Int. J. Nonlinear Anal. Appl. 12 (2021) No. 2, 1255–1265
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/IJNAA.2021.5228

Improvement in credit card fraud detection using ensemble classification technique and user data

Evan Madhi Hamzh Al Rubaie[a,*]

[a] College of engineering, University of Babylon, Babylon, Iraq.

(Communicated by Madjid Eshaghi Gordji)

Abstract

Financial fraud is a serious problem in banking system. Credit card fraud is growing with increasing Internet usage, as it becomes very simple to collect user data and do fraud transaction. Fortunately, all records including fraud and legit transactions are present in the financial record. Improved data mining techniques are now capable to find solutions for such outlier detections. Financial data is freely available in many sources, but this data has some challenges like,1) the profile of legit and fraudulent behavior changes constantly, 2) there is a class imbalance problem in dataset, because less than 3% transaction are fraud, 3) transaction verification latency is also one more problem. All this data issues are handled using pre-processing techniques like cleaning and reduction. Main aim of this research is to find out, output attribute 'is Fraud', with better time complexity. To this end, K-means, Random Forest and J48 algorithm is used, and its accuracy rates are compared to find best fit pre-processing and machine learning algorithm. It is observed that accuracy rate of Random Forest is 93.8% when both global and local dataset is used.

Keywords: Credit card transaction, Global dataset, User dataset, J48, K-means clustering, Random forest, Ensemble method

1. Introduction

Credit card company's attractive offer is leading people to use it over a debit card. There is huge online business were credit card is the only mode of transaction, Credit card is proven to be a comfortable and convenience method of payment, as it usually supports international transaction, limit change facility, cash withdrawal and spending facility even you do not have enough funds in account. These are some advantages of credit card over other transaction method. Because of this easiness, it is observed that percentage of usage of credit card is increasing tremendously. Hackers and thief are using different methods for stealing credit card information this includes, phishing

*Corresponding Author: Evan Madhi Hamzh Al Rubaie
Email address: eng.evan.rubae@uobabylon.edu.iq (Evan Madhi Hamzh Al Rubaie[a,*])

Received: March 2021 *Accepted:* June 2021

attack, malware, information leak over phone calls, social networking [9]. Many of these transactions are carried out over Internet and it can be very easily hacked by hackers. from past few years, credit card fraud is increased and it is being problem for end user and service providers. From survey it is observed that total banking fraud is increased in last 3 years, and specifically the count of cash fraud has reduced to 5% from 9.2 % since 2015 to 2018, while credit card fraud is still at its pick level which is 5.2% [5]. Online agency and banking system have somehow achieved considerable reduction in cash fraud, but credit card fraud is still a challenge to banking system. One way to handle this fraud is using data mining tool to find fraud transaction and respond to is as fast as possible so that transaction can be avoided.

In this research data is collected from a certain bank, this dataset contains over 1 billion credit card transactions, and different data mining approaches are applied to get satisfied result to detect online fraud. The purpose of this research is to detect online fraud with high accuracy rate and with good time complexity function and algorithms. This paper is divided into 7 parts, which includes introduction, literature review, challenge in detecting online fraud, data mining steps including data pre-processing and model creation, result analysis is also included which is very prominent for deciding which data pre-processing technique and data mining algorithm should be implemented and finally conclusion and future works parts cover the synopsis which is observed after result analysis and future work contains the other techniques which can be implemented to improve the accuracy rate and time complexity of this methodology.

2. Literature survey

Srivastava et al. [13] describes Fraud detection using Markov Model. In this research they used pre-processing techniques and Hidden Markov Model to detect fraud transaction. Accuracy of this algorithm is improved after applying HMM model to transaction dataset. In [11], Sahin and Duman proposed an approach which takes transaction history as an input data and applies different machine learning and decision tree algorithms. C5.0 decision tree is used and along with it for classification support vector machine with different functions such as polynomial, Kernel functions are used and result analysis of this algorithm is compared to detect fraud. Wang et al [14] Proposed a credit card fraud detection based on whale algorithm optimized BP neural network for fraud detection. It uses whale algorithms to get balanced weight for BP neural network algorithm. MATLAB simulation is used to detect accuracy rate of this method. It shows that WOA-BP algorithm has fast convergence rate and higher accuracy which helps to detect credit card frauds.

