DETECTION OF FAKE PROFILES USING MACHINE LEARNING

Diksha Gupta

Student,

Sri Sai College of Engineering & Technology, Badhani, Pathankot.

ABSTRACT

Fake profiles in modern era are cause of concern in almost every field of life. Credit card, money laundering and bank fake profiles are common and technology has to play important part in overcoming this issue. This paper provides insight into financial fake profiles leading from malicious users in trading network. To this end several techniques are researched over. To start with price based fake profile detection is discussed and then similarity matrix, linear binary patterns, support vector machine and random forest in the field of fake profile detection are elaborated. This paper highlights pros and cons of each of such techniques. Dataset required determining classification accuracy of these approaches is synthetically driven. Execution time while determining fake profiles is critical entity and similarity matrix approach is fast and accurate as compared to random forest, support vector and linear binary patterns.

Parameters: Classification Accuracy, Execution time

Implementation tool: Matlab 2018

Achievement: support vector machine results are closer to similarity matrix based approach in terms of classification accuracy but execution time of similarity based approach is much less and hence this algorithm is considered better in determining financial fake profiles.

Keywords: *Financial Fake profiles, LBP, RF, SM, classification accuracy, execution time*

i. Introduction

Fake profiles in financial applications are common and avoidance is compulsory. Fake profiles in such areas not only divert mass communication towards other investment alternatives and finance in market dries up. This is one of leading issues causing economic crises. Detecting financial fake profiles and blocking source of such fake profiles is important. Technology can help check fake profiles and cause stability in trading environment. Simplified mechanism based on misleading price on goods can be used as a feature to detect fake profiles but that will work only for stable financial environment. Unfortunately trading environment is ever fluctuating market where price as feature may not work accurately. To overcome issue of this sought, statistical features can be used in place of single feature in term of price. Model reflecting simplified price based model is in figure 1.

In this strategy, price from dataset is extracted. The labelling information regarding price tag serve as original price. To perform testing this labelling information is compared with extracted price tag. This extracted price tag and original price becomes two column feature vector. Threshold value indicating fake profileulent and normal transaction is established in the form of classes. Difference in extracted and original price tags if violates threshold value, fake profile is detected otherwise transaction is complete. Fluctuating trading environment could hamper classification accuracy of this mechanism. Handling changing environment requires multiple attribute feature extraction and selection mechanism. The next approach that improves results of single feature prediction model is linear binary pattern analysis. **Linear binary pattern** mechanism is implied on dataset to extract statistical features.(Tejashwini 2017) These statistical features include mean, median, mode, correlation, regression, entropy and kurtosis. All the extracted features are represented through feature vector. Forming feature vectors with heterogeneous values requires large buffer size. After extraction of features, testing process plays its part. Testing also extract features that are compared against the training features to classify the result. In-depth of this strategy is given in section 2. The problem with this approach is linear extraction of features and execution time in classifying the result is too high. The

primary reason for slow execution time is extraction of all the features having high or low frequency values. To overcome this problem, support vector machine can be used.

