Terrorist Group Prediction Using Data Classification

Faryal Gohar[1], Wasi Haider Butt[2], Usman Qamar[3]
Department of Computer Engineering, College of Electrical & Mechanical Engineering
National University of Sciences and Technology
Rawalpindi, Pakistan
faryalgohar88@gmail.com[1], wasi@ceme.nust.edu.pk[2], usmanq@ceme.nust.edu.pk[3]

ABSTRACT

Terrorist attacks are the challenging issue across the world and need the attention of the practitioners to cope up deliberately. Predicting the responsible group of an event is a complicated task due to the lack of in depth terrorist historical data. Data mining classification techniques are largely used to resolve the problem. This research proposes a novel ensemble framework for the classification and prediction of the terrorist group that consists of four base classifiers namely; naïve bayes (NB), K nearest neighbour (KNN), Iterative Dichotomiser 3 (ID3) and decision stump (DS). Majority vote based ensemble technique is used to combine these classifiers. The results of individual base classifiers are compared with the majority vote classifier and it is determined through experiments that our approach achieves a considerably better level of accuracy and less classification error rate as compared to the individual classifiers.

KEYWORDS

Decision stump, ID3, K nearest neighbour, Naïve bayes, Ensemble classifier**.**

1 INTRODUCTION

The verdicts and prediction of terrorist group is an intricate task which depends upon various indications and factors. Terrorist attacks are leading issue in the present situation and are the central point of concentration for the whole world due to its complicated, synchronized and well-planned terrorist actions [1]. Data mining is a process to mine useful information from the huge datasets that was not possible with traditional techniques. Data mining analysis techniques are broadly used in the field of national defense, government, finance, medicine, crime analysis, expert prediction, engineering and insurance [2]. There are mainly two approaches used in data mining: supervised and unsupervised learning. In supervised learning, a data set is used to train by using some training model whereas in unsupervised learning technique no training set is used [3].

Classification and prediction is the prominent approach in data mining that is used in various fields. It is a predictive model that predicts the future trends based on some training datasets.

This research paper proposes a framework for terrorist group prediction that is based on data mining classification techniques, namely, Naïve Bayes (NB), Decision Stump (DS), K- Nearest Neighbour (KNN) and Iterative Dichotomiser 3 (ID3). The proposed model combines the different data mining models using the majority vote base classifier and it is valuable to determine the terrorist group.

1.1 **Motivation**

Prediction about the responsible group after a terrorism incident can be very useful for law enforcement agencies in order to device a reactive strategy. The findings and forecast of terrorist group is an obscure task which depends upon various factors. Terrorist attacks are leading issue in the present situation and are the focal point of concentration for the whole world due to its complicated, synchronized and well-planned terrorist actions. Predicting the responsible group for the occurred event is a complicated task due to the lack of in depth terrorist historical data. As per best of our

knowledge data classification has never been used for prediction of responsible terrorist groups.

The main motivation of this research is to find out and transform the facts into useful information that can let the security proficient's to make intellectual judgments.

The major contributions of this research are:

- The proposed model predicts the terrorist group that is responsible of an attack.
- Different set of classifiers are used as base classifiers namely, Naïve Bayes (NB), Decision Stump (DS), K-Nearest Neighbour (KNN) and ID3 and are then combined to make majority vote base ensemble.
- The proposed technique is specific in nature that can only be applied for the classification of terrorist group.
- The focal point of this research is to measure the performance of the classifier.
- Comparison of individual classifiers and ensemble classifier is performed and the one (ensemble classifier) showing the best results is proposed.

The further work is explained as follows; section 2 describes the literature review, section 3 describes the detailed architecture of the proposed framework, section 4 explains the results and experiments performed on data set with multiple classifiers and finally the section 5 illustrates the future work and concludes the whole work done.

2 LITERATURE REVIEW

There are various classification methods proposed by the researchers in machine learning, statistics and pattern recognition [7]. This section reviews the different data mining techniques that are being used for the classification and prediction and the prior work done on the respective topic. The techniques that are reviewed are Naïve Bayes, KNN, ID3 and Decision Stump [4], [7], [8].