3. Challenges in credit card detection

Credit card fraud detection was never being a simple task. There are many researchers still going on to detect fraud and avoid the losses of customer. In this task of data analysis, many researches are facing challenges with respect to data and data format all these challenges are listed below and using these challenges, proper pre-processing and machine learning algorithm is applied on collected dataset.

a. Skewed distribution of Dataset

First step of proper data mining algorithm is data pre-processing, to detect proper pre-processing techniques, it is very much essential to visualize and analyze dataset before applying any data mining technique. Data visualization step of data mining algorithm proves that data is not balanced. Legit

transaction is around 97%, where fraud transactions are covers less than 3% of dataset, which gives the problem of unbalanced data. This problem of data distribution is known as skewed distribution of datasets. This is the first challenge in handing credit card transaction history dataset [2, 8] .

b. Overlapping of data

For every user, credit card transactions are generated at huge amount, these transactions are very similar to daily or routine transaction, there can be a lot of similarities between the dataset with output class as fraud and dataset with output class as legit. This is a big challenge as it can be concluded from data visualization that difference between real and fake transaction is very less in transaction history. This overlapping of fake data over real data is known as data overlapping issue [9].

c. Sudden Variation in Transaction Type

There can be huge difference between two consecutive transactions and hence a false positive rate can increase with increase in accuracy maintaining FP rate at lower level is one more challenge for transaction fraud detection [16].

d. User Data vs. Global Data

There can be a case where a user frequently does transaction with higher amount, and which is very much identical to transaction from fraud dataset, which means a user's whose transactions are in reality is legit but it appears as fraud over a time. Data mining is based on rule extraction and model creation from previous work it is observed that such models cannot detect such users who does real transaction which appears like fake transaction. For example, a person who visit multiple country and transact amount which fluctuate in different trips.

e. Transaction verification latency

A time complexity requires for algorithm to create model and give the output to end user. For quick detection of fraud, transaction verification latency should be as low as possible. If transaction detected as frauds on spot, then government authority can help to prevent that transaction service. Transaction verification latency is one more important measure which should be considered while developing machine learning model [12].

f. False positive rate

In literature review, we studied different algorithms and techniques to detect fraud in credit card transaction. After analysis, it is observed that, the factor which is affecting this technology is false positive rate. Many times, false positive alarms appeared in case of fraud detection. This is one more challenge to reduce or at least maintain false positive rate.

4. Data collection and preprocessing

Data Pre-processing contains data collection and data pre-processing. Authenticated and well classified data is the key for data mining. Data should be in readable format, so that knowledge discovery and model formation become more convenient when machine learning is applied. Machine learning algorithm does the task of finding hidden pattern, but machine learning model becomes more reliable when data pre-processing is carried out considering algorithm for analysis. To handle User Data vs. Global Data challenges described in previous section, two different datasets are used. These datasets are user data and global data. Global Dataset is constant throughout all research while user transaction dataset varies with user. Global Dataset is collected from Bank, while user data is collected from user or from Bank. Global dataset contains transactions for all users over a time, while User dataset contains transaction for specific user. Figure 1 shows a sample for a global dataset.