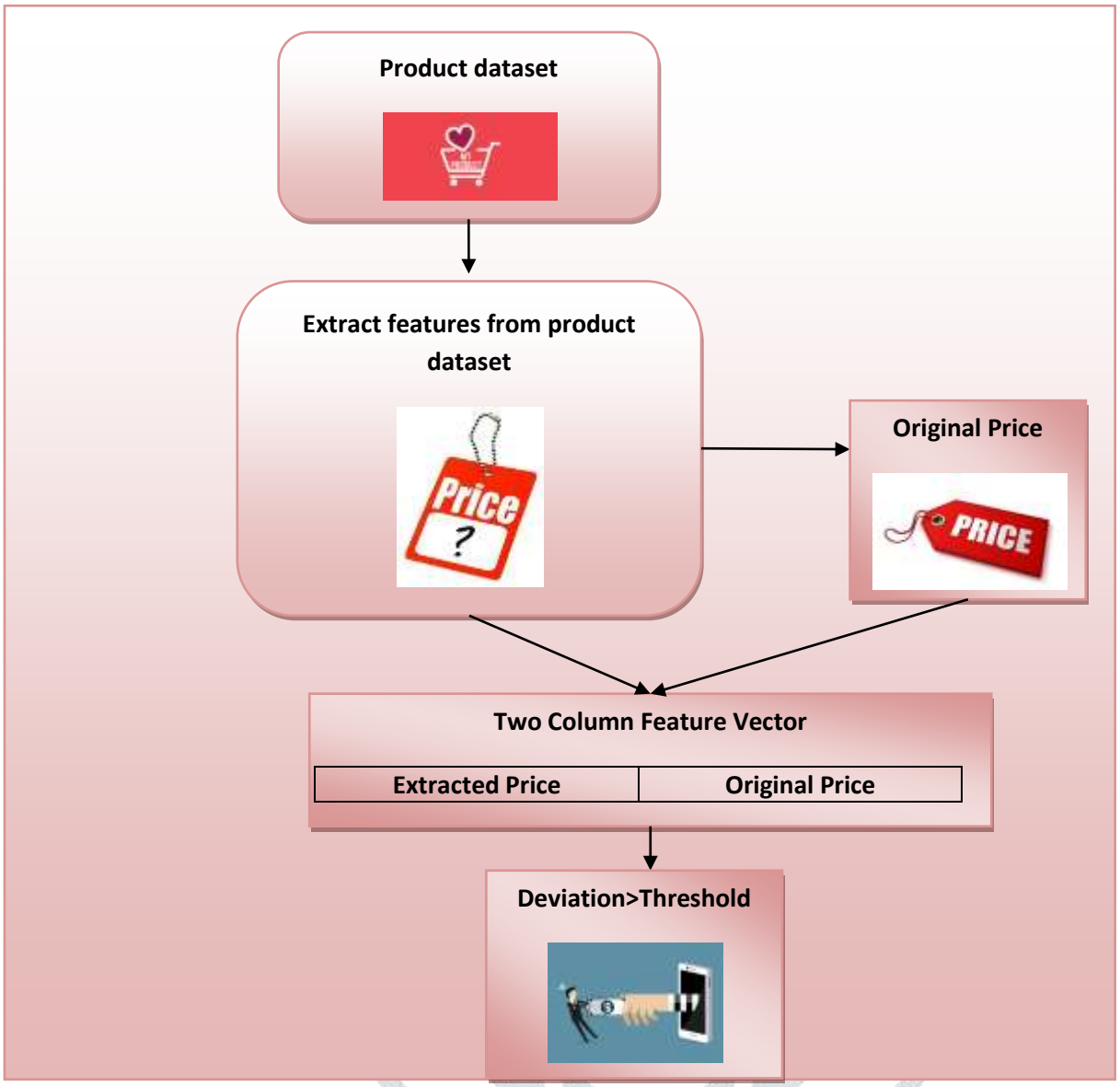

Figure 1: Price based fake profile detection mechanism

SVM is supervised machine learning algorithm that has associated learning algorithms. This algorithm classifies the result based upon non probabilistic binary mechanism. It assign obtained values either to one class or the other. Accordingly two hyper planes are formed and aggregate values from features belong to one hyper-plane or other.(Abiramy et al. 2019) This classifier can only classify presented data into two distinct categories but if more than two classes are to be evaluated than it is not possible through this classifier.

Random forest algorithm produces output based upon schedules that is randomly formed. Multitude of decision trees are formed using this mechanism. Each tree consumes training time and reducing this time is objective of researches suggested by(Freeman and Hwa 2013; Patgiri et al. 2019). The formed schedule can consume least execution time but classification accuracy may not be stable in each case due to randomization. Rest of the paper is organised as under: section 2 present in-depth study of LBP, SVM, Random forest and similarity based matrix, section 3 gives the comparative analysis from result obtained through distinct algorithms and present discussion of result obtained, section 4 gives the methodology that can further enhanced the result, section 5 gives conclusion and future scope and last section present references.

ii. BACKGROUND ANALYSIS

This section study different mechanisms that can contribute towards scanning of fake profiles within credit card along with prevention if detection consumes least execution time.First of all we will discuss linear binary pattern to detect malicious entry within dataset. Dataset formation is synthetic having structure listed in table 1.

Transaction_Id	Customer_ID	Number_of_stocks	Time_of_transaction	Location	Transaction_Success	Payment	Fake profile
1	75-0394959	Often	10:25 AM	Portugal	FALSE	Cash	FALSE
2	57-5862341	Once	2:25 AM	Ethiopia	TRUE	Cash	FALSE
3	40-9613879	Never	11:41 PM	Venezuela	TRUE	Debit_Card	TRUE
4	34-2794758	Seldom	7:53 PM	Armenia	FALSE	Credit_Card	FALSE
5	77-1386847	Once	11:03 AM	China	FALSE	Debit_Card	TRUE
6	27-9262596	Never	4:20 AM	Brazil	FALSE	Debit_Card	FALSE
7	92-9503116	Daily	8:07 PM	Norway	TRUE	Debit_Card	FALSE
8	65-4630371	Once	1:27 AM	Philippines	FALSE	Debit_Card	FALSE
9	14-8751538	Monthly	11:15 AM	Nigeria	FALSE	Credit_Card	FALSE
10	95-8257618	Never	3:18 AM	Bosnia and Herzegovina	TRUE	Credit_Card	TRUE
11	98-3079133	Yearly	1:47 AM	El Salvador	FALSE	Credit_Card	TRUE
12	46-3058861	Once	9:42 PM	Indonesia	TRUE	Cash	FALSE
13	00-7821561	Monthly	8:58 PM	Nigeria	FALSE	Credit_Card	TRUE
14	78-1715464	Yearly	11:23 AM	Brazil	TRUE	Cash	TRUE
15	50-4279225	Daily	2:45 AM	China	TRUE	Cash	FALSE
16	23-7619174	Once	1:41 PM	Ethiopia	TRUE	Cash	TRUE
17	13-5015076	Once	8:16 PM	China	FALSE	Debit_Card	TRUE
18	63-6339431	Seldom	4:44 PM	Brazil	TRUE	Credit_Card	FALSE
19	77-8604671	Never	12:31 PM	United States	TRUE	Credit_Card	TRUE
20	07-2469001	Never	12:14 AM	Philippines	FALSE	Debit_Card	TRUE
21	57-2688287	Monthly	9:45 PM	Indonesia	FALSE	Cash	TRUE
22	04-5158392	Often	10:53 PM	China	TRUE	Debit_Card	FALSE
23	08-4463031	Yearly	1:11 AM	United States	TRUE	Credit_Card	TRUE
24	98-0352583	Yearly	6:27 PM	Portugal	TRUE	Debit_Card	FALSE
25	00-7772925	Weekly	6:47 AM	Brazil	FALSE	Credit_Card	FALSE
26	11-6510661	Often	2:05 PM	France	TRUE	Credit_Card	TRUE
27	10-4398660	Never	12:35 AM	France	FALSE	Credit_Card	TRUE
28	88-8169884	Seldom	10:28 AM	Canada	FALSE	Credit_Card	FALSE
29	90-4112043	Weekly	3:44 PM	Azerbaijan	FALSE	Debit_Card	FALSE