Bayesian classifier is the supervised machine learning technique used to take decisions under the uncertain conditions. The concept of probability is used to classify the new entities. It looks into the past to predict the future [10]. According to the author J. Han [8] the attribute value of a given class is conditionally independent of the values of other attributes. In Bayesian analysis the final classification is produced by combining the both of information i.e. the prior and the likelihood, to form the posterior probability using the so called Bayes. According to the author, Tom.M.Mitchel in practicality there are some complexities with Bayesian classifier for instance, it requires prior information of probabilities and in absence of that they are frequently predicted on the basis of background knowledge and earlier available data about the original distributions. The other problem is the computational cost that is required to find out the Bayes finest hypothesis in common case. In certain conditions this cost can be minimized [14].

K nearest neighbor algorithm (KNN) is known as lazy learner that makes predictions based on KNN labels assigned to test sample [16]. K nearest neighbour is determined by measuring the distance between the new entered query and the previously known samples. Distance between the query instances is calculated by using the Euclidean distance formula most often. The bulk of K nearest neighbors is taken for the prediction for the entered query once the K nearest neighbor is assembled. The outcome of novel instance query is classified on the basis of mass of KNN in case, if the new query instance doesn't belong to the similar group [17]. KNN is famous for its simplicity, applicability, spontaneous maintenance. It supports the multiple data structures and can be expressed easily without the model of training. This algorithm carries some drawbacks as well like the selection of K value is done arbitrarily that greatly affects the accuracy of the algorithm. Its computational complexity is very

high due to its global search every time during the classification process. Many researchers Gora [19] Yingbing [20] Chen ZhenZhou[21] and , Xianqiong[22] have worked to overcome these defects.

ID3 is one of the popular DT algorithms that deal with the nominal data sets. It does not deal with the missing values [11]. ID3 is a classical version of the decision tree induction and its improved versions are; SPRINT, SLIQ and CART. It mainly works on the selections of attributes at all the levels of decision tree that bases on (Quinlan) information entropy [12]. Basically top-down greedy search approach is followed to constrict the tree, where each attribute is tested on every tree node [13]. In ID3 every non-leaf node is tested to get the maximum value of information where the average depth of decision tree is less. This algorithm is also a good selection where you need the accuracy as it improves the accuracy and speed of classification. It is good when dealing with large scale problem. As this algorithm works on the bases of information entropy hence it lacks on some points like it becomes the reason to build too large decision trees that leads to the poor structure so it gets difficult to determine constructive rules [12]. Furthermore, it has some other pitfalls such as; it does not have the quality of backtracking during the search. Once a property is selected for the test, it does not re-consider the selection so it makes the convergence easier for the outcome of local optimal instead of global optimal. ID3 is not suitable for the incremental learning because it does not recognize training samples incrementally. According to Linna and Xuemin ID3 is sensitive to noise [9].

Decision stump is decision tree machine learning technique that works with only one split. That is, it is a decision tree with one internal node (the root) which is immediately connected to the terminal nodes (its leaves). A decision stump makes a prediction based on the value of just a single input feature. Sometimes they are also called 1-rules. Decision stumps are often used as components (called "weak learners" or "base learners") in machine learning ensemble techniques such as bagging and boosting.

The author Abishek Sachan and Devshri Roy [1] has propsed a TGPM to predict the terrorist group in India using the historical data. The database is taken from GTD that includes the terrorist attacks in india from 1998 to 2008. The researchers have used the terrorist corpus, parameter's weight and value as input. The unsupervised learning clustering technique is used to form the clusters of the data. The selected parameters are attack type, weapon type, group type, hostage/kidnapping, location and sucide attacks etc. The mathematical equation is also used to perform some main steps. Parameter's weight effect the performance however, the overall performance attained by the proposed model is 80.41 %.

The author Pawan H. Pilley and S. S Sikchi [24] have reviewed the terrorist group prediction model and analysis is performed using CLOPE algorithm. Historical data is used to detect the terrorist group and an association is made between terrorist group and the attacks occurred before. CLOPE clustering algorithm is used to make the clusters of the data that is particularly for the categorical features. It is concluded through analysis that terrorist group can be predicted using the historical data.

Nicholas Stepenosky, D. Green, J. Kounios, C. M. Clark, and R. Polikar [25] state in their research that collection or combination of classifiers is now more popular than individual classifiers because of their better performance and superiority over individual classifier system. Ensemble techniques of classifiers include bagging, boosting, voting techniques. These techniques are quite effective on many applications. The main idea behind combining classifiers is that individual classifiers are diverse and chances of errors are higher while combining classifiers can reduce errors and results in better performance through averaging.