111	10102	basic	23365	cash_in	25530	48895	2155962	2132597	10107	legit
112	10101	basic	1324	cash_in	11032	12356	4251060	4249736	10107	legit
113	10106	basic	15685	cash_in	18969	34654	920884	905199	10101	legit
114	10105	basic	23643	cash_in	25471	49114	3015982	2992339	10101	legit
115	10100	basic	8006	cash_in	12410	20416	9585804	9577798	10103	legit
116	10100	basic	15964	cash_in	18912	34876	1780903	1764939	10103	legit
117	10108	moderate	23922	cash_in	25413	49335	3876001	3852079	10103	legit
118	10103	moderate	8284	cash_in	12351	20635	545825	537541	10105	legit
119	10103	moderate	16242	cash_in	18853	35095	2650921	2634679	10105	legit
120	10102	moderate	24200	cash_in	25355	49555	4746020	4721820	10105	legit
121	10106	basic	16521	cash_in	18795	35316	3510941	3494420	10107	legit
122	10107	moderate	18430	cash_in	21760	40190	7865153	7846723	10103	legit
123	10102	moderate	10750	cash_in	15200	25950	6630075	6619325	10106	legit
124	10105	basic	1161	cash_in	5675	6836	1040784	1039623	10103	legit
125	10101	basic	26666	cash_in	28203	54869	930271	903605	10105	legit
126	10104	moderate	17077	cash_in	18678	35755	5240978	5223901	10103	legit
127	10105	basic	18986	cash_in	21643	40629	9595190	9576204	10108	legit
128	10107	moderate	9398	cash_in	12119	21517	4005900	3996502	10105	legit
129	10108	basic	11307	cash_in	15084	26391	8360112	8348805	10101	legit
130	10102	moderate	1718	cash_in	5559	7277	2770821	2769103	10107	legit
131	10101	moderate	9676	cash_in	12060	21736	4865920	4856244	10107	legit
	10106	moderate	24037	cash_in	28997	53034	1535743	1511706	10101	fraud

Figure 1: A sample for Global dataset

Data from Dataset: Data collected form user is then divided into 2 classes depending on user specified. Global Dataset and User Dataset contains same output class as fraud and legit. Global Dataset has classes such as user ID, type, amount, old Balance, new Balance, new Destination, destID, destOldBal, destNewBal, location, steps. User Data set contains features like user ID, type, amount and location.

4.1. Data pre-processing

Some pre-processing is required before applying machine learning algorithm on data. in this step, raw data is converted into a dataset for knowledge discovery. some data pre-processing steps are explained as follows:

- Cleaning: Credit card transaction data sometimes contains transactions which were failed for some obvious reason. This transaction practically doesn't make any sense. All such transaction should not be considered in model formation task. This data is removed before applying the algorithm. Removal of such noisy data comes under cleaning step of data mining [9].

- Integration: In this step, data is collected from different tables. This data is then integrated into one csv file. This process is known as Data Integration.
- Reduction: In data mining procedure few repeated data are present in the dataset; this data needs to be removed. Removal of such repeating data is called as Data Reduction; this involves dimensionality reduction, aggregation and clustering. Credit card transaction contains many similar types of transaction. If the count of such transaction exceeds the threshold data repetition value, then using reduction method, such transaction is removed. This is important step for handling Overlapping of data issue present in transaction dataset [8].
- Sampling: Sampling is a process in which data of few same samples is converted under one labeled to minimize the variation present in the datasets. This research requires sampling the input 'steps' into basic, moderate, high values. If user has tried to transact in first attempt, then data is sampled to basic, for second to fourth attempt then it is sampled as moderate, for more than 5 attempts value is flagged to high.

4.2. Model design

Model presented in this research consist of two phases. These two phases start their work at same time using threads, to reduce the time required to build and test model. First phase of this model consists of model building using Global data as mentioned in data collection part of this research, while second phase is based on User Data. In both of this phase different machine learning models are created on Global and Local Data. Phase 1) Random Forest on Global Dataset. In the First phase of this research, Global dataset is used for building machine leaning model. As explained in Challenge section of this paper, there is high level of skewed distribution of dataset exists in Global Data. To handle this issue, it is very much essential to use cross-validation or ensemble methods, where n number of models are created and voting method are used to find the best model. This is a technique to avoid problem of unbalanced data. Random Forest is a collections of decision trees which is built on the same principle and hence it is used in this research.