30	39-8317772	Monthly	2:22 PM	Portugal	TRUE	Credit_Card	TRUE
31	00-6869730	Yearly	1:06 AM	Thailand	TRUE	Cash	TRUE
32	93-6088635	Yearly	11:57 AM	Poland	FALSE	Credit_Card	TRUE
33	95-5978269	Daily	1:44 PM	China	FALSE	Credit_Card	FALSE
34	74-5915247	Seldom	2:37 AM	Greece	FALSE	Cash	TRUE
35	63-3343814	Yearly	5:26 AM	Indonesia	FALSE	Credit_Card	TRUE
36	03-7278964	Yearly	4:42 PM	Brazil	FALSE	Credit_Card	TRUE
37	20-0082567	Often	6:07 AM	China	TRUE	Credit_Card	TRUE
38	22-0973044	Seldom	8:26 AM	Paraguay	FALSE	Cash	FALSE
39	13-4672642	Never	8:26 PM	Sweden	FALSE	Credit_Card	TRUE
40	64-8712919	Often	6:42 PM	China	TRUE	Credit_Card	TRUE
41	99-5132950	Monthly	8:51 AM	Lithuania	TRUE	Credit_Card	FALSE
42	97-4586597	Yearly	10:40 PM	China	FALSE	Credit_Card	FALSE
43	61-0637441	Seldom	8:13 PM	Philippines	FALSE	Debit_Card	FALSE
44	44-6019354	Weekly	2:20 PM	Indonesia	TRUE	Cash	FALSE
45	84-4972883	Weekly	7:27 AM	China	FALSE	Debit_Card	FALSE
46	89-7112086	Daily	2:27 PM	Kazakhstan	TRUE	Credit_Card	TRUE
47	24-7640206	Yearly	3:08 PM	Syria	FALSE	Debit_Card	FALSE
48	17-1401232	Daily	6:55 AM	Brazil	FALSE	Debit_Card	FALSE
49	82-0438126	Often	6:58 PM	China	TRUE	Debit_Card	TRUE
50	53-9121454	Often	5:29 PM	Azerbaijan	TRUE	Debit_Card	TRUE

Table 1: Syntahtically driven dataset for financial fake profile detection

- Linear Binary Pattern in financial fake profile detection

Using machine learning financial fake profile can be detected especially in the field of trading. Limited work is done towards detection of such fake profiles. (Hu et al. 2018) proposed squirrel cage linear binary pattern mechanism in the detection of video anomaly. This mechanism can be incorporated within detection of fake profiles within financial transactions. This mechanism can effectively extract features and form feature vector. This feature vector then can be compared against test data to determine fake profiles. In most of existing researches (Anjos et al. 2014; Cao et al. 2019) LBP mechanism is used to detect fake profiles on image dataset. This work implement linear binary pattern on text data to form feature vector and to derive conclusion on fake profileulent transactions. (Kumar et al. 2019)

The methodology of work fetch the data from dataset and perform grouping based on common distance mechanism. suppose fetched data have sequence of points (1,3) and (2,4) then it will be represented using LBP through equation 1

$$F(x) = \frac{x-2}{1-2} * 3 + \frac{x-1}{(2-1)} * 4$$

Equation 1: representation of data fetching and representing it with composite function F(x)

In general if ‘n’ values are fetched then LBP is represented with the equation 2

$$F(x) = \frac{(x-x_2)(x-x_3)---(x-x_n)}{(x_1-x_2)(x_1-x_3)---(x_1-x_n)} * y_1 + \frac{(x-x_1)(x-x_3)---(x-x_n)}{(x_2-x_1)(x_2-x_3)---(x_2-x_n)} * y_2 \pm - \mp \frac{(x-x_2)(x-x_3)---(x-x_n)}{(x_1-x_2)(x_1-x_3)---(x_1-x_n)} * y_n$$

Equation 2: General equation for LBP

In case 'x' represents time value then 'x' can yield class value as fake profileulent or not depending upon the extracted features corresponding to time. This means for any distinct real values x_i along with any different attribute values may or may not be distinct y_i, there exists a unique polynomial P(x) having deg(P)<n. In case all the 'n' values are distinct then 'n' different classes for classification can be yielded. This mechanism applied to dataset for table 1 give results as shown in figure 2.

Financial fake profiles specified from labeling dataset is give distinct values of F(x). All these different values form different classes. Test data is checked and feature vector is again formed. The feature vector of training and testing data is compared with each other to determinefake profiles within test data. Classification accuracy that is obtained by subtracting actual and approximate results is substantial (80% on an average) and execution is high. To overcome issues of LBP , support vector machine on text dataset is implied.(Tejashwini 2017)

