alternative Artificial Intelligence methods such as Neural Networks in [5], Data Mining Techniques [6], and Genetic Algorithm [7]. "These techniques, although common and quite widely used do present some hard to avoid downsides" [8].

Case-based reasoning an emerging technology has grown from rather specific and isolated research area to a field of widespread interest [9]. However, there are few research in applying these technique in the context of detecting financial fraud patterns. According to Kapetanakis et. al. [8] the possible reasons may be the focus of the relevant literature on optimising existing approach, lack of maturity of CBR research with reference to the transaction application scope and lastly, the established view of Financial Fraud Detection problem as one seeking precision optimisation, rather than seeking new ways of identifying and representing activity patterns. This provides motivation for exploring the performance of CBR methodology in financial fraud detection.

The contribution of this paper is to evaluate the performance of CBR as a knowledge discovery tool in Identifying fraudulent patterns among financial transaction. The remainder of this paper is structured as follows: Section 2 will give an overview of related work; Section 3, explains the methodology followed throughout this research and Section 4, will present the experimental results. Finally, Section 5 concludes this by summarising the outcomes and future work.

2 Related Work

Artificial intelligence (AI) techniques have been successfully applied to fraud detection and credit scoring, and the field of AI has applied to the financial

domain is both well-developed and well documented. As an emerging methodology, case-based reasoning (CBR) is making a significant contribution to the task of fraud detection. CBR systems are able to learn from sample patterns of credit card use to classify new cases, and this approach also has the promise of being able to adapt to new patterns of fraud as they emerge [10]. As applied to the financial domain, CBR systems have a number of advantages over other AI techniques such as a reduction in knowledge-elicitation effort from complex and complicated transaction situations, the ability to learn by acquiring new cases over time without having to add new rules or modify existing ones, and the ability to provide justification by offering past cases as precedence rather than justifying a solution by showing a trace of the rules that led to decision [11,12].

Kapetanakis et. al. [8], applied CBR Financial Transactions Intelligent Monitoring System (named CBR-FTIMS) to demonstrate the use of a CBR workflow approach in identifying abnormal financial transactions. They showed that CBR and workflow representation can be applied successfully over a case base of transactions, where classification has been applied in advance, and contribute to the ranking of an unknown cluster of cases. However, applying simplified CBR generates high number of false positive alarm.

Cheol-Soo et. al. [13], proposed an analogical reasoning structure for feature weighting using a new framework called the analytic hierarchy process (AHP)-weighted k-NN algorithm. The AHP-weighted k-NN algorithm use hierarchical or network structures to represent a decision problem and then develop priorities for the alternatives based on the decision maker's judgements throughout the system. This addresses the issues of how to structure a complex decision problem, identify its criteria (tangible or intangible), measure the interaction among them and finally synthesize all the information to arrive at priorities, which depict preferences. The proposed AHP weighted k-NN model was used to perform an intelligent system for bankruptcy prediction. The proposed AHP weighted k-NN algorithm achieved classification accuracy higher than the pure k-NN algorithm. However, the CBR modeling is not sufficient, since techniques with multi-step, meta-reasoning are required.

Wheeler et. al. [10], applied a Multi-agent Case-based reasoning approach to the problem of reducing the number of final-line fraud investigation in credit approval process. From the results, the adaptive CBR algorithm was found to have the best performance, and these results indicate that an adaptive solution can provide fraud filtering and case ordering functions for reducing the number of final-line fraud investigations necessary. However the model needs to be tested with similarly complex data sets from other real world domains.

3 Proposed Approach

This section will describe the methodology adopted for this study in terms of the classification models used, experiment data and it's processing. According to the findings of related research, neural networks (NN) and logistic regression (LR) are often used in many banking related knowledge discovery fraud activities.

This has given them a well established popularity and ability to be used as a control method by which other techniques are tested [14].

3.1 Logistic Regression

Logistic regression technique is a widely used statistical model used in solving diverse classification algorithm problems. It has been widely applied in different domains and it is used for modeling the relationship between a categorical outcome variable(s), which is usually dichotomous, such as a credit card transaction being fraudulent or non-fraudulent and a set of predictor variables [15].