Lior Rokach [26] in his research states that the idea of ensemble classifiers is used to build a model by combining different classifiers for

better performance. Different type of techniques can be used for combining classifiers one of which is majority voting. In majority voting technique classification of an unlabeled instance is performed according to the class that obtains the highest number of votes.

3 MATERIAL AND METHODS

3.1 Data Set Information

The GTD data set used in this research is taken from an open source of the National Consortium for the Study of Terrorism and Responses to Terrorism (START) initiative at University of Maryland USA, which broadcasts the terrorism incidents reports about the globe from 1970 to 2012. This dataset provides the information on every event with respect to the date and locality of the event, armaments used, figure of fatalities and the most importantly the responsible group. The features of this data set are [5],[6];

- It's an open source, most comprehensive and world's largest dataset available on terrorism incidents
- It contains the information about more than 87,000 terrorist events
- It includes the vast information on 120 variables
- Since 1970 it contains the information over than 13,000 eliminations, 38,000 bombings and 4,000 kidnappings.
- This dataset is under the supervision of counseling board of 12 terrorism research experts
- The main intended user is academic community and they recognize the GTD well.

The selected attributes for the research are month, city, country, weapon type, attack type, target and group name. These attributes are selected on the basis of their relevance to the predicted attribute (terrorist group). The attribute group is consist of 1400 diverse terrorist groups however, in order to get the accurate results only the most frequently active groups are considered and this is done by applying thresholding. The most active groups that are considered are six in number. Furthermore, data set is partitioned into training and testing parts. A class label is assigned to the data set that is the terrorist group. Data mining classifiers are used to train the base classifiers and their evaluation is performed using the testing data set.

4 PROPOSED FRAMEWORK

The proposed framework consists of four classifiers namely; naïve bayes, K nearest neighbor (KNN), Iterative Dichotomiser 3 (ID3) and decision stump. Majority vote based ensemble is used to combine these classifiers.

4.1 Base Models for Classifiers

The base classifiers and ensemble approach is described below:

4.1.1 Naïve Bayes (NB) Classifier

Baysian classifier works as follows [15];

- Let X be an n-dimensional vector consists of attributes X=(x1, x2, x3,....., x_n) depicting n number of measurements performed on given sample.
- We suppose that there are m number of classes, (C1, C2, C3,....., C_m) and to calculate the probability that the vector X belongs to the class C_m an unknown data set is assigned only when;

  $P(Ci \mid X) > P(C_k \mid X)$ for $1<=k<=m, k\neq i$
- Hence the probability $P(Ci \mid X)$ is maximized and it is known as the maximum posterior hypothesis of the class for which it is maximized.