- Support vector machine on fake profile detection

Support vector machine is fast and reliable mechanism to classify the data into different classes. SVM is based upon formation of hyperplane. Each different hyperplane represent one class. Intensity of values obtained from dataset decides which hyperplane is penetrated and according class is predicted. Kernel trick mechanism is applied to classify non linear data that is required in the prediction of fake profiles in financial data.(Neto et al.) Formation of hyperplane and classification process is describe by considering functional aspects represented with F(x). Single valued classification model is described as

$$F(x) = \sqrt{5}$$

Equation 2: Hyperplane formation equation

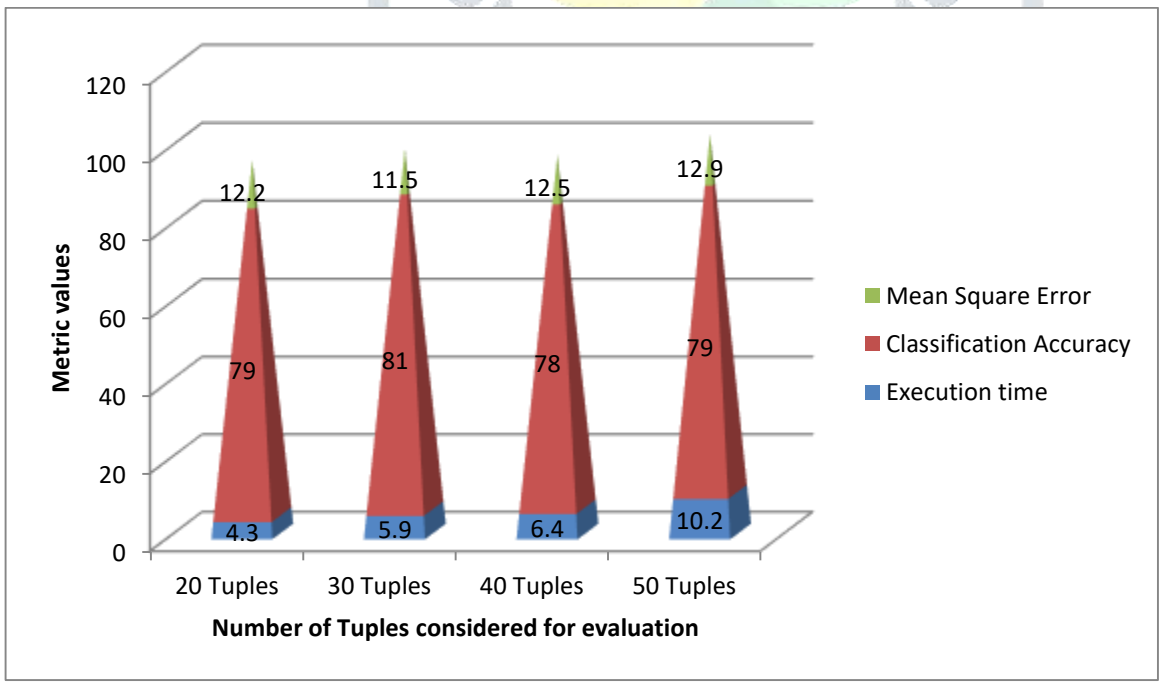

Figure 2: Result obtained from LBP on text dataset

To determine hyperplane and classification process, real valued approximation close to zero is critical. Thus $F(x) = \sqrt{5}$ can be represented as $x^2=5$ and $x^2-5=0$.

Evaluating this equation could give us one class for prediction. To predict class of test data and form hyperplane root determination equation is applied.

$$X_{n+1} = X_n - \frac{f(x_n)}{f'(x_n)}$$

Equation 2: optimization equation to determine feature vector corresponding to hyperplane

F'(x) represents derivative corresponding to equation 2. The hyperplane values obtained through this mechanism is optimal as it is a iterative process and each iteration yield unique values. Once repeated values obtained from each iteration then hyperplane is labeled with that value to predict fake profile or normal class. Execution time from this process is high but classification accuracy is improved. Result section from this mechanism is given in figure 3.

- Random Forest approach for fake profile detection

Random forest algorithm used to detect fake profiles in financial industry is proposed by (Liu et al. 2015). Ratio of debt to equity is used to detect fake profiles within trading environment. Ten fold cross validation approach is followed to detect fake profiles with accuracy. Hold out ratio of 0.3 is used for training and 0.7 is used for testing. Random forest approach also uses management expense ratio for prediction of fake profiles. Working of this model is highlighted considering large number of decision trees. Each tree act as a prediction node.(Nipane et al. 2016)Node with highest vote gives the prediction. Correlation between individual tree is low. Uncorrelated trees can produce more accurate result as compared to individual tree. The reason is that individual tree protect other trees from their individual errors. Multiple distinct decision tree are formed and then these are combined to generate more accurate tree.

