Journal of Science and Technology ISSN: 2362-1443
2023, Vol. 3, No. 1, 40-45

Comparative Study of k-Nearest Neighbor (KNN) and k-Means Algorithm in Fraud Detection

Jendi Bade Shrestha,
Nepal College of Information Technology,
jendibade@gmail.com

Suresh Pokharel
Nepal College of Information Technology,
suresh@ncit.edu.np

Article History:
Received: 11 July 2023
Revised: 8 October 2023
Accepted: 3 December 2023

Keywords*—*K-Nearest Neighbor; K-Means; classification; clustering

***Abstract*—Fraud detection especially in credit card is one of the challenging issues in these days. Finding irregularities is even more difficult due to high volume of data during the transaction. Many data mining techniques are applied by researchers for solving these problems. In this research, we explore K-Nearest Neighbor (KNN) and k-means algorithms which are widely used classification and clustering algorithms respectively. These algorithms are used in this research to find out the better among them. Moreover, we also optimize the system by finding the most dominant and influencing factor responsible for fraud which will help in effective fraud detection in case of credit card.**

I. INTRODUCTION

Everyday huge amount of data and information are generated in any organization. The computational complexity and time complexity is very high if the task is performed in conventional way. So, the job has to be done by the use of clustering or classification algorithms which are widely used data mining techniques.

When there are lots of data in the dataset, Clustering and Categorization have problem on efficient categorization. Some of the attribute plays vital role in creating such problem. So the thesis will focus on finding such dominant attribute. During clustering and categorization by using KNN and K-means algorithm the main issue is to find the proper value of K for which accuracy and efficiency is high, so the research will also be focused on finding the value of K for fraud detection problem. This research focus various categorizations and clustering techniques used for data categorization and will find out the best between KNN and K Means algorithms on different basis. It will also be concentrated on the comparison between them and find out the best among them.

II. LITERATURE REVIEW

The data mining functionalities are used to specify the kind of patterns to be found in the data-mining task. The data mining functionalities mainly include association rule mining, classification, prediction & clustering. Association analysis is used for discovering interesting relations between variables in large databases, which in given in the form of rules to user. Classification predicts the class labels. Prediction is used to access the value of an attribute that a given sample is likely to have. Clustering is the process of grouping the data into classes or clusters so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Classification is supervised learning algorithms in contrasts with clustering, which are unsupervised learning algorithm [1].

A lot of research work has been done on classification and clustering of data sets in different fields such as medical data classification, news mining, weather forecasting, stock market prediction, text mining, fraud detection and many more. R. Gayathri, A. Malathi [2] applied data mining approach for credit card fraud detection. The five most frequently used classification techniques were applied in fraudulent detection. Neural Network, Decision Tree, Naïve Bayes, k-NN and Support Vector Machine are taken in to consideration discussed on each technique and their limitations. The accuracy of most of these classifiers is in the range of 66. 6% to 77.7%. Hybrid K-means and Decision tree [4] achieved the classification accuracy of 92.38% using 10 fold cross validations, cascaded learning system based on Generalized Discriminate analysis (GDA) and Least Square Support Vector Machine (LS_SVM), showed accuracy of 82.05% for diagnosis of Pima dataset [5].

A. G, Karegowda , M.A. Jayaram, and A.S. Manjunath [3] applied k-nearest neighbor (k-NN) classifier for classification of standard medical database for diabetic. Incorrect labeled instance are eliminated using K-means clustering followed by feature extraction using GA_CFS that had classification accuracy 79.50% using Cascaded GA_CFS_ANN, relevant feature identified by Genetic algorithm with Correlation based feature selection is given as input to ANN, 77.71% [3] using GA optimized ANN, 84.10% using GA optimized ANN with relevant features

identified by decision tree and 84.71% [3] with GA optimized ANN with relevant features identified by GA_CFS.

P. W. Buana, D.R.R. Sesaltina Jannet, I.K.G.D. Putra [3] combine traditional KNN algorithm and K-Means cluster algorithm for news mining where they applied grouping all the training samples of each category of K-means algorithm, and take all the cluster centers as the new training sample. The modified training samples are used for classification with KNN algorithm. Finally, calculate the accuracy of the evaluation using precision, recall and f-measure. The results showed that the combination of the proposed algorithm in that study had a percentage accuracy 87%, an average value of f-measure evaluation= 0.8029 with the best k-values= 5. Accuracy had been compared for different values of k along with the different news category.