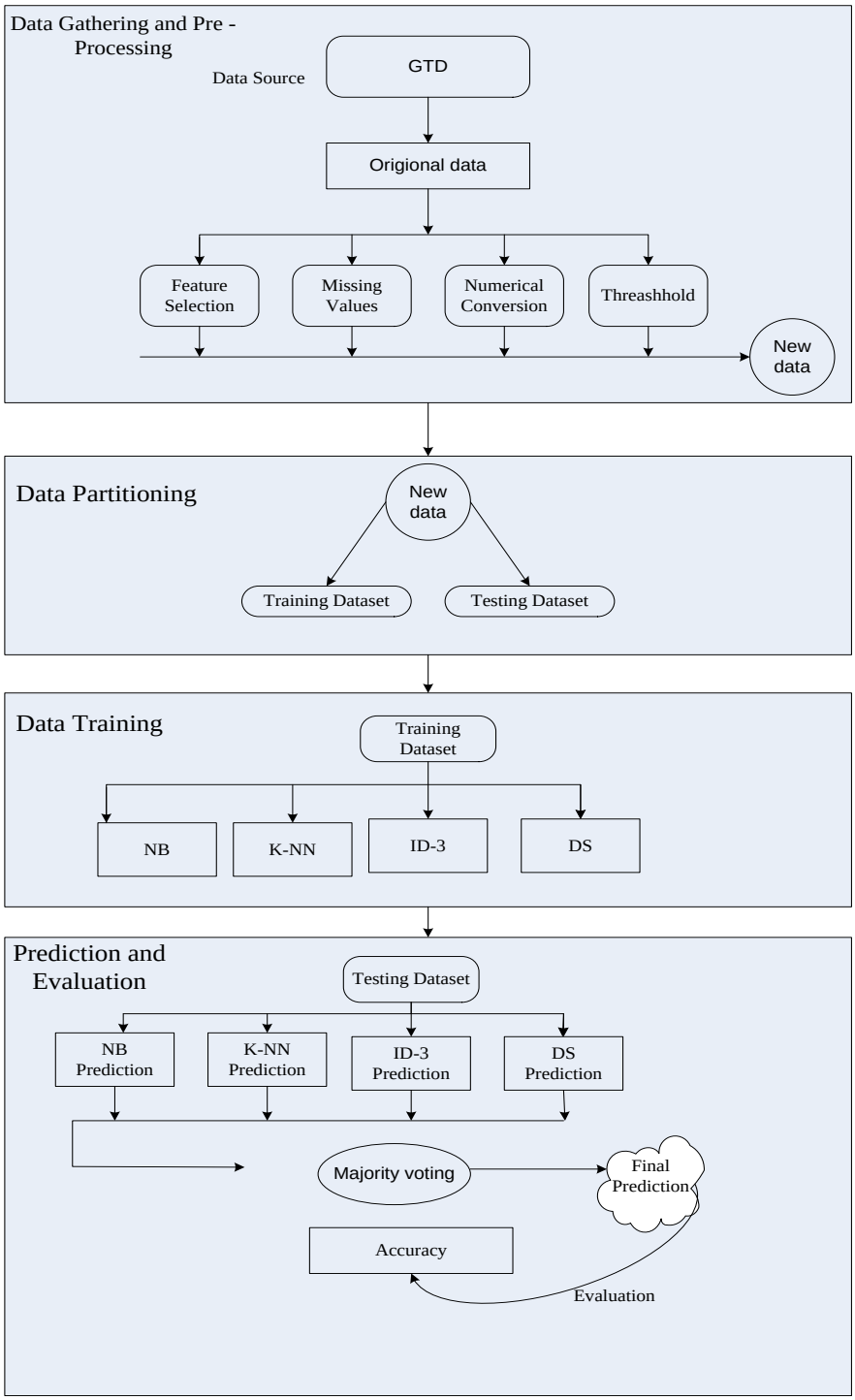

Figure 1: Architecture of proposed scheme

- The P(X|Ci) is merely required to be maximized as P(X) is continuous for every class.

$$P(C_i|X) = \frac{P(X|Ci) \cdot P(Ci)}{P(X)}$$

Where Ci is representing the class and X is known as an undefined data tuple. P(X|Ci) is posterior probability of X conditioned on C_i, P(Ci) and P(X) are the prior

probabilities and the P(Ci|X) is the posterior probability of Ci conditioned on X.

4.1.2 K Nearest Neighbour (KNN) Classifier

KNN is one of the top ten data mining algorithms used for the classification and regression. It is known as lazy learner that makes predictions based on KNN labels assigned to test sample [16]. KNN works as follows [18]:

- It assembles a training set of data D
- It selects the initial value of K, As there is no standard way followed to set the value of K, so it is selected randomly based on experimental results. The value of K is fixed according to the required results of the sample data. This is why selection of K-value is still an issue.
- It measures the distance between the sample point X and to its K neighbors by using the Euclidian distance formula. The distance between these samples is defined as:

$$d(X, Y) = \text{sqrt}\left(\sum_{i=1}^{n}(xi - yi)\right)$$

4.1.3 Iterative Dichotomiser (ID3) Classifier

This algorithm consists of a partitioned dataset D, a list of attributes that describes the tuples and an attribute selection method that identifies the course of action to select an attribute that best classify the provided tuples according to class. This heuristic method makes use of Gini Index or information gain for the attribute selection measure. The resultant tree of Gini Index is binary whereas others like information gain allow two or more branches to be grown from a node.

According to the Tom [14] attribute selection for the test at every node in the tree is the core option in the ID3 algorithm. During this selection the most valuable attribute is selected that becomes useful for the classifying data. This is done on the basis of a statistical property called information gain.

4.1.4 Decision Stump (DS) Classifier

Decision stump is a decision tree with one internal node (the root) which is immediately connected to the terminal nodes (its leaves). It's a tree graph where the leaves represent the classification results and the nodes represent some predicate. It's a tree traverse process where the start is taken from root and then we descend further [23].