Let y_i be the dependent variable with outcome $y_i = 1$ (for fraudulent transaction), probability P_i and $y_i = 0$ (non-fraudulent transaction) with probability $1 - P_i$. The probability P_i is then modeled in relation to the predictor variables. Therefore, relating the logistic regression model to the probability that a transaction is fraudulent:

$$\text{logit}(P_i) = \beta_0 + \beta_1 x_1 i + \beta_2 x_2 i + \ldots + \beta_k x_k i \tag{1}$$

Where $x_1, x_2, \ldots\ldots, x_k$ are the predictor variables and P_i is the probability that a transaction is a fraudulent $y_i = 1$. β_0 is a constant and $\beta_1, \beta_2, \ldots, \beta_k$ are coefficients of the dependent variables y_i. i is the number of observed cases in the dataset.

Transforming P_i:

$$\text{logit}(P_i) = \log\left[\frac{Pi}{1 - P_i}\right] \tag{2}$$

The regression coefficients β_k are derived by means of Maximum likelihood estimation (MLE).

3.2 Neural Network

Neural network (NN) is an adaptive system that is designed to model the way in which the human brain performs a particular task or function of interest using electronic component or software simulation. NN topologies are made up of neurons, which are linked together in layers with modifiable weighted interconnections. It also has the ability to modify its topology which is motivated by the fact that the human brain can die and new synaptic connections can grow [16]. In this study, Multilayer Perceptron (MLP) architecture and error backpropagation algorithm was applied to minimize the error at output network and to compute the error for the experiment sigmoid function was applied. In a MLP experiment the number of hidden layers and neurons in each layer has significant influence on the performance of network. When a small number of nodes is used, it makes it insufficient in generating a generalize rules for the training sample, while more number of nodes in the hidden layer increase the power and flexibility of the network for identifying a complex patterns. However, an overly large hidden layer leads to over fitting and memorizing the training set [17]. In order

to determine the number of hidden neurons, experiments with various values of neuron was carried out and the neuron yielding the best accuracy was then used for evaluation of the test set. The weights during the experiments were generated randomly between the range $[-1, 1]$ and the termination criteria used was two hundred and fifty iterations of the network. Once the termination criteria is met the training stops and the network is tested with the test data.

3.3 Case-Based Reasoning

Case-based reasoning methodology can be used as a classification technique; it classifies an unlabeled case by retrieving closely matching labelled cases and reusing their labels. In the study we use the $k-$nearest neighbor $(k-NN)$instance-based learning for the case classification. The $k-NN$ requires defining the case representation and the similarity function, which may employ algorithms for feature selection or weighting [18]. In order to compute the similarity between the input dataset and previously experienced case instances, three $(k = 3)$ neighboring datasets were chosen and the distance metrics was define using Euclidean distance.

$$\mathrm{D}(x, y) = \sqrt{\sum w(x_i - y_i)^2} \tag{3}$$

Where D is the Euclidean distance between the new case x_i and retrieved similar case y_i from the training dataset. While w is the weights, it represents the importance of the attributes for comparing the two cases. The weights for the attributes can be standardized and represented as numerical values between 0 and 1. However, in this study the weights are assumed to be normalized and are of equal importance.

3.4 Experiment Dataset

Generally, there are no publicly available data sets for studying fraud detection within the financial service sector. Obtaining real data from companies for research purposes is almost impossible due to legal, corporate policies and competitive reasons. However, researchers in the past circumvented theses availability problems by simulating data which matches closely to actual data. Barse et. al [19] justifies that synthetic data can train and adapt a system without any data on known frauds, variations of known fraud and new frauds can be artificially created, to a specific environment. The proposed approach was tested using a simulated dataset known as BankSim by [20]. The BankSim is a Bank payment Simulation, based on a sample of aggregated transaction data provided by a bank in Spain. This data contains several thousand logs of transaction data covering six months, from November 2013 to April 2013. The aggregated data was obtained by using three types of queries: (1) Consumption habits; (2) Customer classification and origin; and (3) Source of Transactions. The simulation generated 587,443 normal and 7,200 fraudulent transactions, with total amount of stolen money summing up to around 3.8 million Euros which corresponds to 17% of the total amount of payments. The data set is highly imbalanced, with ratio of 98.78% : 1.21% non-fraudulent to fraudulent transactions respectively.