Fuzzy clustering techniques are quite popular in various research on the data mining domains, P. K. Jena, S. Chattopadhyay [7] applied fuzzy logic in k-nearest neighbor classification and in C-means algorithm where Fuzzy clustering techniques handle the fuzzy relationships among the data points and with the cluster centers and the distance measures compute the load of fuzziness. Investigation on the effects of cluster fuzziness and three different distance measures, such as Manhattan distance (MH), Euclidean distance (ED), and Cosine distance (COS) on Fuzzy c-means (FCM) and Fuzzy k-nearest neighborhood (FkNN) clustering techniques, implemented on Iris and extended Wine data. The quality of the clusters is assessed based on (i) data discrepancy factor (i.e., DDF, proposed in this study), (ii) cluster size, (iii) its compactness, (iv) distinctiveness, (v) execution time taken, and (vi) cluster fuzziness (m) values and the result showed that FCM handles the cluster fuzziness better than FkNN. MH distance measure yields the best clusters with both FCM and FkNN. Finally, best clusters are visualized using a Self Organizing Map (SOM).

J. Kim, B.Kim, S. Savarese [8] used a general model in order to compare two different classification methods, K-Nearest-Neighbor (KNN) and Support-Vector-Machine (SVM) and observed that the SVM classifier outperformed the KNN for classification task for images. Classification with a 5-fold validation set, each fold with approximately 2800 training images and approximately 700 testing images was performed. Each experiment had different images in training and testing compared to the other experiments so as to prevent the overlapping of testing and training images in each experiment. G. Kalyani, K. K. Jyothi, V. N. Rao and D.Rambabu [11] discussed several ways in which an offender performs a credit card fraud and also reviews of diverse algorithms such as Hidden Markov Model, Bayesian Learning, Genetic Algorithm, Neural Network, Artificial Immune System, Support Vector Machine, K- nearest neighbor algorithm, Fuzzy Logic Based System and Decision Tree that can be used to overcome the frauds in plastic digital and virtual currencies. Their work identifies the perspective use of several methods for credit card fraud identification. M. V.

Kumar and B. K. Sriganga [12] also revies on several data mining techniques inside fraud detection in bank that highlight on common insider frauds occurring in banks and also tries to categorize them into different types. They categorize different types of frauds, their definitions, factors affecting them and the challenges faced in detecting them. The work also lists out different data mining techniques with their generic use, also with respect to the insider fraud detection and explained the best available data mining techniques, proposed by many researchers and currently employed in different industries.

K.K. Tripathi, M. A. Pavaskar did survey on credit card fraud detection methods [14]. Their study focused mainly on response of techniques among techniques based on Artificial Intelligence, Data mining, Neural Network, Bayesian Network, Fuzzy logic, Artificial Immune System, K- nearest neighbor algorithm, Support Vector Machine, Decision Tree, Fuzzy Logic Based System, Machine learning, Sequence Alignment, Genetic Programming. There can be many ways of detection of credit card fraud. If one of these or combination of algorithm is applied into bank credit card fraud detection system, the probability of fraud transactions can be predicted soon after credit card transactions by the banks. And a series of anti-fraud strategies can be adopted to prevent banks from great losses before and reduce risks. Among the several techniques the selection of the techniques based on the data quality and applicability of data on time. V. I. Memon, G. S. Chandel [15] applied k-Means (KM), K-nearest neighbor (KNN) and Decision Table Majority (DTM) (rule based) approaches for anomaly detection classify them into four categories according to risk level.

III. METHODOLOGY

The overview of the research is as shown in the figure below. The following diagram describes the approach that is applied for effective and appropriate finding with the data mining tools and techniques.