4.3. Random Forest

Random Forest is based on ensemble classification method, where n number of decision trees are created for building machine leaning models. In basic classification algorithms only one decision tree is prepared, and whole algorithm is based on that model, but in case of Random Forest, n is a user defined count, which describes the number of decision trees needs to be built to create algorithm model. The main objective of Random Forest is to handle issue of data over fitting and Data imbalance. These issues are solved in Random Forest because it creates n number of tree and then voting methods, avoid selecting one specific dataset. It uses different set of data from given dataset to create forest of decision tree, hence it is known as Random Forest [7]. Figure 2 shows a simple representation for Random Forest. To build Random Forest, initially entropy rate of each class from given dataset is calculated. Entropy rate is defined as H(S) as in equation 1.

$$IG(A,S) = H(S) - \sum_{t \in T} A(q)H(t) \qquad (1)$$

where t stands for subset of $S.A(q)$ is count of elements from in t with respect to s. This step Is repeated for n times to create n number of decision trees and then voting method is used to get final tree. If data is $m = 1, \ldots, m$: then prediction for $z^{'}$ unseen samples is calculated by equation 2 :

$$\hat{f} = (\frac{1}{M}) \sum_{m=1}^{m} fp(z^{'}) \qquad (2)$$

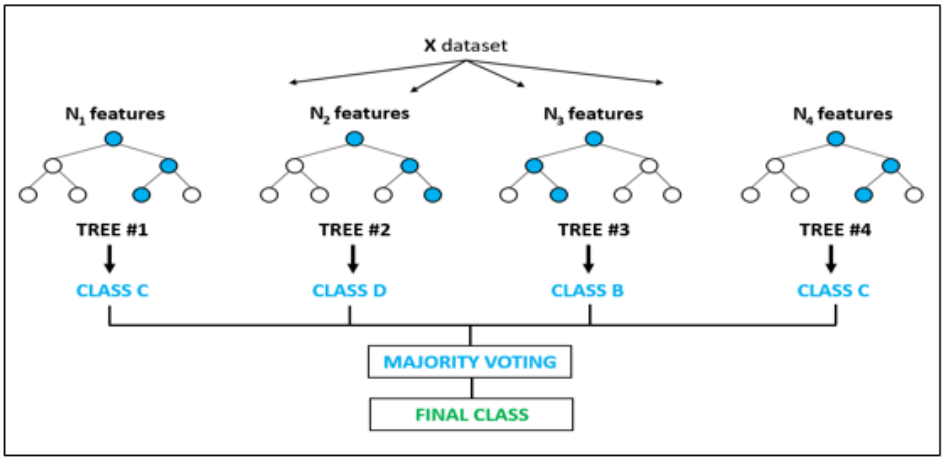

Figure 2: Scheme of a Random Forest

Random Forest can handle numeric, discrete, missing values, and continuous values. Tree pruning helps to avoid unnecessary length of tree and formation of many different trees deals with data imbalance and data over fitting issue [7].

4.4. K-means clustering

Clustering is a technique in which similar objects are grouped to form clusters. These groups are called as clusters as shown in figure 3. Many data mining projects use clustering method for data prediction. Clustering is used in computer graphics, pattern recognition, data compression and image analysis. There are different clustering techniques present which is computed by different formula for finding distance to crate clusters. These are density-based clustering, centroid-based and distribution-based clustering [1, 15].

