TECH-KNOWLEDGE JOURNAL

Fraud Classification and Detection Model Using Different Machine Learning Algorithm

Noor Khalid Hussein*, Ayad Rodhan Abbas, Bashar Saadoon Mahdi
Department of Computer Science, University of Technology, Baghdad, Iraq
* cs.19.82@grad.uotechnology.edu.iq

Keywords: **Classification, Machine Learning, Credit Card, Fraud Detection, WEKA.**

ABSTRACT

Recently, fraud technologies have become more advanced and easier to fraud. Therefore, different machine learning techniques have been applied and developed to recognize fraudulent credit card transactions. The main problem to fail any detection techniques on any fraud operation is the accuracy of results. This paper discusses how to improve fraud detection performance using machine learning algorithms by choosing the most appropriate algorithm for inclusion in fraud detection systems. It also provides a comprehensive study of Taiwan's customer database and how classifiers interact with it by applying 30 different classification algorithms. Moreover, using the WEKA tool for applying machine learning algorithms with the voting method to choose the right classification. The experimental results reveal that using the LMT algorithm will be the best one where achieved 82.0867 % accuracy.

1.0 INTRODUCTION

Fraud is a label for any malicious activity that wants to harm people by stealing their money, identity or other things, while digital money has become common use in these days, the fraud operations also become more powerful and more efficient to keep pace with this development [1]. For any conducting electronic financial transactions (digital money) in the last few years, credit card has become one of the widely used to that purpose, to protect these credit cards the systems for any institution like banks, companies or other else must have high security techniques to detect any fraud operations [2]. The protect operations of credit card have a big challenge to prevent fraud occurring on them and machine learning one of the wildest techniques that used to detect fraud transactions on a credit card.

Machine learning is abroad scientific field which is based on concepts of computer science, mathematics, statistics, engineering, and many other fields of mathematics and science. The main idea of using machine learning methods is to recognize if the transactions are fraudulent or not, there is a four type of machine learning supervised learning, Semisupervised Learning, Unsupervised Learning, and Reinforcement learning, in this research focus on supervised learning [3], supervised learning depends on the historical behavior of transaction data for all users in the system to predict the rules base, this rule base is used to check any new transaction and defined it either fraudulent or safety transaction [4]. To apply machine learning algorithms there are many tools like ORANGE, O3, WEKA, etc. However, this research uses WEKA as a tool to apply different types of machine learning algorithms.

Waikato Environment for Knowledge Analysis (WEKA), the Java™ programing language is used to written code for WEKA is a collection of algorithms of machine learning that deal with data mining;

WEKA tool can execute data by one of these ways [5, 6]. the main purposes of using WEKA is the ease of implementation, saving time, contain all algorithms including old ones and new ones. Thirty different classification algorithms on our credit card dataset are used and the comparing results dependent on accuracy to find the best algorithm that achieves high accuracy. In this study, using 30 algorithms categorization in 6 categorize such as naïve Bayes, decision trees, rules, lazy, meta, and function classifiers. Then select the best classifier.

The rest of this paper, section II presents the related work, where section III presents the dataset description, our proposed method presented in section IV, result and dissection present in section V, and finally conclusion present in section VI.

2.0 RELATED WORKS

Frauds are increasing significantly because of this, every day the losses increase more and more this led to an increase in fraudrelated studies to reduce them as much as possible, some of the studies like John O.Awoyemi [7] use three techniques KNearest Neighbor, Naive Bayes, and Logistic Regression techniques in binary classification for imbalance fraud data of credit card. and comparing results according to accuracy, sensitivity, and Matthews's correlation coefficient (MCC). Yashvi Jain [1] explained more than the first study and use Artificial Neural Networks (ANN), Support Vector Machine (SVM), Bayesian Network, Hidden Markov Model, K Nearest Neighbour (KNN), Decision Trees and Fuzzy Logic Based System. Yashvi Jain measured accuracy, false alarm rate, and detection rate among them.