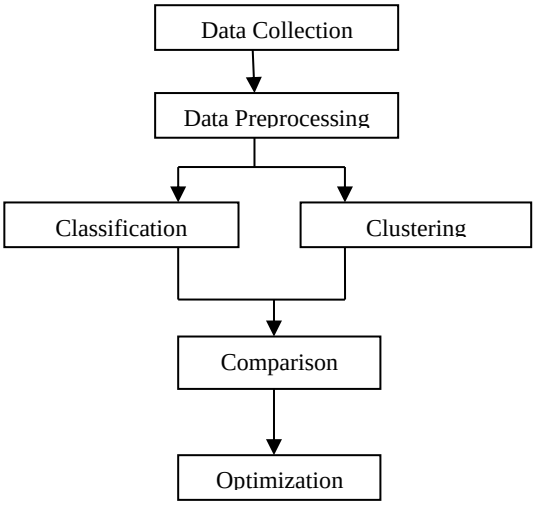

Fig. 1. Figure 1: Block diagram of Research Overview

A. *Data Collection:*

For this research work, data is downloaded from the internet. The data is of the German credit fraud dataset.

Different attributes of the data set are Over_draft, credit_usage, credit_history, purpose, current_balance, average_credit_balance, employment, installment_rate, personal_status, other_parties, residence_since, property_magnitude, cc_age, other_payment_plans, housing, existing_credits, job, num_dependents, own_telephone,foreign_worker, class.

B. *Data Preprocessing*

For data preprocessing the numeric missing data are replaced by mean value and for non-numeric it is replaced by most frequent one.

C. *Classification*

Classification assigns items on a collection to target categories or classes. For this research Instance Based classification is applied. K-Nearest Neighbor (KNN) algorithm is selected for the classification. Value of k is randomly and iteratively changed to find the appropriate classification. For fraud detection, sample data are classified with different values of number of neighbor. With varying number of neighbors for grouping the fraud data, the significant attribute need to be identified for fraud detection from training data.

D. *Clustering*

Cluster is a collection on data objects in which the objects are similar to one another within the same cluster and dissimilar to objects of another cluster. Using K-Means algorithm value of k is randomly and iteratively changed to find the appropriate clustering. Both Euclidean and Manhattan distance are applied for distance measure and hence the training data is clustered to find the abnormal data.

E. *Result Comparison*

At this stage, result obtained from both classification and clustering is compared. Confusion matrix data is used for the accuracy and other parameter comparison. Recall and precision are identified for both cases.

F. *Optimization*

When the preprocessed data is classified and clustered by using KNN and K-means algorithm respectively optimal result is not obtained in numerous iterations. So, for optimum classification and clustering on finding fraud detection, appropriate value of k is identified.

G. *Validation Criteria*

After model building, knowing the power of model prediction on a new instance, is very important issue. To measure the performance of a predictor, there are commonly used performance metrics, such as confusion matrix. In classification problem, the primary source of performance measurements is confusion matrix.

TABLE I. FORMAT OF CONFUSION MATRIX

	Classified Positive	**Classified Negative**
Positive Instances	True Positive	False Positive
Negative Instances	False negative	True Negative

Using the confusion matrix value parameter used for the statistical measure are TP Rate, FP Rate, Precision, Recall, True Negative Rate, Prevalence, Error Rate, Accuracy.

Once the result is obtained using KNN and K-Means for classification and clustering respectively, validation is applied using test data. Sample data is divided into training data and test data.

For further validation another dataset of insurance is also taken which have large number of instances comparatively.

IV. EXPERIMENT AND RESULT

A. *Experimental Data*

The dataset is German Credit fraud data which consists of 1000 instances with 21 fields.

B. *Experimental Environment*

The experimental environment that is set up for carrying out research is as below.

- Hardware: Intel(R) Core(TM) i5 – 3337U CPU @ 1.80 GHz, 4GB RAM
- Operating System: Windows 8 32-bit
- Tool: WEKA 3.7.

C. *Preprocessing*

Data is first fit for data preprocessing to identify the quality of data and data fields. Altogether 20 individual attributes are examined with their range of values. Missing values, uniqueness are also checked for the attributes values. Data preprocessing is applied to on credit fraud data to check out the individual attribute's data range. Data is also processed for missing values. Missing values are replaced with mean value for numerical values and with most frequent value for non-numeric value.