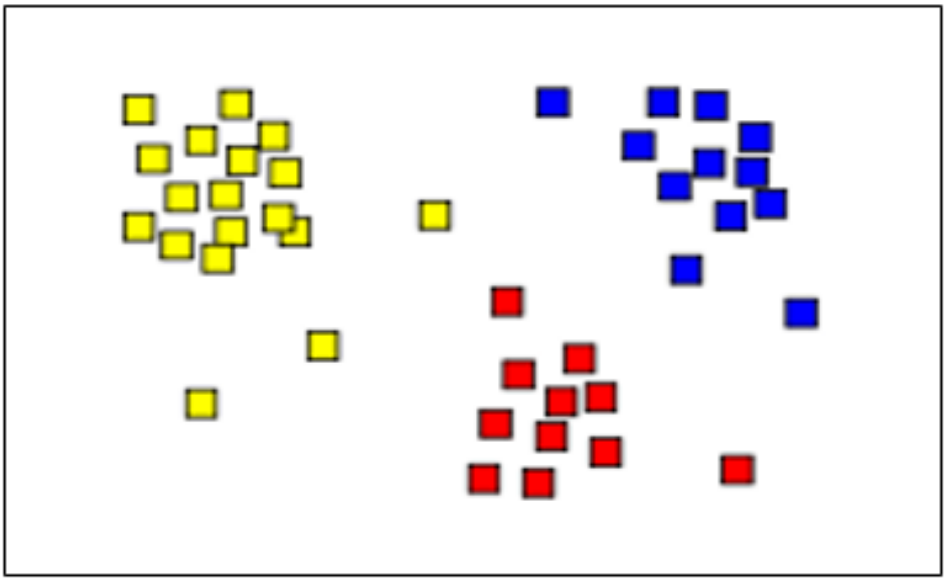

Figure 3: An example of three clusters

In figure 3, objects which belongs to similar types are clustered together to form one clusters. In this way data is divided into different clusters. When new input is given to the system, it checks the

distance from all cluster, the cluster which has more similarity with these test data is then selected as output class. In this research K-means clustering algorithm is used for building and testing Users credit card data. There are many distance measures, which are used to find clusters. In this research Euclidean distance measure is used. Formula for Euclidean distance measure as in equation 3.

$$d(p,q) = d(q,p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \quad (3)$$

where p and q are points with $p_1, p_2, \ldots, p_n$ and $q_1, q_2, \ldots, q_n$.

K-means Algorithm consists of the following steps:

1) n is taken as input from user, which is used to create n number of clusters.
2) In the dataset n number of points is randomly selected as Centroid.
3) Then from in points and centroid Euclidean distance is measured to place it in nearest cluster.
4) The centroid for each cluster is re-positioned to get correct centroid.
5) These steps are followed in iterative manner until centroid gets static. These steps insure that all clusters which are formed are balanced.

4.5. Model based on user and Global ataset

As explained in data collection section, two different datasets are used to create two models. These datasets are User Dataset and Global Datasets. First phase of this model consists of model building using Global data. While the second phase of this model consist of model building using User data.

a. Phase I – Random Forest

After pre-processing, data is forwarded to Random Forest to create a decision tree for decision making. This data is highly imbalance and hence ensemble method is required to use this in scenario. Random Forest algorithm is already based on ensemble method, and the accuracy of Random Forest is better, also when consider timing, other algorithm needs to implement bootstrap aggregation, boosting or bagging process externally while in case of random forest, these steps are internally present so there is no need of special algorithm for bagging or boosting as shown in figure 4.