Sangeeta Mittal [8] try to evaluate the performance of some machine learning techniques, these techniques some of them supervised and other are unsupervised, the evaluation depends on the ability of the algorithm to find frauds correctly, for this research Random Forest (RF), Neural Networks (NN), Deep Learning (DL), Support Vector Machine (SVM), Naive Bayes (NB), Logistic Regression (LR), Extended Gradient Boosted Tree (XGBT), Quadratic Discriminant Analysis (QDA) and KNearest Neighbour (KNN) are used. Heta Naik [9] analyze an online dataset and comparative accuracy for algorithms Naïve Bayes, AdaBoost, Logistic regression, and J48.

Dejan Varmedja [2] uses various algorithms of machine learning, like Multilayer Perceptron (MLP) (LR), Naïve Bayes (NB), Random Forest (RF) and Logistic Regression to find the suitable algorithm for fraud operations on a credit card. In this work using all algorithms in the WEKA tool, these algorithms are 30 and select the best one, as presented in this section there are no researchers used all these algorithms on this dataset.

3.0 DATASET DESCRIPTION

This The database used in this research was taken from the website https://archive.ics.uci.edu. The database is made up of data for clients in Taiwan. This database is characterized by being the most diverse among databases used for fraudulent purposes. In addition, it contains many challenges, which makes most researchers avoid using it. This dataset has size 5.28 MB (5,539,840 bytes) and its contents 25 different features with a different type of represented. The dataset has two label classes and has 30000 records as described in Table 1.

Table 1: Dataset structure

No	Features		
	Features	*Data type*	*Description*
1	Customer ID	Auto number	Sequins no for each user.
2	Bank account	Number	how many have in credit card (NT dollar) for both individual and family Accounts (one person in the family and the family).
3	Gender	Number	"1" for male, "2"for female.
4	Education	Number	"1" for school, "2" for collage , "3" for high school , "4" for other statues.
5	Social state	Number	"1" for married, "2" for single, "3" for other statues.
6	Age	Number	Represented by year NO.
7	history for last "6" payments monthly	Number	Six Features columns represent six last month start from September to April (NT dollar).
8	bill statement amount	Number	Six Features columns represent six last month start from September to April (NT dollar).
9	the amount for previous payment	Number	Six Features columns represent six last month start from September to April (NT dollar).
10	Classes	Number	"1" for yes, "0" for no.

4.0 PROPOSED METHODOLOGY

In this section, introduce the proposed methodology which can be referred to as Figure 1. The idea is firstly preprocessing, and feed the split data then split our data into training and testing, in this study for split the data using 10fold cross validation method. Then build our model used six methods such as naïve Bayes, decision tree, rules, lazy, meta, and function classifier all these methods in WEKA tool. the proposed method is evaluated using an evaluation metric to check the prediction for each method then select the best method for the Fraud detection problem.

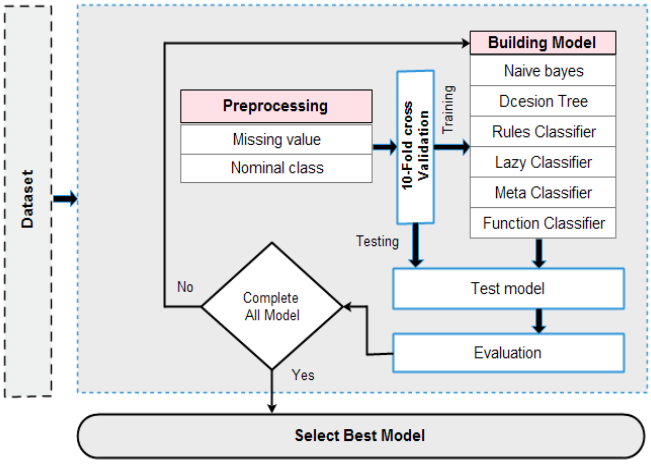

Fig. 1: Proposed method

4.1 Preprocessing

Preprocessing is one of the most important steps for any dataset to enhance the results. Data preprocessing can affect hardly the performance of any supervised machine learning algorithms. In this study, applying replace missing values and numerical class to nominal class. Missing value occurs when the value could not be recorded, the value did not have relevant with a particular case, or the user just ignored them [10]. In this work, replace missing value by used mean and modes. Some classification algorithms work with a numerical class, nominal class, or both [11]. In this work change the class from numerical to nonmale as original classes are 1 and 0 where change 1 to yes and 0 to no.