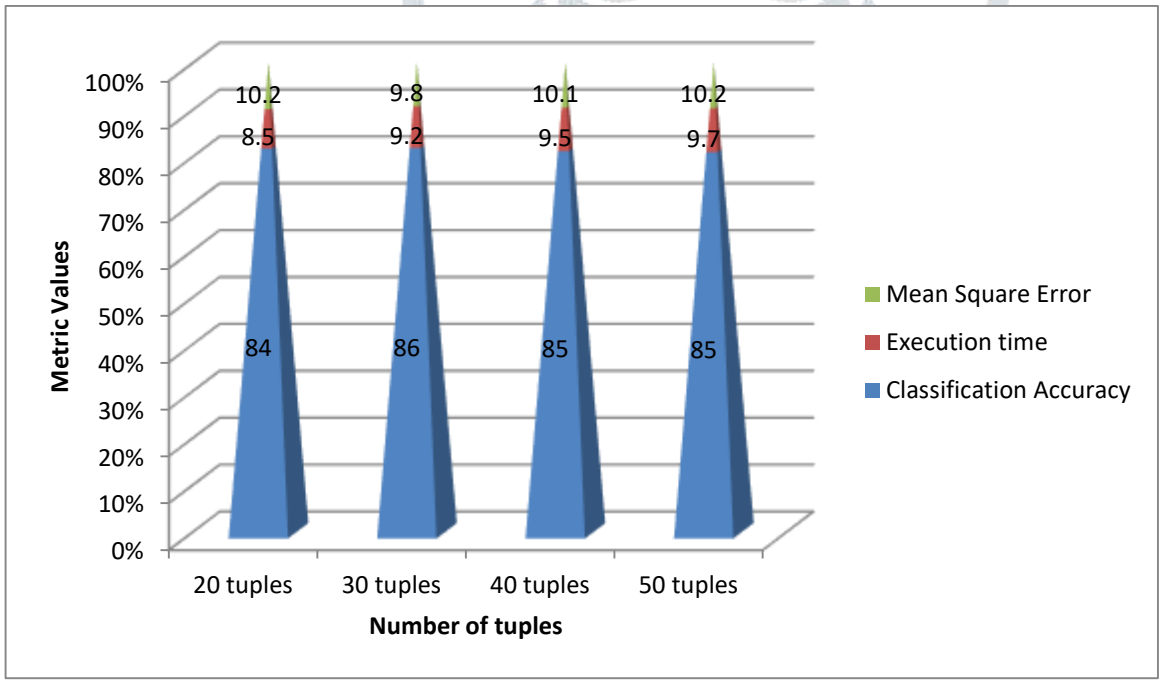

Figure 3: Classification accuracy, execution time and mean square error comparison

Formation of random forest build trees by extracting different features and each class is labeled with composite feature vector. This composite feature vector is compared against test feature vector to derive the conclusion regarding fake profiles. The process of feature vector formation is given in figure 4

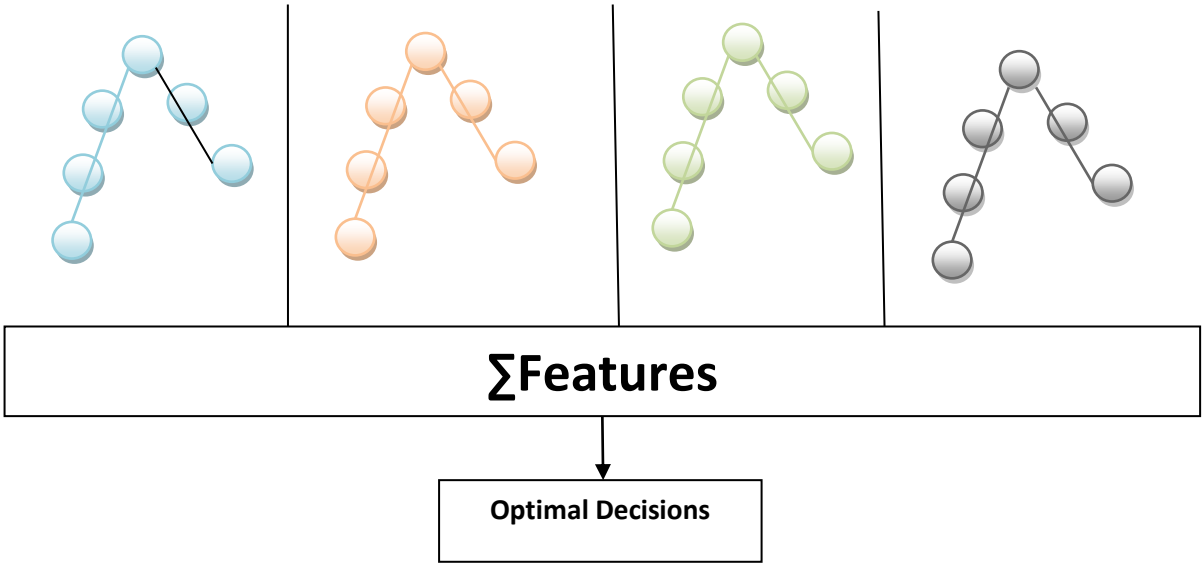

Figure 4: Random forest based approach for decision making

Each colored circle in figure 4 indicates different feature tree.(et al. 2012) The optimal decision is selected by combining multiple features together to generate unique optimal and fittest value. This approach when applied to the dataset given in table 1 result generated is given in figure 5 in terms of execution time and classification accuracy. Metrics considered for evaluation of fake profiles includes classification accuracy indicating accuracy and reliability of result along with execution time. Both of these metrics are achieved high fitness but execution time can be further reduced.