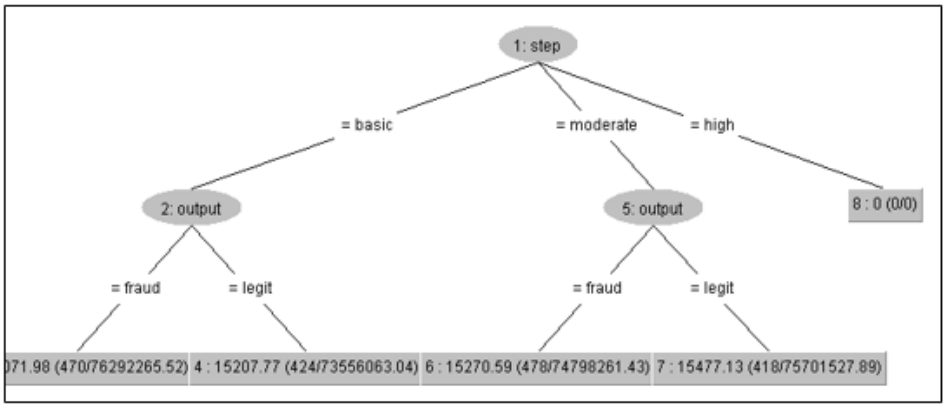

Figure 4: one of the formed in Random forest

As shown in Figure 4, Random Forest formed by global dataset. Output class of this decided on tree consist of only two parameters which is either fraud or legit, and hence there is high possibly that tree length is unnecessarily high, random forest uses tree pruning method to remove unnecessary tree branches.

b. Phase – II

In Phase II of this research, User Dataset is used to get decision from dataset, output of this phase also contains only fraud and legit parameters. After processing of data, it is forwarded to K- means algorithm. In k-means clustering n is used, which is used to create k number of decision trees. User defined number of clusters is an advantage of K-means clustering over other clustering algorithm. In this research 2 types of clusters are formed which is responsible either legit or fraud transaction as shown in figure 5.

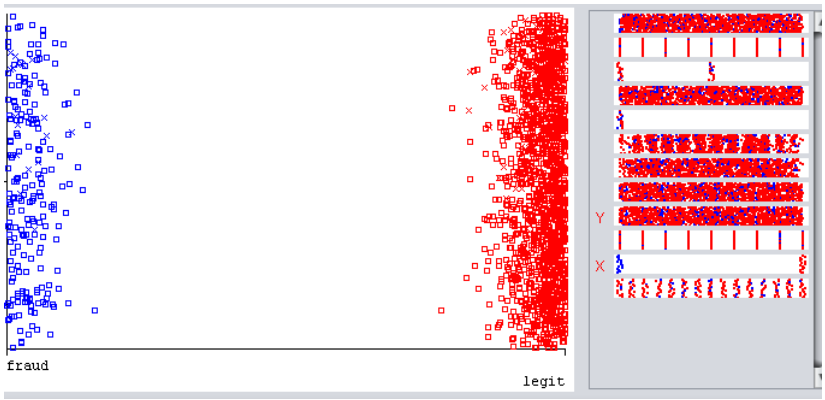

Figure 5: one of the formed in Random forest

As illustrated in Figure 5, k-means clustering is applied on User dataset. Output class of this decision tree consist of only two parameters which is either fraud or legit. Important factor in phase II which is to considerer is its accuracy rate, it is very important to get highly well classified data, with probability distribution greater than 0.8 for output class.

c. Model formation

As explained in the previous sections, prepossessing, phase–I and Phase-II are used in this model. Using Global and user data is a key for reducing false positive rate. In this system Data is collected from data source and then it is forwarded for pre-processing methods, then it is divided into two sections for classification in phase I and Phase II. In Figure 6, it is observed that, data is forwarded to machine learning algorithm for model formation; data is simultaneously creating models for these two phases, as it is not depended on each other. Once model is formed for Phase I, input data is tested on Phase I, if the input dataset is found to be legit, then there is no need to check for Phase II, as the transaction is legit, but in case of fraud classification, phase II output is considered before final output. After Phase II, model formation next step is to calculate its accuracy using cross validation technique, if output class is fraud or legit and the probability distribution of output class is greater than 0.8 in that condition only Phase II output is considered for Analysis, in other Condition where output probability is less than 0.8, phase II return value as 'un-classified'. There is a situation where user has highly undistributed data, and in few cases user dataset can have very few transactions, in that case obvious probability distribution comes out to be less than 0.8, and hence such output is considered to be invalid and system returns its value as 'un-classified. In such situation only, phase I output is considered as real output of this system. While if both output result as fraud then transaction is considered to be a fraud transaction, while if the phase II return output as legit, then it means even if this transaction is appearing to be a fraud transaction globally, but this specific user continually does such transaction, and therefore there is no meaning in classifying this transaction as Fraud transaction. This Technique is used to avoid problem of False Positive rate and Overlapping of data.