4.2 The Proposed Ensemble Scheme

The proposed ensemble scheme for majority vote based joins the four classifiers. The proposed method illustrates a novel classifier on the basis of predicted values taken from numerous predecessor classifiers. This novel classifier is then used to analyze the testing data set. As the performance of individual classifiers is also measured therefore, this ensemble classifier attains the improved results as compare to the individual one. The proposed model mainly caters two steps. In first step individual classifiers are evaluated whereas in second step these individual classifiers are combined together to create a new ensemble method. These classifiers differ on the basis of their decisions.

4.2.1 The Proposed Ensemble Classifier (Majority Vote Based)

The proposed approach presents the majority vote based system that classifies and predicts the terrorist group. The framework mainly focuses on accuracy measure of the correctly classified attributes and classification error rate of the incorrectly classified attributes. It mainly involves the certain steps of preprocessing, data partitioning, data training and prediction and evaluation. Figure 1 illustrates the steps in detail.

4.2.1.1 Data Gathering and Preprocessing

The first step of this architecture is about data gathering and preparation that is gathered from the well known data source of GTD that broadcasts the terrorism incidents reports in relation to the globe from 1970 to 2012. The collected dataset provides the information on every event with respect to the date and locality of the event, armaments used; figure of fatalities and the most importantly the responsible group. There are few steps involved in preprocessing process that are followed in order to increase some quality attributes like the efficiency, specificity, sensitivity and the most importantly the accuracy of the classification and prediction process. Following are the steps which have been followed in data preparation process:

- **Data cleaning:** This is a core step of the preprocessing of the data where the removal or reduction of the inconsistent and noisy data is performed. Missing values of the data set are also treated there. These missing records are identified and removed from the data sets.
- **Feature Selection:** To increase the probability of the accurate results it is necessary to include only relevant data and to exclude the redundant attributes. Feature selection is performed to select the only relevant attributes of the data set. Only seven attributes are selected for the process where one attribute acts as a class label and rest of them are the regular attributes.
- **Data Conversion:** Data conversion is the process of converting the data from one form to the other required form. The GTD based data is converted from categorical to numerical version to improve the results. All the steps of data conversion are performed by using the Microsoft Excel.
- **Thresholding:** Thresholding is a point or a certain place from which a level starts or ends. It is applied on the labeled class named “Group name” in order to keep the most active groups. As the provided data consists of more than 1400 groups so, to come up with better results thresholding is applied. Frequently used groups are placed in a separate data set.

Classifiers are trained through training data set whereas evaluation and prediction is performed through testing data set.

4.2.1.2 Training Module

Once the data is preprocessed it is passed to training module to classify the individual classifiers in order to make them valuable in prediction. Terrorist groups are then identified through these trained classifiers. Once the individual classifiers are trained they are then combined together as a base classifier for the ensemble majority vote based approach.

4.2.1.3 Prediction and Evaluation Module

Prediction and evaluation is performed in this module. Trained classifiers are applied on new data to predict the class label first and then final results are predicted using majority vote base technique. Results of individual classifiers are compared to the proposed technique to perform the evaluation.

5 RESULTS AND DISCUSSIONS

Experiments are performed on dataset and results are analyzed respectively. This research consists of multiple classes of terrorist groups. Experiments are performed and results are evaluated and analyzed for the proposed approach as well as for the individual classifiers. Test data is classified into multiple classes in respect to the event occurred. Table 1 shows the classifiers and their accuracy level and table 2 shows the classification error rate of the classifiers. When all these base classifiers are combined we achieve the higher rate of accuracy and less no of classification errors. So the proposed technique considerably is better than the individual classifiers.

TABLE 1 PERFORMANCE LEVEL OF CLASSIFIERS

S.NO	CLASSIFIER	ACCURACY
1	Naïve Bayes	92.75 %
2	KNN	83.43 %
3	Decision Stump	91.30 %
4	Decision Tree ID3	84.97 %
5	MV (NB, KNN, ID3, DS)	93.40%

TABLE 2 CLASSIFICATION ERRORS OF CLASSIFIERS

S.NO	CLASSIFIER	ERROR RATE
1	Naïve Bayes	7.22 %
2	KNN	16.57 %
3	Decision Stump	8.70 %
4	Decision Tree ID3	15.03 %
5	MV (NB, KNN, ID3, DS)	6.64%

Figure 2 shows the graphical representation of the achieved results. It is shown through comparison of individual and majority vote based classifiers that the proposed techniques have better results than the individual one.