4.2 Building Model

We building 30 algorithms for every categorize type in WEKA. Every model was tested, results were derived and the best model was selected. Thereafter, they evaluated every model and compared the results of each algorithm and determining the best model that could be used.

Classification techniques on of the most popular method that used in machine learning and deals with different types of the dataset, the main goal for use Classification Techniques to classified data accordion to some conditions to be more useful in detection operations. The data will be more understand and more useful to predict events, rules, and so on. Classification techniques contain two main approaches supervised learning technique that can detect fraud operation dependent on fraudulent patterns of the last transactions and unsupervised technique that can detect fraud operation dependent on comparative and testing computed data to find unexpected transactions the details of it describe by Fabrizio Carcilloa, YannA¨el Le Borgne, Olivier Caelenb, Yacine Kessacib, Fr´ ed´ eric Obl´eb, Gianluca Bontempi [12]. Each supervised and unsupervised learning techniques have a lot of algorithms that deal with the dataset differently, in this part will describe the main techniques used in this research.

4.2.1 Naive Bayes

Naive Bayes is a type of algorithms in machine learning, Naive Bayes is a supervisor algorithm and very widely used for classification, also it a simple to understood and use. The basis for this algorithm depends on the probabilistic theory[13]. For more details of Naive Bayes as described [14]. as shown in equation 1.

$$p(c / f) = \frac{p(f / c)}{p(f)} \quad , \text{P (f)>0}$$

$$p(c / f) = \frac{p(f / c)}{p(f)}$$

where in f refers to the features, and c denotes the class.

4.2.2 Decision Tree

Decision trees it just like our brains when trying to decide so it's easy to understand, simple to belting. The decision tree techniques work with a continuous or independent data set and give all possible solutions. Decision trees have the same contents of a real tree, that contains is (root node, branches, and left nodes) [15]. Many algorithms of decision trees are used in this paper such as (J48, Decision stump, Hoeffding tree, J48, LMT, Random forest, Random Tree, and REP Tree) [16].

4.2.3 Rules Classifier

Rule classifier used to predict continues to rule, the metric depending on "AND" logical method for linking attributes together, rule classifier use Information Gain to Reduced Error Pruning (REP). the rule in classification can be in different forms but they are ordered, so whenever the first rule is fired the classification stop the possessing and find results [17]. Rule classifier has many algorithms some of them used in this paper like (JRip, Decision table, One R, PART, and Zero R) [18].

4.2.4 Lazy Classifier

The lazy classifier is used with difficult decision spaces and polygonal shape that other algorithms can't be explained easily but it's expensive technique because it works with parallel hardware and storage with efficient technique, these requirements are needed to build the system and its expensive. Lazy classifier gives little information about data structure [5] There are many algorithms of Lazy classifier techniques used in this paperlike (IBK, KStar, and LWL).

4.2.5 Meta Classifier

Meta classifier its many classifiers marred together, that because it split the dataset into many training subsets, each training subsets will be classified with different classifiers, at the last, these results will integrate to find the final result of meta classifier. meta classifier used if the dataset contains attributes with any number of values, by using meta classifier space and time complexities will decrease [19, 20]. Different types of algorithms used in this paper (ADaBoostM1, Bagging, CVParameter Selection, Logit Boost, Stacking, Multischeme, Multiclass classifier, Multiclass classifier updatable, and Vote)[21].

4.2.6 Function Classifier

Function classifier depending on neural network and regression. Accuracy in this model depends on the amount of training data, biological neurons is the base of built statistical models. Some of these types of classifiers used raw input data and some of them tend to be overtraining [22]. The algorithms used in this paper are (Logistic and Voted perceptron).