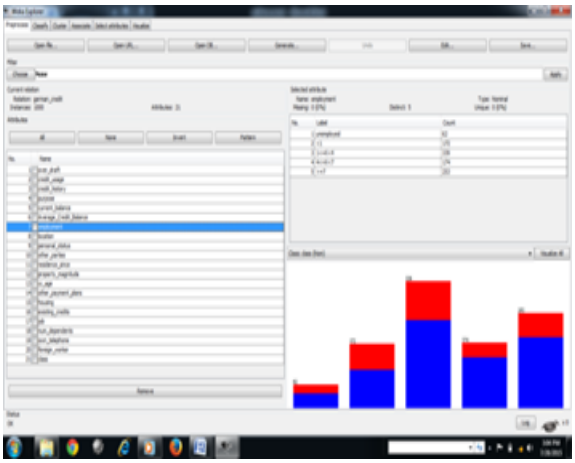

Fig. 2. Preprocessing of Data

Fig. 3. Classification of Dat

D. *Experiment and Results*

At next stage, the data has been loaded for classification. Among several algorithms of classification, KNN is selected for the classification. Two values for preliminary investigation are done with neighbor value (k values) as 3 and 5. For both cases different parameters are generated as shown below along with the time required to generate the result.

TABLE II. SUMMARY OF EVALUATION OF IB1 INSTANCE BASED CLASSIFIER USING 3 NEAREST NEIGHBOR FOR CLASSIFICATION

Time taken to build model	0.02 seconds
Correctly Classified Instances	86 %
Incorrectly Classified Instances	14 %
Kappa statistic	0.6457
Mean absolute error	0.2102
Root mean squared error	0.318
Relative absolute error	50.027 %
Root relative squared error	69.3906 %
Total Number of Instances	1000

With the increases in the value of k the time for classification decreased. The Incorrectly Classified Instances percentage also increases with the increase in the number of neighbor. Different resulting parameters found are shown in corresponding sections.

TABLE III. CONFUSION MATRIX FOR 3 NEAREST NEIGHBOR FOR CLASSIFICATION

	Good	**Bad**
Good	661	39
Bad	101	199

TABLE IV. SUMMARY OF EVALUATION OF IB1 INSTANCE BASED CLASSIFIER USING 5 NEAREST NEIGHBOR FOR CLASSIFICATION

Time taken to build model:	0.01 seconds
Correctly Classified Instances	82.3 %
Incorrectly Classified Instances	17.7 %
Kappa statistic	0.5357
Mean absolute error	0.2531
Root mean squared error	0.3505
Relative absolute error	60.2387 %
Root relative squared error	76.4831 %
Total Number of Instances	1000

TABLE V. CONFUSION MATRIX FOR 5 NEAREST NEIGHBOR

	Good	**Bad**
Good	660	40
Bad	137	163

The classification is experimented by iteratively increasing different values of K (nearest neighbor). It is observed that the K= 7 has the least error rate for the given training data.

TABLE VI. CONFUSION MATRIX FOR 5 NEAREST NEIGHBOR

	Good	**Bad**
Good	800	34
Bad	23	143

TABLE VII. SUMMARY OF EVALUATION OF IB1 INSTANCE-BASED CLASSIFIER USING 7 NEAREST NEIGHBORS

Time taken to build model:	0.01 Seconds
Correctly Classified Instances	94.3 %
Incorrectly Classified Instances	5.7%
Kappa statistic	0.5533
Mean absolute error	0.2243
Root mean squared error	0.3106
Relative absolute error	53.3747 %
Root relative squared error	67.789 %
Total Number of Instances	1000

From above results, it is noted that the accuracy varies with the change in value of K. For KNN classification, choosing the appropriate values is always the challenging issues. During the experiment the lower and higher values of K has lower accuracy. For this sample fraud detection data, value of K= 7 is the most appropriate for KNN classification. During the validation 5 fold and 10 fold method are applied that gives the incorrectly classified data on the lower range which shows the significance of the classification of fraud data using KNN that is the KNN is highly suitable for the classification by defining appropriate K value in fraud detection.

For clustering, k-means is selected. At preliminary investigation two different tests are performed. Values of k are set as 2 and 4. When 2 and3 clusters are selected, there is no any sign of bed credit card, but when the cluster number is increased to 4 then the bed credit along with attributes are identified.