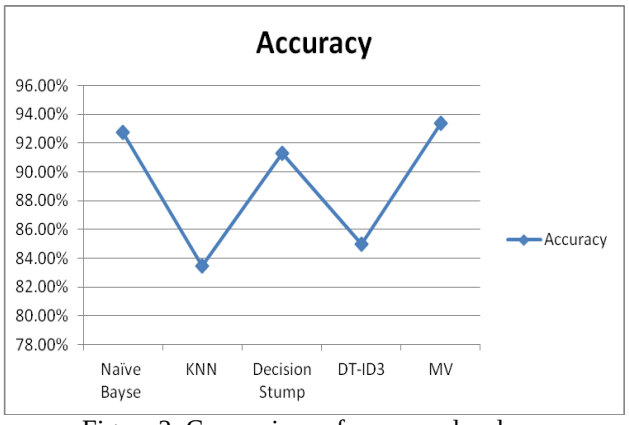

Figure 2: Comparison of accuracy level

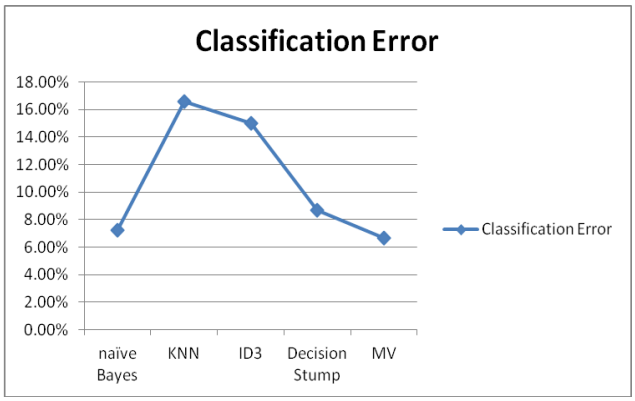

Figure 3 Comparison of classification error rate

The figure 3 is representing the graphical view of classification error rate of incorrectly classified attributes. The error rate of proposed technique is less than the individual classifiers.

To classify and predict the terrorist group all the preprocessing steps, data partitioning and testing plays an important role. As a result, the proposed technique successfully classifies and predicts the terrorist group.

6 FUTURE WORK

A data mining classification ensemble approach is introduced in this research for the classification and prediction of the terrorist group. The results of individual base classifiers are compared with the majority vote classifier and it is found that our novel approach presents the significantly better results. It is determined through experiments that our approach achieves a considerably better level of accuracy and less number of incorrectly classified attributes as compared to the individual classifiers. This work can further be extended by applying other data mining supervised and unsupervised learning techniques. Moreover other ensemble approaches can also be applied for comparison.

REFERENCES

[1] A. Sachan and D. Roy, "TGPM: Terrorist group prediction model for counter terrorism", in International Journal Of Computer Applications (0975 – 8887) Vol. 44– No10, April 2012.

[2] L. Li and X. Zhang, "Study of data mining algorithm based on decision tree", in 201O International Conference On Computer Design And Appliations (ICCDA 2010).

[3] I. SH and S. SA., "Intelligent heart disease prediction system using data mining techniques", in International Journal of Healthcare & Biomedical Research, Vol. 1, pp. 94-101, 2013.

[4] S. Özekes and O. Osman, "Classification and prediction in data mining with neural networks", in – Journal Of Electrical & Electronics Engineering, vol. 2003 : 3 : 1 (707-712). .

[5] K. Singh and S. Bhasin, "Modification of gtd from flat file format to OLAP for data mining", in International Journal Of Innovative Technology & Creative Engineering (ISSN:2045-8711) VOL .1 NO.4 APRIL 2011.

[6] S. Ejaz Hussain, "Terrorism in pakistan: Incident Patterns, Terrorists' Characteristics, and The Impact of Terrorist Arrests on Terrorism" (2010).Publicly accessible Penn Dissertations.Paper 136.

[7] H. Jantan, A.R Hamdan and Z. A Othman, . "Classification for talent management using decision tree induction techniques", 2nd Conference of IEEE, data mining and optimization Kajand, 2009. ISBN 978-1-4244-4944-6.

[8] J. Han, Data Mining Concepts and Techniques, second edition, pg no 310 - 312.