4.3 Evaluation Metric

Researchers have used the same validation method for determining the classifier percentages. They used the dataset which had an approximately similar size and class distribution. For every fold, the classifier was trained with the help of the 10. Here, the researchers have explained the performance measurements used for the machine learning classification issue[23, 24]. According to the confusion

matrix, several measurements could be used for examining the performance of the model with regards to the accuracy, which was determined using the below mentioned in Table 2. The recall was used for determining the accuracy of every class known. Precision was also inaccurately classified using the equation below. This helped in calculating the F1 scores.

Table 2: Metric equations

Metric name	**Equation**
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Recall	$\frac{TP}{TP+FN}$
Precision	$\frac{TP}{TP+FP}$
F1score	$2*\frac{\mathrm{Re}\,call*\mathrm{Pr}\,ecesion}{\mathrm{Re}\,call+\mathrm{Pr}\,ecesion}$

Accuracy is usually determined by using the confusion matrix. Here, the confusion matrix was dependent on the choice of the datasets. The researchers used the contingency table for improving performance and accuracy.

5.0 RESULTS AND DISCUSSION

The researcher applied the WEKA tool version 3.9.4 which offered the steps of the processes like preprocessing, and the visualization and validation results. This study applied machine learning algorithms. The researcher used default parameters for machine learning as it is. Then used the confusion matrix for determining accuracy. The researchers implemented 30 tests on 25 features and 300000 instances. The result was split into six tests for each main category. Table 3 shown the accuracy of Naïve Bayes. Table III shown the first categorization of our method.

Table 3: Naive Bayes

Algorithm	Precision %	Recall %	Fscore %	Accuracy %
Bayes Net	0.795	0.810	0.771	81.0267
Naïve Bayes updateable	0.770	0.694	0.717	69.3833
Naïve Bayes	0.770	0.694	0.717	69.3833
Naïve Bayes multinomial text		0.779		77.88

Table 3 shown the naïve Bayes method, where Bayes Net achieve 81.0267 % accuracy higher than other methods. Naïve Bayes updateable and Naïve Bayes both achieve the same accuracy 69.3833%. naïve Bayes multinomial doesn't show value in both precision and Fscore because of memory limitation with the WEKA tool. Table 4 utilization decision tree method.

Table 4: Decision tree

Algorithm	Precision %	Recall %	Fscore %	Accuracy %
Decision stump	0.803	0.820	0.794	81.96
Hoeffding tree	0.796	0.815	0.796	81.4967
J48	0.779	0.800	0.782	80.04
LMT	0.804	0.821	0.799	82.0867
random tree	0.737	0.735	0.736	73.5067
Randomforest	0.801	0.818	0.798	81.84
REP Tree	0.793	0.813	0.792	81.2667

Table 4 shown LMT the higher accuracy where achieved 82.0867%. the lower accuracy achieved 73.5067% with random tree. Other algorithm almost achieved same accuracy 81% except J48 algorithm where achieved 80.04%. The following Table 5 shown the rule classification methods.

Table 5: Rule classification

Algorithm	Precision %	Recall %	Fscore %	Accuracy %
decision table	0.802	0.820	0.798	81.95
JRip	0.801	0.818	0.798	81.8333
oneR	0.803	0.819	0.793	81.9233
Part	0.794	0.814	0.792	81.3733
zero R	-	0.779	-	77.88

Table 5 shown almost the same accuracy 81% except with Zero R where achieved 77.88% accuracy. The decision table achieved 81.95% high accuracy than others. The following table utilized Lazy classification methods.

Table 6: Lazy classification

Algorithm	Precision %	Recall %	Fscore %	Accuracy %
IBK	0.728	0.730	0.729	72.9533
kStar	0.698	0.712	0.705	71.2267
LWL	0.801	0.819	0.796	81.86

Table 6 demonstrates the LWL algorithm achieved 81.86 higher accuracies than others. The following table is shown Meta methods that we utilization in this study.