3.5 Data Pre-processing

To be aligned with the literature, we selected the most popular features in our sampling dataset.The data contains some missing values and outliers; the missing values were generated using classification and regression tree algorithm ($C\&RT$) with 50% of the sample size. To optimise the performance of the models, 10-fold cross validation was applied to the data set and all the three classifiers were experimented. The data sets were partitioned into training set for training the model and test set for testing the model. The percentage of the training and testing data was varied in order to study the variations of performance caused by changing the ratio of training to testing partitions of the dataset. For the selection of samples in training the percentage of each class was balance statistically, while for testing portions the percentage of each class in each portion is preserved. The training and testing portions used were in these ratios 10:90, 30:70, 50:50, 70:30 and 90:10. The average error for all the ten folds was computed and the performance of each model was measured by using the performance metrics mentioned in the next section.

4 Results and Evaluation

The performance of computational intelligent systems can be measured in many different ways such as absolute ability, probability of success, visual mediums and more. A number of these performance metrics were identified from literature [14]. Accuracy and Area Under Curve (AUC) are among the most widely used classifier performance evaluator, however they are not adequate enough for fraud detection problems where there is significant class imbalance between the non-fraud and fraud cases [17]. In the study, since part of the aim is to reduce the number of final line case investigation by the investigator, true positive and false positive rate metrics was adopted. In order to balance the trade-off for the class imbalance sensitivity, specificity and precision was also used in the experiments.

4.1 False Negative, False Positive Rate and Accuracy Evaluation

This section presents results from our experiments in terms of how the classifiers were able to accurately identify the value of normal and fraudulent transactions. The false negative rate is the number of transactions that are fraudulent but mistakenly classified as normal, while false positive rate is the number of normal transaction that are mistakenly classified as fraudulent. Fig. 1 shows the comparison with 10-fold cross validation of the training and test data variations:

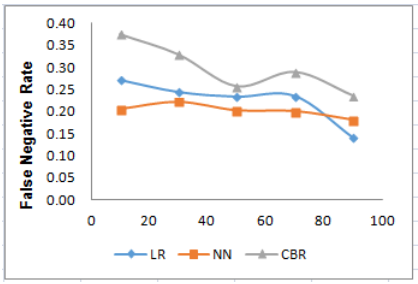

Fig. 1. False Negative Rate (FNR) of the different classifiers

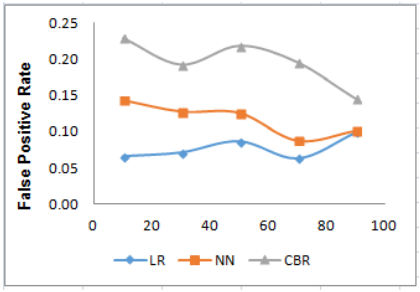

Fig. 2. False Positive Rate (FPR) of the different classifiers

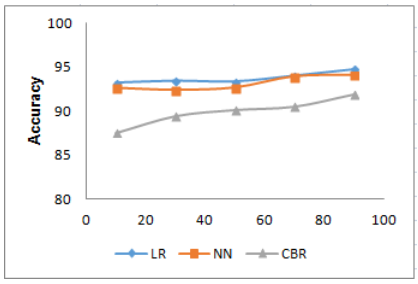

Fig. 3. Accuracy of the different classifiers

Fig. 1 compares the false negative rate for the three classifiers. The false negative rate for LR and CBR declined suddenly as percentage ratio for the training dataset was increased, while NN declined steadily. Also, the false negative rate for both LR and CBR suddenly increased dramatically when 70% of the dataset was used for the training and continued to decrease as the percentage ratio is increased. Therefore, NN has shown a better performance in correctly classifying fraudulent transactions.