[9] L. Li, X. Zhang, "Study Of Data Mining Algorithm Based On Decision Tree", Changchun institute of technology Changchun, China, 201O International Conference On Computer Design And Appliations (ICCDA 2010) vol 1.

[10] M. J. Islam, Q. M. Jonathan Wu, M. Ahmadi and M.A. Sid-Ahmed, "Investigating the Performance of Naive- Bayes Classifiers and K- Nearest Neighbor Classifiers", 2007 International Conference on Convergence Information Technology, IEEE DOI 10.1109/ICCIT.2007.148.

[11] A. Cufoglu, M. Lohi and K. Madani, "A Comparative Study of Selected Classifiers with Classification Accuracy in User Profiling, 2008 IEEE DOI 10.1109/CSIE.2009.954.

[12] D. Chen and Z. Liu, "An Optimized Algorithm of Decision tree Based on Rough Sets Model", 2010 International Conference on Electrical and Control Engineering, 2010 IEEE DOI 10.1109/iCECE.2010.8046.

[13] A. B. E. D. Ahmed and I. S. Elaraby, "Data Mining: A prediction for Student's Performance Using Classification Method", World Journal of Computer Application and Technology 2(2): 43-47, 2014.

[14] T. M. Mitchel, "Machine Learning", Publisher: McGraw-Hill Science/Engineering/Math; (March 1, 1997), ISBN: 0070428077.

[15] Y. Yang and S. Elfayoumy, "Anti-Spam Filtering Using Neural Networks and Baysian Classifiers", Proceedings of the 2007 IEEE International Symposium on Computational Intelligence in Robotics and Automation Jacksonville, FL, USA, June 20-23, 2007.

[16] Z. Yan and C. Xu, Combining, "KNN Algorithm and Other Classifiers", Proc. 9th IEEE Int. Conf. on Cognitive Informatics (ICCI'10) F. Sun, Y. Wang, J. Lu, B. Zhang, W. Kinsner & L.A. Zadeh (Eds.) 978-1-4244-8040-1/10/$26.00 ©2010 IEEE.

[17] X. Xiao and H. Ding, "Enhancement of K-nearest Neighbor Algorithm Based on Weighted Entropy of Attribute Value", 2012 5th International Conference on BioMedical Engineering and Informatics (BMEI 2012).

[18] S. yan, L. Shipin and W. Yiukun, "KNN Classification Algorithm Based On The Structure Of Learning [J]". Computer Science, 2007,34 (12).

[19] C. D. Amato, ,D. Malerba , F. Esposito and M. Monopoli, "Extending The K-Nearest Neighbour Classification Algorithm To Symbolic Objects", Atti del Convegno Intermedio della Societ à Italiana diStatistica"Analisi Statistica Multivariata per le scienze eco2nomico2sociali,le scienze naturali e la tecnologia". Italia :Napoli ,2003.

[20] S. Yingbing and L. Qingsun, "KNN algorithm based on feature weighted", Hanna University of Science and Technology,2008, Vol. 26 No. 4,352-355.

[21] C. ZhenZhou, L. Lei and Y. Zhengan, "KNN Algorithm Based On SVM Weighting[J]", Sun Yat-sen University Natural Science Edition, 2005, 44 (1) : 17 - 20..

[22] T. Xianqiong and Z. Zhongmei, "Improved Algorithm Based On The KNN Information Entropy Of The Property Value [J]", Computer Engineering and Application 2010,46(3):115-117.

[23] G. Yang, Z. Zhou and X. Yu, "Hyperspectral Imagery Classification Based On Gentle Adaboost

And Decision Stumps", in Journal of IEEE, vol 978-1-4244-4994-1/09 ©2009.

[24] P. H. Pilley and S. S Sikchi, "Review Of Group Prediction Model For Counter Terrorism Using CLOPE Algorithm", in International Journal Of Advance Research In Computer Science And Management Studies, vol 2, issue 1, January 2014, ISSN: 2321 – 7782.

[25] N. Stepenosky, D. Green, J. Kounios, C. M. Clark, and R. Polikar, "Majority Vote And Decision Template Based Ensemble Classifiers Trained On Event Related Potentials For Early Diagnosis Of Alzheimers's Disease", In Proceedings of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pages 901–904, 2006.

[26] L. Rokach, "Ensemble-Based Classifiers", Artif Intell Rev (2010) 33:1–39 DOI 10.1007/s10462-009-9124-7, 2009.