Table 6: Meta classification

Algorithm	Precision %	Recall %	Fscore %	Accuracy %
ADaBoostM1	0.801	0.818	0.791	81.81
Bagging	0.800	0.818	0.798	81.78
CVParameterSelection	-	0.779	-	77.88
LogitBoost	0.800	0.817	0.789	81.6767
Multiclassclasifier	0.795	0.810	0.771	81.0267
Multischeme	-	0.779	-	77.88
multisclassclasifierUpdatable	0.792	0.810	0.772	80.9667
Stacking	-	0.779	-	77.88
Vote	-	0.779	-	77.88

Table 7 shown the ADaBoostM1 method achieved 81.81 accuracies than others. CVParameter selection, Multischeme, Stacking, and Vote are achieved the same accuracy of 77.88%. table VIII using Function classification.

Table 7: Function classification

Algorithm	Precision %	Recall %	Fscore %	Accuracy %
Algorithm	Precision %	Recall %	Fscore %	Accuracy %
Logistic	0.795	0.810	0.771	81.0267

Table 7 shown Logistic algorithm achieved 81.0267 accuracy than Votedperceptron where achieved 77.80%. the following Figure shown the model accuracy.

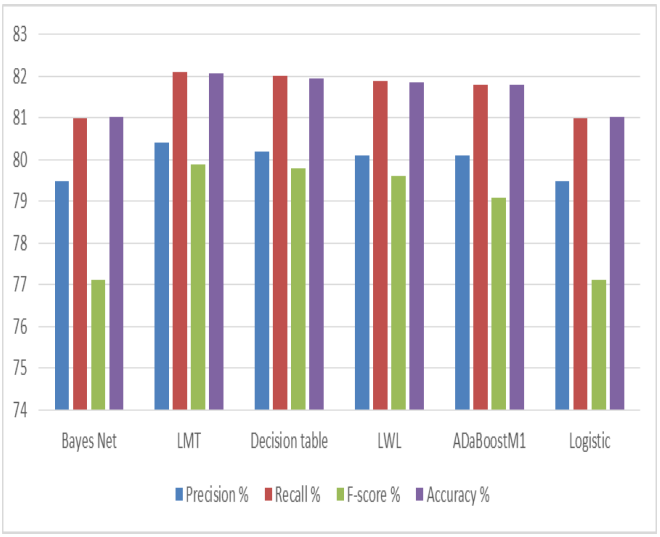

Fig. 2: Model accuracy

6.0 CONCLUSION AND FUTURE WORK

This research focus on classification methods to detect frauds operation and how choosing the right classification algorithm that fitting with the dataset can give an efficient system, choosing the right classification method lead it to improve the efficiency of the system. As a future work for this research, and after studying and knowing the results of most classification algorithms and comparing their performance to find the most suitable and best for the Taiwan database, it will be appropriate in the future to conduct a study to find a suitable improvement method applied to the algorithms with higher voting in this research, which helps in improving the performance of the algorithms and thus improving the performance of detection systems Fraud.

REFERENCES

[1] Y. Jain, S. NamrataTiwari, and S. Jain, "A comparative analysis of various credit card fraud detection techniques," Int J Recent Technol Eng, vol. 7, pp. 402407, 2019.

[2] D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic, and A. Anderla, "Credit Card Fraud DetectionMachine Learning methods," in 2019 18th International Symposium INFOTEHJAHORINA (INFOTEH), 2019, pp. 15.

[3] A. A. Soofi and A. Awan, "Classification techniques in machine learning: applications and issues," Journal of Basic and Applied Sciences, vol. 13, pp. 459465, 2017.

[4] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," BMC Medical Informatics and Decision Making, vol. 19, pp. 116, 2019.

[5] R. kumari Dash, "Selection of The Best Classifier from Different Datasets Using WEKA," International Journal of Engineering Research & Technology (IJERT).

[6] R. Bal and S. Sharma, "Review on Meta Classification Algorithms using WEKA," International Journal of Computer Trends and Technology (IJCTT)–Volume, vol. 35, 2016.