Fig. 2 compares the false positive rate for the three classifiers. The false positive rate suddenly increased dramatically for the three classifiers when 50% of the dataset was used for the training, continued to decrease as the percentage ratio is increased to 70% for LR and NN. While on the other hand after the sudden increase at 50%, the false positive rate for CBR decreased steadily as the percentage ratio is increased. Therefore, LR and NN has shown better performance in correctly classifying normal transactions. Fig. 3 shows results from the comparison of the classification accuracy for the three classifiers. In over all, the results obtained shows that LR has a better performance in fraud detection with steady increase as the percentage ratio is increased compared to NN and CBR.

4.2 Precision, Sensitivity and Specificity Evaluation

This section presents results from our experiments comparing the performance of Logistic regression (LR), Neural network (NN), and Case-based reasoning (CBR) classifiers developed from 10-fold cross validation of the training and test data. Results are shown in Fig. 4, Fig. 5 and Fig. 6 for the performance evaluation metrics used in measuring classification performance. The performance was measured on the problem of measuring classification performance, while balancing the trade-off for the class imbalance in the results.

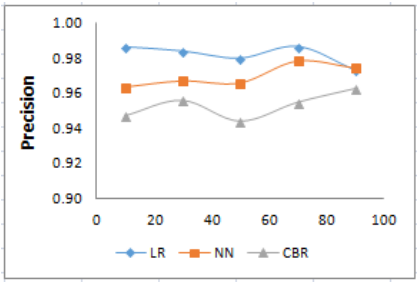

Fig. 4. Precision of the different classifiers

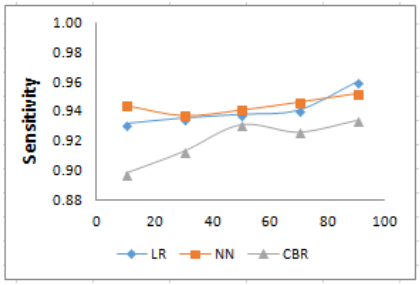

Fig. 5. Sensitivity of the different classifiers

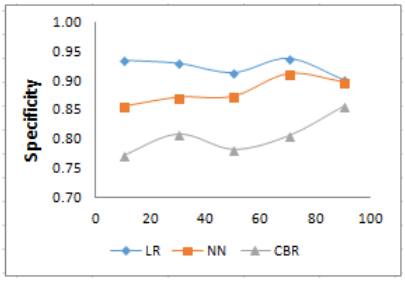

Fig. 6. Specificity of the different classifiers

From Fig. 4, Fig. 5, Fig. 6 respectively it can be seen that variation of the percentage ratio for the training dataset has significant influence on the behavior pattern of the various performance metrics used. Case-based reasoning shows a low precision, sensitivity and specificity rate as the percentage ratio for the training dataset in increased. It can be observed that LR shows overall better performance. The steady increase in the sensitivity of LR as the percentage ratio for the training dataset was increased indicates that LR classifier maintained a similar ranking of cases, irrespective of the level of under sampling of non-fraudulent cases in the training data.

5 Conclusions

In this paper we proposed a Multi-intelligent Fraud Detection System using logistic regression (LR), neural network (NN), and case based reasoning (CBR). To prove the efficiency of our method, we used synthetic simulated data in evaluating their performance. The recognition performance shown by Logistic regression classifier is better compared to NN and CBR, with a steady increase in precision, sensitivity and specificity as the percentage ratio for the training and test data was varied.

During the analysis, it was observed that the data set used is characterized with Linearity, incomplete information, fuzziness and uncertainty which could be due to the sensitivity, legal, corporate and societal impact of the having confidential information such as this in the public domain. These makes it difficult in exploring most fraud detection issues in greater depth, particularly with focus on tracking and monitoring transaction sequence with the intent of identifying the similarities and characteristics of different types of fraud using controlled experiments. Therefore, we can reach a conclusion in addition to the suggestion made by [8] that part of the reason why there are few research in applying CBR to the context of detecting financial fraud patterns can be due to incomplete information, fuzziness and uncertainty in the available data sets used for experimentation. For the future work, we plan to model some of the fuzziness and uncertainty to create a knowledge pool of different types of Fraud patterns and apply CBR in computing the similarities and characteristics of the constructed case base using controlled experiments.