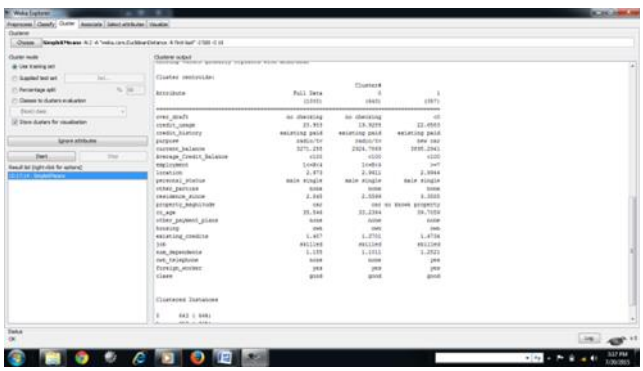

Fig. 4. Clustering of Data

TABLE VIII. SUMMARY OF CLUSTERING IN TO 2CLUSTERS USING K-MEANS ALGORITHM

Cluster	Number of instances
Cluster 0 (Good)	643
Cluster 1 (Good)	357

TABLE IX. SUMMARY OF CLUSTERING IN TO 4 CLUSTERS USING K-MEANS ALGORITHM

Cluster	Number of instances
Cluster 0 (Good)	253
Cluster 1 (Bad)	**140**
Cluster 2 (Good)	257
Cluster 3 (Good)	350

The experiment is repeated with several runs to identify the key attributes that help to identify the fraud cases. After number of iteratively applying the clustering using k-means algorithm, it is notified that employment year and credit card age are two distinct attribute that shoe the featured characteristics in fraud cases.

E. Result

KNN algorithm is used for classification whereas K-Means algorithm is used for clustering. Both classification and clustering are used for grouping the data. Classification is considered as supervised algorithm and clustering is considered as unsupervised algorithm. The KNN algorithm is one of the simplest of strategies for classification. The KNN algorithm considers each unknown object in the test data set, and it finds the k nearest examples in the training set. Whichever label was most common among those top k examples within the training set, that is the label which is assigned to the unknown object in the test set. In K-means algorithm k numbers of clusters are generated by iteratively assigning objects near to the center object.

While the data set was analyzed with k = 7 for KNN algorithm using WEKA, the result is obtained as:

TABLE X. CONFUSION MATRIX FOR KNN WHERE K=7

	Good	Bad
Good	800	34
Bad	23	143

Total number of Bad instances: 166

Classified as Bad: 177

Analyzing Data with k = 4 for K-Means algorithm. It is obtained as:

TABLE XI. CONFUSION MATRIX FOR K-MEANS WHERE K=4

Cluster	Number of instances
Cluster 0 (Good)	253
Cluster 1 (Bad)	**140**
Cluster 2 (Good)	257
Cluster 3 (Good)	350

From above two tables it is noted that in bad instances generated by K-Means algorithm is very much similar to bad instance identified by KNN algorithm (140 in comparison to 176). But in case of KNN the data classified as bad also includes incorrectly classified data that is good data classified as bad. That indicated data classified with KNN as bad is not purely bad data where as in clustered with K-Means may miss some of the data but has no wrongly grouped data.

With this result it can be concluded that correctness in grouping the data is better in case of K-means algorithm whereas KNN is simple to group since it uses only k nearest neighbor for grouping.

From these results it is observed that among the different search for nearest neighbor, linear search is the fastest on classification. As the value of K increases for classification, the absolute error increases. So, classification using KNN, lower range of value for K is preferred.

At second phase the samples are clustered using K-means algorithm. In this case two different distance formula (Euclidean distance and Manhattan distance) are applied. The results obtained are as below.

Thus, it is concluded that, several iterations with different combinations of evaluation parameters are applied for analysis. It is analyzed that clustering using K-Means having Euclidean distance has less incorrectly clustered instances in comparison to the Manhattan distance but it is observed that when Manhattan distance is used, it required less iteration for clustering than to Euclidean distance.

Since the classification and clustering are two different approaches for grouping of data. The KNN algorithm of classification and K-Means algorithm of clustering are used for sample data analysis. The results are compared to analyze the effectiveness and efficiency of these algorithms. It is observed that KNN algorithm is more suitable for grouping of data when the grouping parameters are well defined.