5. Experiments

After training machine learning model, its accuracy analysis is essential before deploying it to real world problem. There is various method of back testing the accuracy of algorithm, which include out of the bag error, cross validation, percentage split in which test and training data is split into specific ratio, to find the percentage instances which are correctly getting classified. In this experimental analysis cross validation technique and out of bag error estimation is used for finding accuracy of Phase I. Phase II, is depend on user data, and hence its model is always dynamic, which means there is need to compute separate accuracy of phase II. Whole system accuracy is computed by applying this system to real time dataset with missing output value.

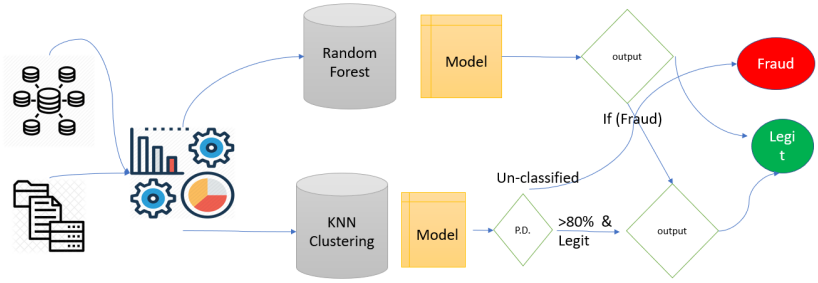

Figure 6: the proposed credit card fraud detection system

A. Phase I Analysis:

For Global dataset and Random Forest model, cross validation technique is used to find out correctly and incorrectly classified instances. For same dataset, this test is carried out on different classification and clustering algorithms including Random Forest, J48, SVM, and K-means. some of the best results and its statics are presented in Table 1.

Table 1: Accuracy of different data mining algorithms

Algorithm	Total Instances	Correct Instances	Incorrect result	Time Complexity	Percentage Accuracy
J48 Decision T	1000000	893000	107000	$\sim$ 3 M	89.3%
SVM	1000000	899000	101000	$\sim$3 M	89.9%
K-means Clustering	1000000	856000	144000	$\sim$2 M	85.6%
Random Forest	1000000	921000	79000	$\sim$3-4 M	92.1%

It is clear that Random Forest with dataset and pre-processing technique have higher accuracy rate in comparison with J48, SVM, K-means. It is observed that J48 decision tree also has good accuracy rate and adding ensemble methods with algorithm can improve the accuracy, but then again external bagging or boosting will be required, which will be time consuming. From the above analysis it is also observed that time complexity of the given model is not good, but then accuracy is higher. True positive and false positive rate for Random Forest were 0.901 and 0.56, respectively.

B. Phase II Analysis:

Phase II depends on user data, and hence its model is always dynamic, which means there is a need to compute separate accuracy of phase II. If system accuracy of Phase I and Phase II combined is better than Phase I alone, we can assume that User Transaction has equal importance as Global Transaction.

C. System Analysis:

Accuracy of whole system is measured by applying new real time transaction on the model. This new transaction contains fraud and legit, both types of transactions. There were around 1000 transactions were tested on this system. detailed results are explained in Table 2.

Table 2: Real time system analysis

System	Total Instances	Correct Instances	Incorrect result	Time Complexity	Percentage Accuracy
Without User Data	1000	912	88	$\sim$ 3 -4 M	91.2%
Proposed System	1000	938	42	$\sim$ 5 M	93.8%

From Table 2, it is observed that accuracy rate of the proposed system is slightly greater than same system without using User data and clustering algorithm. It is observed that real time accuracy of this system is 93.8% but the time complexity increased to further extend, as clustering and then analyzing output class probability takes more operational time. The considerable thing here in this

analysis is accuracy of this model is increased because of increase in false positive rate from 0.56 to 0.92. It is observed that some users have such transaction request which appears fraud, but this is how those users transact, this gap is filled using User Data Phase to the previous system.