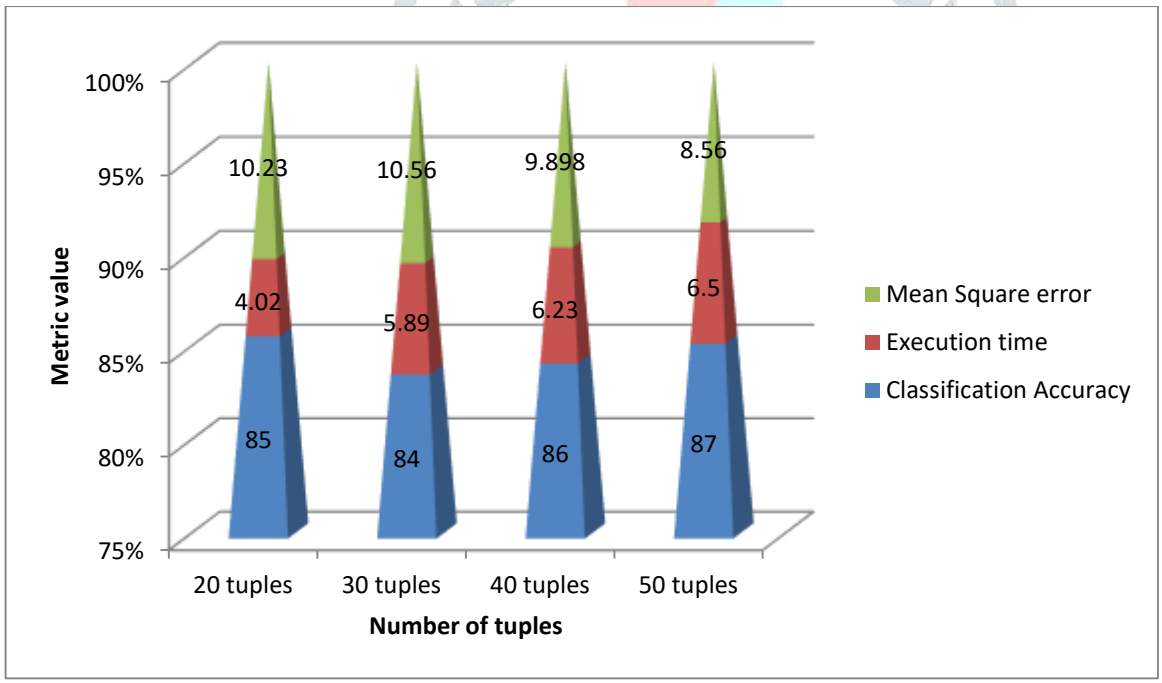

Figure 5: Random forest approach for predicting fake profiles

Random forest mechanism generates optimal results at certain interval of time but may not yield best result at other times. To overcome the problems of random forest similarity based mechanism is employed to first decrease size of dataset based on similarity and then applies decision tree approach for optimal results.

- Similarity based approach for fake profile detection

Similarity based approach reduce dimensionally of data retrieved and hence execution time can be significantly reduced. Similarity based is proposed by (Huang et al. 2018). Table 1 showing dataset is first fed into CoDetect model and then low frequency terms are eliminated. The items from dataset is denoted with $x_1, x_2, - - --, x_n$ and frequency

is denoted with $f_1, f_2, - - - f_n$then threshold value is compared against frequencies. Frequencies less then threshold values are rejected and other values are retained.

Items	Frequency
X1	F1
X2	F2
---	--
Xn	Fn

Table 2: items with frequency count from dataset

If F>Threshold then corresponding 'X' are retained. After retaining values decision trees are formulated again and then features with maximum count determines fake profileulent transactions. This is represented in figure 4. Result obtained from this approach in terms of classification accuracy and execution time in figure 6.

The result obtained from all the approaches are compared in the next section. Comparison of result indicates similarity based approach is better as compared to LBP, SVM and plane random forest approach.

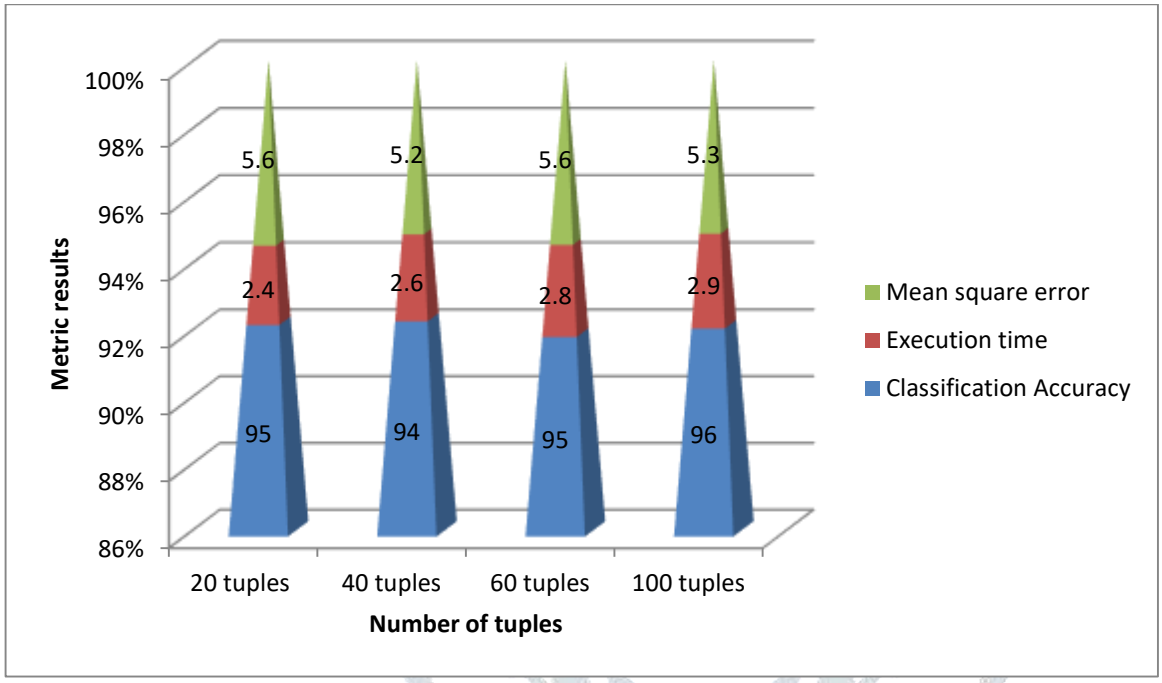

Figure 6: results from similarity based approach

iii. Result Comparison from LBP, SVM, Random forest and Similarity based approach

The result obtained from different approaches in fake profile detection is presented in this section. Result improvement by 5% is observed and technique similarity based approach is obtained to be optimal. Comparison of result is given in figure 7.