Acknowledgement. We would like to thank Dr Edgar et. al [20] for providing us with a data set that was used in our experiment analysis.

References

1. Bolton, J., Hand, D.J.: Statistical Fraud Detection A Review: Statistical Science, 17(3):235-249 (2002).
2. Yang, Q., Feng, B., Song, P.: Study on anti-money laundering service system of online payment based on union-bank mode.In:International Conference on Wireless Communications, Networking and Mobile Computing, WiCOM , pp. 4986-4989, (2007).
3. Ekin, T., Leva, F., Ruggeri, F., Soyer, R.: Application of Bayesian methods in detection of healthcare fraud. Chemical Engineering Transactions, Vol.33, pp.151-156, (2013).
4. Razi, M., Athappilly, K.: A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models. Expert Systems with Applications, 29(1):65-74 (2005).

5. Maria, K., Chris, C., Michalis, A.: Neural Network the Panacea in Fruad Detect? Emerald Insight, 25(7):659-678, (2010).
6. Efstathios, K., Charalambos, S., Yannis, M.: Data Mining techniques for the detection of fraudulent Financial statements. Expert Systems with Applications, 32(4):995-1003, (2007).
7. Chuang, C.L.: Application of hybrid case-based reasoning for enhanced performance in bankruptcy prediction. Information Sciences, 236:174-185, (2013).
8. Kapetanakis, S., Samakovitis, G., Gunasekera, B., Petridis, M.: Monitoring Financial Transaction Fraud with the use of Case-based Reasoning, In: 17th UK Case-Based Reasoning Workshop, Cambridge, UK (2012).
9. Aamodt, A., Plaza, E.: Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches; AI Communications, 7(1):39-59, (1994).
10. Wheeler, R., Aitken, S.: Multiple algorithms for fraud detection; Knowledge-Based Systems Elsevier, Vol.13,pp. 93–99 (2000).
11. Chi, R. H., Kiang M.Y.: An integrated approach of rule-based and case-based reasoning for decision support.In: Proc. 19th ACM Annual Computer Science Conference pp. 255–267, (1991).
12. Watson, I.: Applying Case-based Reasoning: Techniques for Enterprise System, Morgan Kaufman, San Mateo, CA (1997).
13. Cheol-Soo, P., Ingoo, H.: A Case-Based Reasoning with the feature weights derived by analytic hierarchy process for bankruptcy prediction, Expert Systems with Applications Vol. 23, Iss. 3, pp. 255–264, (2002).
14. Jarod .W, Maumita .B, Rfiqul .I: Intelligent Fraud Detection Practices: An Investigation. Charles Sturt University, Australia (2014).
15. Archer K.J, Leweshow .S: Goodness-of-fit test for a logistic regression model fitted using survey sample Data. The Stata Journal, Vol.6, No.1, pp. 97-105,(2006).
16. Haykin .S, , Neural Networks: A Comprehensive Foundation. Pearson Prentice Hall, 2nd edition, (1999).
17. Bhattacharyya et. al.: Data mining for Credit Card: A Comparative Study. Decision Support Systems vol.50, pp. 602-613, Elsevier,(2011).
18. McDowell et. al.: Case-Based Collective Classification. American Association for Artificial Intelligence (2007).
19. Barse, E., Kvarnstorm, H., Jonsson, E.: Synthesizing test data for fraud detection systems. 19th Annual Computer Security Application Conference, pp. 265-277, Springer (2003).
20. Lopez-Rojas .E, Axelsson .S: BankSim: A Bank Payments Simulator for Fraud Detection Research. Blekinge Institute of Technology, Sweden (2014).