6. Conclusion

In this research, it is observed that when steps are higher, output class is mostly fraud. This is a common observation of this research. It is observed that using Random Forest or ensemble method, data imbalance issue can be handled and accuracy of algorithm can be increased. It is found that Random Forest performs better on Global Transaction data when compared to other algorithm with accuracy rate of 92.1 %. Few transactions can be protected from being false positive result, by using user's specific data along with Global data. Transaction time required for this system is more when compared to other system based on Global Transactions.

References

[1] Z. Arbabi, K. Yeganegi and A. Obaid, Application of neural networks in evaluation of key factors of knowledge management system, Case Study: Iranian Companies Based in Alborz Province, International Conference for Modern Applications of Information and Communication Technology. Baghdad, Iraq, 2020.

[2] V. Babar and R. Ade. A novel approach for handling imbalanced data in medical miagnosis using undersampling technique. Communications on Applied Electronics, 5(7) (2015) 36-42.

[3] Credit card fraud detection: A hybrid approach using fuzzy clustering and neural Network, IEEE Second International Conference on Advances in Computing and Communication Engineering, Dehradun, India, 2015, pp. 494-499.

[4] Y. Gmbh and K. G. Co, Global online payment methods: Full year 2016, Tech. Rep. 3 (2016).

[5] C. Greene and J. Stavins, The 2016 and 2017 Surveys of Consumer Payment Choice: Summary Results, Research Data Reports, 2018.

[6] H. T. Kam, A data complexity analysis of comparative advantages of decision forest constructors, Pattern Analysis and Applications 5 (2002) 102-112.

[7] R. A. Kamble, Short and long-term stock trend prediction using decision tree, IEEE International Conference on Intelligent Computing and Control Systems, Madurai, India, 2017.

[8] A. Mishra and C. Ghorpade, Credit card fraud detection on the skewed data using various classification and ensemble techniques, IEEE International Students' Conference on Electrical, Electronics and Computer Science, Bhopal, India, 2018.

[9] K. Modi and R. Dayma, Review on fraud detection methods in credit card transactions, International Conference on Intelligent Computing and Control, India, 2017.

[10] A. D. Pozzolo, O. Caelen, R. A. Johnson and G. Bontempi. Calibrating probability with undersampling for unbalanced classification, IEEE Symposium Series on Computational Intelligence, Cape Town, South Africa, 2015, pp. 159-166.

[11] Y. Sahin and E. Duman. Credit Card Fraud by Decision Trees and Support Vector Machines, Proceeding of the International MultiConference of Engineers and Computer Scientists, Hong Kong, 2011, pp.16-18.

[12] N. Soltani, M. K. Akbari and M. S. Javan, A new user-based model for credit card fraud detection based on artificial immune system, The 16th CSI International Symposium on Artificial Intelligence and Signal Processing, Shiraz, Iran, 2012, pp. 29-33.

[13] A. Srivastava, A. Kundu, Sh. Sural and A. Majumdar. Credit card fraud detection using hidden markov model, IEEE Transactions on Dependable and Secure Computing, 8(1) (2008) 37-48.

[14] C. Wang, Y. Wang, Z. Ye, L. Yan, W. Cai and S. Pan, Credit card fraud detection based on whale algorithm optimized BP neural network, The 13th International Conference on Computer Science & Education (ICCSE), Colombo, Sri Lanka, 2018.

[15] K. Yeganegi, D. Moradi and A. J. Obaid, create a wealth of security CCTV cameras, International Conference for Modern Applications of Information and Communication Technology. Baghdad, Iraq, 2020.

[16] W. F. YU and N. Wang, Research on credit card fraud detection model based on distance sum, IEEE International Joint Conference on Artificial Intelligence, Hainan, China, 2009.

[17] L. Zheng, S. Wang and S. Xuan, Random forest for credit card fraud detection. IEEE 15th International Conference on Networking, Sensing and Control, Zhuhai, China, 2018.