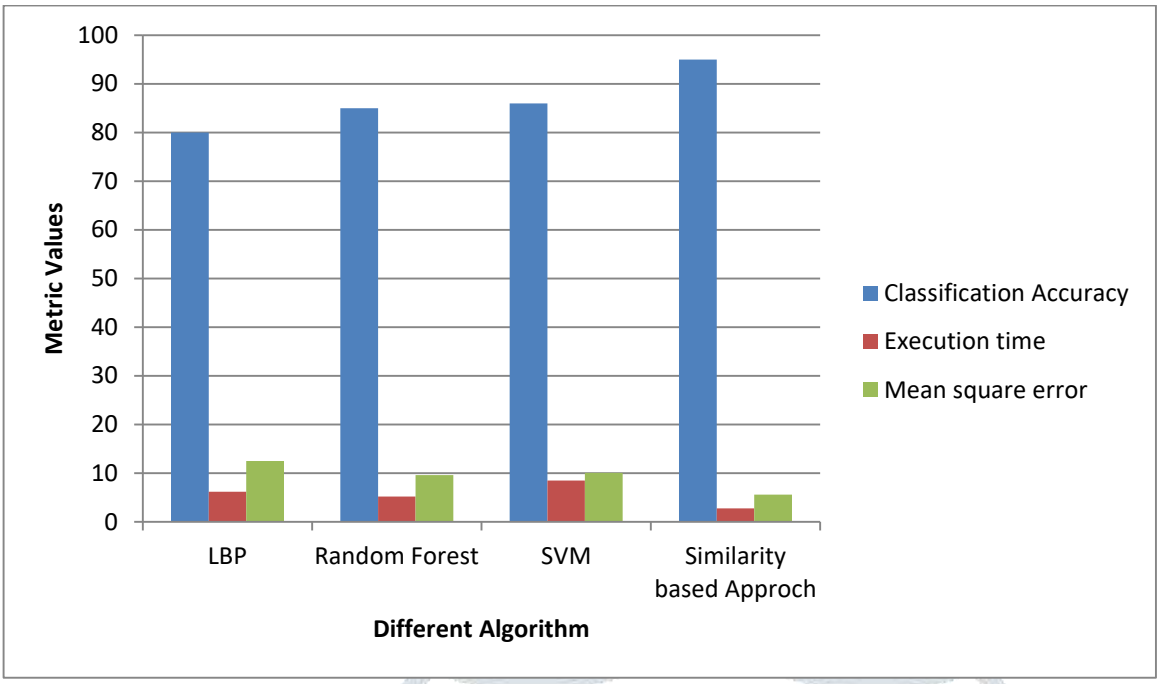

Figure 7: Comparison of result from LBP, RF, DVM and similarity based approach

Execution time is significantly reduced since significant values are retained and insignificant value is eliminated. Classification accuracy is improved due to optimal decision tree mechanism. this approach can be further improved by changing hold out ratio.

iv. Methodology to be followed for result improvement

The methodology to be followed must accommodate pre-processing mechanism. This pre-processing mechanism must eliminate noisy data. This noisy data may include missing data. In addition to missing data insignificant values must be eliminated in order to increase the execution speed. After that feature vector must be formed using similarity based random forest approach.(Abdelhamid et al. 2014)The proposed model that can further improve the result of similarity based approach is given in figure 8.