[7] J. O. Awoyemi, A. O. Adetunmbi, and S. A. Oluwadare, "Credit card fraud detection using machine learning techniques: A comparative analysis," in 2017 International Conference on Computing Networking and Informatics (ICCNI), 2017, pp. 19.

[8] S. Mittal and S. Tyagi, "Performance evaluation of machine learning algorithms for credit card fraud detection," in 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2019, pp. 320324.

[9] H. Naik and P. Kanikar, "Credit card fraud detection based on machine learning algorithms," Int J Comput Appl, vol. 182, pp. 812, 2019.

[10] D. Gimpy and M. Rajan Vohra, "Estimation of missing values using decision tree approach," Int J Comput Sci Inf Technol, vol. 5, pp. 52165220, 2014.

[11] R. Bouckaert, E. Frank, M. Hall, R. Kirkby, P. Reutemann, A. Seewald, et al., "WEKA Manual for Version 378, 2013," Available at Accessed July, vol. 21, 2013.

[12] F. Carcillo, Y.A. Le Borgne, O. Caelen, Y. Kessaci, F. Oblé, and G. Bontempi, "Combining unsupervised and supervised learning in credit card fraud detection," Information Sciences, 2019.
[13] E. A. Baruah, S. Baruah, and J. Goswami, "A Comparative Analysis of Different Classification Algorithms based on Students' Academic Performance Using WEKA."
[14] D. H. Abd, A. T. Sadiq, and A. R. Abbas, "Political Articles Categorization Based on Different Naïve Bayes Models," in International Conference on Applied Computing to Support Industry: Innovation and Technology, 2019, pp. 286301.
[15] H. H. Patel and P. Prajapati, "Study and Analysis of Decision TreeBased Classification Algorithms," 2018.
[16] I. S. AlMejibli and D. H. Abd, "Mushroom Diagnosis Assistance System Based on Machine Learning by Using Mobile Devices," Journal of AlQadisiyah for computer science and mathematics, vol. 9, pp. Page 103113, 2017.
[17] S. R. Kalmegh, "Comparative Analysis of the WEKA Classifiers Rules Conjunctive rule & Decision table on Indian News Dataset by Using Different Test Mode," International Journal of Engineering Science Invention (IJESI), vol. 7, pp. 23196734, 2018.
[18] D. H. Abd and I. S. AlMejibli, "Monitoring System for Sickle Cell Disease Patients by Using Supervised Machine Learning," in 2017 Second AlSadiq International Conference on Multidisciplinary in IT and Communication Science and Applications (AICMITCSA), 2017, pp. 119124.
[19] T. S. Devi and K. M. Sundaram, "A comparative analysis of meta and tree classification algorithms using WEKA," Int. Res. J. Eng. Technol. (IRJET), vol. 3, pp. 7783, 2016.
[20] S. Agarwal and C. R. Chowdary, "AStacking and ABagging: Adaptive versions of ensemble learning algorithms for spoof fingerprint detection," Expert Systems with Applications, vol. 146, p. 113160, 2020.
[21] D. Abd, J. K. Alwan, M. Ibrahim, and M. B. Naeem, "The utilization of machine learning approaches for medical data classification and personal care system management for sickle cell disease," in 2017 Annual Conference on New Trends in Information & Communications Technology Applications (INTACT), 2017, pp. 213218.
[22] M. Dua, "Attribute Selection and Ensemble Classifier based Novel Approach to Intrusion Detection System," Procedia Computer Science, vol. 167, pp. 21912199, 2020.
[23] D. H. Abd, A. T. Sadiq, and A. R. Abbas, "Classifying Political Arabic Articles Using Support Vector Machine with Different Feature Extraction," in International Conference on Applied Computing to Support Industry: Innovation and Technology, 2019, pp. 7994.
[24] I. S. AlMejibli, D. H. Abd, J. K. Alwan, and A. J. Rabash, "Performance evaluation of kernels in support vector machine," in 2018 1st Annual International Conference on Information and Sciences (AiCIS), 2018, pp. 96101.