An Experimental Study of Classification Algorithms for Terrorism Prediction

Ghada M. Tolan and Omar S. Soliman

***Abstract*—Terrorist attacks are the biggest challenging problem for the mankind across the world, which need the wholly attention of the researchers, practitioners to cope up deliberately. To predict the terrorist group which is responsible of attacks and activities using historical data is a complicated task due to the lake of detailed terrorist data. This research based on predicting terrorist groups responsible of attacks in Egypt from year 1970 up to 2013 by using data mining classification technique to compare five base classifiers namely; Naïve Bayes (NB), K-Nearest Neighbour (KNN), Tree Induction (C4.5), Iterative Dichotomiser (ID3), and Support Vector Machine (SVM) depend on real data represented by Global terrorism Database (GTD) from National Consortium for the study of terrorism and Responses of Terrorism (START). The goal of this research is to present two different approaches to handle the missing data as well as provide a detailed comparative study of the used classification algorithms and evaluate the obtained results via two different test options. Experiments are performed on real-life data with the help of WEKA and the final evaluation and conclusion based on four performance measures which showed that SVM, is more accurate than NB and KNN in mode imputation approach, ID3 has the lowest classification accuracy although it performs well in other measures, and in Litwise deletion approach; KNN outperformed the other classifiers in its accuracy, but the overall performance of SVM is acceptable than other classifiers.**

***Index Terms*—KDD, precision, recall, terrorist group.**

I. Introduction

Terrorist attacks are biggest, challenging, and leading issue in the whole world. It is one of the central points of concentration in all governments. Data mining is popularly known as Knowledge Discovery in Databases (KDD), it is a logical process of discovering new patterns from large data sets involving methods combined with statistics, database systems, support vector machine, artificial intelligence, meta-heuristics, and machine learning. The main goal of data mining is to extract useful, hidden predictive knowledge from large data sets in a human understandable structure and involves database, data management and pre-processing tools, model and interface capabilities, post-processing of discovered structure, visualization and online updating methods for finding hidden patterns, and predictive information that expert may miss because it lies outside their expectations [1], [2]. Data mining and automated data analysis techniques have become used as effective branch of the most important key features for many applications, data mining has a wide number of applications ranging from marketing and advertising of goods, services or products, artificial intelligence research, biological sciences, crime investigations to high-level government intelligence [3]. Recently there has been much concern on using data mining in detecting and investigating unusual patterns, crimes, terrorist activities and preventing the fraudulent behavior [4], some of different techniques used in that regard are entity extraction, clustering techniques, deviation detection, classification techniques, string comparator, and social network [5]. Data mining, Sentiment analysis, text mining, machine learning techniques and predictive analytics are some of methodologies being used to identify and combat terrorism [6].

Classification is an important task of data mining; it is a supervised class prediction technique [1]. The main goal is to accurately predict the class for each data [2], provided that sufficient numbers of classes are available. Classification has been previously used in many branches of research such as terrorism prediction, medical, finance, weather prediction, business intelligence, homeland security. Various approaches are used for classification of datasets, as there are numerous techniques for classification and rule extraction. Classification algorithms can be seen as probabilistic or non-probabilistic classifiers, other classify the classification algorithms as Binary and Multiclass classifier, where Binary classification is the task of classifying the elements of a given set into two groups on the basis of a classification rule.

American Historian and Terrorist Expert Walter Laqueur has counted over 100 definitions of terrorism, and concluded that the "only general characteristics agreed upon is that terrorism involves violence and threat of violence". In Political Terrorism: A New Guide to Actors, Authors, Concepts, Data Base, Theories, and Literature. They counted over 109 definitions of terrorism that covered a total of 22 different situations. Most define terrorism as" the use or threat of serious violence to advance some kind of cause", some state clearly the kinds of group ("Sub-national","non-state") or cause (political, ideological, religious) to which they refer. In our research study a real data set of Egypt is used for terrorism prediction based on data mining classification algorithms with the help of WEKA as one of open software in data mining written in JAVA [7].

The organization of this paper is as follows; Section II covers the literature review. Section III illustrates the methods and techniques used for terrorism prediction, discusses terrorism data set and collection methodology, data pre-processing steps, and classification with WEKA. Section IV explains experimental results, analysis, and performance measures of mode-imputation and Litwise deletion approaches in different classification test options illustrated with figures and tables. Finally, section V covers conclusions

Manuscript received February 14, 2015; revised June 2, 2015.

The authors are with Operations Research Department, Cairo University, Egypt (e-mail: gh.tolan@fci-cu.edu.eg, O.Soliman@fci-cu.edu.eg).

and future work.

II. Literature Review

There are various classification approaches proposed by the researchers in machine learning, statistics, and pattern recognition [8]. 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, C4.5, ID3, and SVM [8]-[10].

Bayesian (Naïve Bayes) Classifier is the supervised machine learning technique used to take decision under the uncertain conditions as well as a statistical method for classification. According to the author, D. Hongbo [11] Naïve Bayes makes the assumption that descriptive attributes are conditionally independent of each other given the class label is known; in other words, Bayesian Classifiers have the ability to predict the probability that a given tuple belongs to a particular class. According to the author, Tom. M. Mitchel [12] 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 original distributions. The other complexity is the computational cost that is required to find out the bayes finest hypothesis in common case, but in certain cases this cost could be minimized. The Advantages of Naïve Bayes Classifier as summarized by V. Batchu [13] and I. Rizwan [1] are; it proves success in solving different classification tasks effectively, as it is robust to isolated noisy data, and also robust against irrelevant attributes. The naïve bayes method can also cope with null values. Because of these advantages, the naïve bayes method is widely used for different applications.