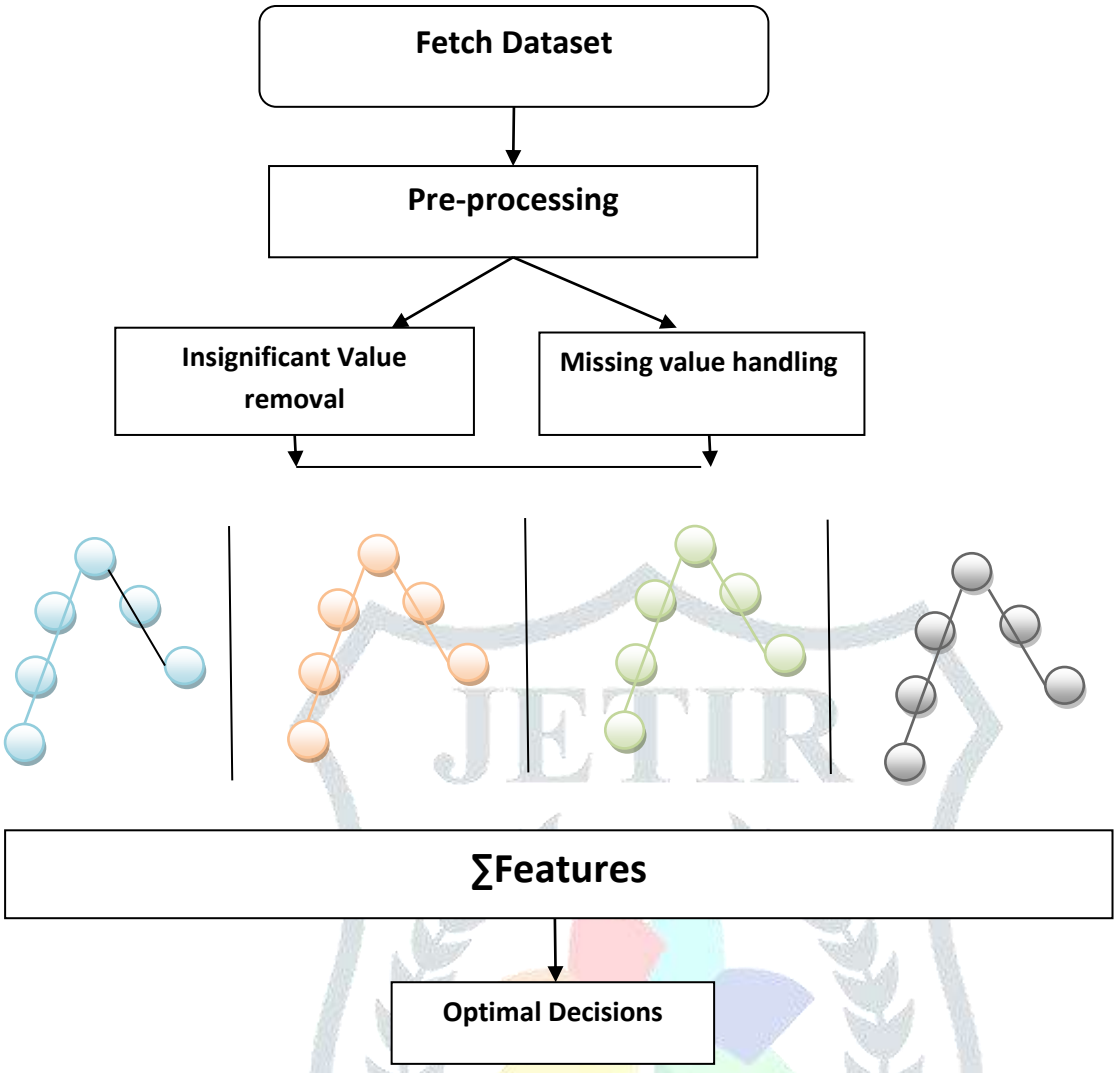

Figure 8: Proposed system to improve classification accuracy

v. Conclusion and future scope

The mechanism based on similarity based approach uses decision tree approach along with dimensionality reduction to improve performance of fake profile detection mechanism. overall result improvement by 5% is observed. Execution time is still a problem that has to be improved further. In order to perform this operation missing value handling along with infrequent value removal can be used to reduce the size of dataset. In addition random forest approach is applied to obtain classification result. Classification accuracy is already upto 96% but execution time can be a problem that can be improved by the use of suggested approach.

vi. References

1. Abdelhamid D, Khaoula S, Atika O (2014) Automatic Bank Fraud Detection Using Support Vector Machines. Int Conf … 10–17
2. Abiramy R, Narayanan K, Anandan R, Swaraj Paul C (2019) Fraud detection for online retail using random forest. Int J Eng Adv Technol 8:1–6
3. Anjos A, Chakka MM, Marcel S (2014) Motion-based counter-measures to photo attacks in face recognition. 147–158 . doi: 10.1049/iet-bmt.2012.0071
4. Cao J, Wang M, Li Y, Zhang Q (2019) Improved support vector machine classification algorithm based on adaptive feature weight updating in the Hadoop cluster environment. PLoS One 14:1–18 . doi: 10.1371/journal.pone.0215136
5. Freeman DM, Hwa T (2013) Detecting Clusters of Fake Accounts in Online Social Networks Categories and Subject Descriptors
6. Hu X, Huang Y, Gao X, Luo L, Duan Q (2018) Squirrel-Cage Local Binary Pattern and Its Application in Video Anomaly Detection. IEEE Trans Inf Forensics Secur PP:1 . doi: 10.1109/TIFS.2018.2868617
7. Huang D, Mu D, Yang L, Cai X (2018) CoDetect : Financial Fraud Detection with Anomaly Feature Detection. IEEE Access 3536: . doi: 10.1109/ACCESS.2018.2816564

8. Kumar MS, Soundarya V, Kavitha S, Keerthika ES, Aswini E (2019) Credit Card Fraud Detection Using Random Forest Algorithm. 2019 Proc 3rd Int Conf Comput Commun Technol ICCCT 2019 149–153 . doi: 10.1109/ICCCT2.2019.8824930
9. Liu C, Chan Y, Hasnain S, Kazmi A, Fu H (2015) Financial Fraud Detection Model : Based on Random Forest. Res Gate. doi: 10.5539/ijef.v7n7p178
10. Neto JJDM, Santos JA, Schwartz WR Meat adulteration detection through digital image analysis of histological cuts using LBP
11. Nipane VB, Kalinge PS, Vidhate D, War K, Deshpande BP (2016) Fraudulent Detection in Credit Card System Using SVM & Decision Tree. 1:590
12. Patgiri R, Varshney U, Akutota T, Kunde R (2019) An Investigation on Intrusion Detection System Using Machine Learning. Proc 2018 IEEE Symp Ser Comput Intell SSCI 2018 1684–1691 . doi: 10.1109/SSCI.2018.8628676
13. Tejashwini SG (2017) Fraud Detection in Examination using LBP method. 2:28–35
14. V. D, R. D (2012) Behavior Based Credit Card Fraud Detection Using Support Vector Machines. ICTACT J Soft Comput 2:391–397 . doi: 10.21917/ijsc.2012.0061