V. CONCLUSION AND FUTURE WORK

Data mining techniques are very useful in identifying the hidden knowledge as well as pattern in data set. KNN is one of the effective algorithms for classification because of its simplicity whereas K-Means is the widely used algorithm for clustering purpose. When different values of k were iteratively applied it is found that both very small and very large values of K are not suitable for accurate classification. Applying these algorithms for fraud cases in credit card gives the preliminary grouping of the fraud cases. Among several attributes it is needed to identify the most significant attributes. For that attributes need to be categorized. At next stage the attributes were categorized and then classified using KNN and clustered using K-Means. The most featured attributes among several attributes are identified. During clustering, it was applied using Euclidean and Manhattan distance and Euclidean distance is found to have less error than Manhattan. While classification using KNN algorithm various search namely linear search, cover search, ball tree search and KD search was used and it is found that linear search performs better than other searches during classification.

The thesis is only limited to German credit fraud data set and one each from the classification and clustering algorithms are used. So in future further the research can be carried out

to some real Nepalese data as well or some other credit card fraud data. Also, other algorithms can be used for the classification and clustering of fraud data.

REFERENCES

[1] J. Han, and M. Kamber, *Data Mining: Concepts and Techniques*, San Francisco, Morgan Kauffmann Publishers, (2001).

[2] R. Gayathri, A. Malathi, *Investigation of Data Mining Techniques in Fraud Detection: Credit Card*, International Journal of Computer Applications (0975 – 8887) Volume 82 – No.9, November 2013.

[3] A.G. Karegowda , M.A. Jayaram, A.S. Manjunath, *Cascading K-means Clustering and K-Nearest Neighbor Classifier for Categorization of Diabetic Patients*, International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-1, Issue-3, February 2012.

[4] B.M Patil, R.C Joshi, Durga Tosniwal, *Hybrid Prediction model for Type-2 Diabetic Patients*, Expert System with Applications, 37, 2010, 8102-8108.

[5] Polat, K., Gunes, S., & Aslan, A., *A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine*. Expert Systems with Applications, 2008, 34 (1), 214–221.

[6] P. W. Buana, D.R.R. Sesaltina Jannet, I.K.G.D. Putra, *Combination of K-Nearest Neighbor and K-Means based on Term Re-weighting for Classify Indonesian News*, International Journal of Computer Applications (0975 – 8887) Volume 50 – No.11, July 2012.

[7] Pradeep Kumar Jena, Subhagata Chattopadhyay, *Comparative Study of Fuzzy k-Nearest Neighbor and Fuzzy C-means Algorithms*, International Journal of Computer Applications (0975 – 8887) Volume 57– No.7, November 2012.

[8] J. Kim, B.Kim, S. Savarese, *Comparing Image Classification Methods: K-Nearest-Neighbor and Support-Vector-Machines*, Applied Mathematics in Electrical and Computer Engineering, ISBN: 978-1-61804-064-0, 133 – 138.

[9] R.Malarvizhi1, A. S. Thanamani, *K-NN Classifier Performs Better Than K-Means Clustering in Missing Value Imputation*, IOSR Journal of Computer Engineering (IOSRJCE), ISSN: 2278-0661, ISBN: 2278-8727 Volume 6, Issue 5 (Nov. - Dec. 2012), PP 12-15.

[10] D. L. Abd AL-Nabi, S. S. Ahmed, *Survey on Classification Algorithms for Data Mining:(Comparison and Evaluation)*, Computer Engineering and Intelligent Systems ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online) Vol.4, No.8, 2013.

[11] G. Kalyani, K. Krishna Jyothi, Prof.T.Venkat Narayana Rao, D.Rambabu, *A Comprehensive Study of Mechanisms Dealing Credit Cards to Defy Social Engineering Crimes*, International Journal of Computer Trends and Technology (IJCTT) – Volume 19 No.1 – Jan 2015.

[12] M. V. Kumar and B. K. Sriganga, *A Review on Data Mining Techniques to Detect Insider Fraud in Bank*, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 12, December 2014.

[13] *Navigating the Challenging environment - India Banking Fraud Survey – 2012.*

[14] K.Tripathi, M. A. Pavaskar, *Survey on Credit Card Fraud Detection methods*, International Journal of Emerging Technology and Advanced Engineering ISSN 2250-2459, Volume 2, Issue 11, November 2012.

[15] V. I. Memon, G. S. Chandel, *A Design and Implementation of New Hybrid System for Anomaly Intrusion Detection System to Improve Efficiency*, Int. Journal of Engineering Research and Applications, SSN : 2248-9622, Vol. 4, Issue 5(Version 1), May 2014, pp.01-07