K-Nearest Neighbor (KNN) Classifier is one of the top ten algorithms used for the classification and regression. It is, also known as lazy learner or instance-based, in that it stores all of the training samples and do not build a classifier until a new sample needed to be classified that makes predictions based on KNN labels assigned to test sample [9], KNN is based on learning by analogy, and it is amongst the simplest of all machine learning algorithms which can be used for prediction, that is, to return a real valued prediction for a given unknown sample [14]. KNN is famous for its simplicity, applicability, spontaneous maintenance. It supports the multiple data structures and can be expressed easily without the training model. Drawbacks of KNN are summarized by S. Neelamegam [14] such as KNN classifiers assign equal weight to each attribute; this may cause confusion when there are many irrelevant attributes in the data. Because KNN is a lazy learner, so it can incurs expensive computational costs when the number of potential neighbors is great, therefore they require efficient indexing techniques. The classification by KNN can be misleading if the chosen value of K is too large than it should be.

Decision Trees (DT) are predictive decision support tools that create mapping from observations to possible consequences, a statistical data mining technique that express independent and a dependent attributes logically and in a tree shaped structure [15]. As a major approach decision tree induction has received a great attention from the researcher in the last two decades [11], as a result there are a number of decision tree induction methods have been developed such as ID3, C4.5, C5, C&RT, and CHAID. According to the authors, D. Hongbo [11] and R. Kalpana [3], the strengths of DT are; it assigns a class label to an unseen record, as well as explains why the decision is made in an easy-to-understand classification rule. DT classifiers unseen records efficiently, and it can handle both categorical and continuous attributes, the attribute selection measures used by DT induction method are capable of indicating the most important attribute in relation to class. The researchers mentioned the weaknesses points of DT; it has high error rates when the training set contains a small number of instances of a large variety of different classes, DT algorithm may not work well on data sets when attribute split in any other shape exist. Decision Trees are automatically quite expensive to build.

ID3 is one of the popular DT algorithms that deal with nominal data sets, does not deal with missing values [16]. ID3 is the 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 base on (Quinlan) information entropy [9].This algorithms is a good selection where the research needs accuracy as it improves the accuracy and speed of classification; it is helpful when dealing with a large scale problem. As this algorithm works on the basis of information entropy hence it lacks on some points as stated by D. Chen [17] like it becomes the reason to build too large decision tree that leads to the poor structure and so it gets difficult to determine constructive rules. Furthermore, it has some other weaknesses such as; it does not have the quality of backtracking during the search, and it is sensitive to noise.

Support Vector Machine (SVM) is a new and promising method for regression, classification, and general pattern recognition. SVM aims to find the best classification function to distinguish between members of the two classes in the training as explained by S. Neelamegm [14]; with other words SVM tries to find a hyper plane to separate the two classes while minimizing the classification error [15]. The authors in [14] state some advantages of SVM as; it considers a good classifier because of its high generalization performance without the need to add a priori knowledge, and it has been successfully applied to a wide range of application areas. But SVM has a weak point which is computational inefficiently, but this problem has been solved by two methods.

The Author I. Rizwan and A. Masrah [1] compare two different classification algorithms namely, Naïve bayes and Decision Tree for predicting "Crime Category" for different states in USA. 10 fold cross validation was applied to the input dataset in the experiment, separately for both NB and DT to test the accuracy of the classifiers which showed that DT algorithm out performed NB algorithm and achieved 83.951% accuracy in predicting "crime Category".

The author G. Faryral, B. H. Wasi [9] have proposed a novel ensemble framework for the classification and prediction of terrorist group in Pakistan that is consists of four base classifiers namely; NB,KNN,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 the new approach achieves a considerably better level of accuracy and less classification error rate as compared to the individual classifiers.

The author Abishek Sachan and Devshri Roy [18] have proposed 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 mathematical equation is also used to perform some main steps. The overall performance attained by the proposed model is 80.41%.

The author Pawan H. Pillry and S. S Sikchi [19] has 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 algorithms 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.

III. Methods and Techniques Used for Terrorism Prediction

A. Terrorism Data Set and Collection Methodology

The GTD data set is an open source, most comprehensive and world's largest dataset available on terrorism incidents used for the experiment, taken from an open source of the National Consortium for the study of terrorism and Responses of Terrorism (START) initiative at University of Maryland USA, which broadcasts the terrorism incidents report about the globe from 1970 to 2013, and includes information about more than 87,000 terrorist events as well as the vast information on 120 variables, and contain information over than 13,000 eliminations, 38,000 bombing and 4,000 kidnappings.

B. Terrorism Data Set Pre-processing

The data set used for our research paper consists of a total of 869 terrorist events (instances), and 23 attributes, the attribute group is consisting of 35 diverse terrorist groups. Before applying classification algorithm(s) usually some pre-processing is performed on the data set. In order to perform data processing, it is essential to improve the data quality [20]. There are a few number of techniques used for the purpose of data pre-processing [11] as data aggregation, data sampling, dimension reduction, feature creation, data discretization, variable transformation, and dealing with missing values. It is necessary in our research to apply the following steps:

1) First step

Data reduction is performed on the terrorism data by selecting the most informative attributes without lose any critical information for classification and so only 6 attributes are selected from 23 attributes, there are different algorithms for attribute or feature selection. For this research a manual selection method was chosen for attribute selection based on our understanding of the application problem. The selected attributes are date, city, weapon-type, attack-type, target-type, and group-name. These selected attributes are related to the predicted attribute (Terrorist Group).

2) Second step

For the missing data values, there are three approaches to handle missing data elements: removal, imputation, and special coding [11]-[21]. In our research we applied two approaches; data removal, and data mode-imputation techniques for the missing data instances to produce two data bases, and then we will apply the selected classification algorithm(s) on each data set and compare between them via the classification accuracy and different performance measures as explained in Fig. 1.

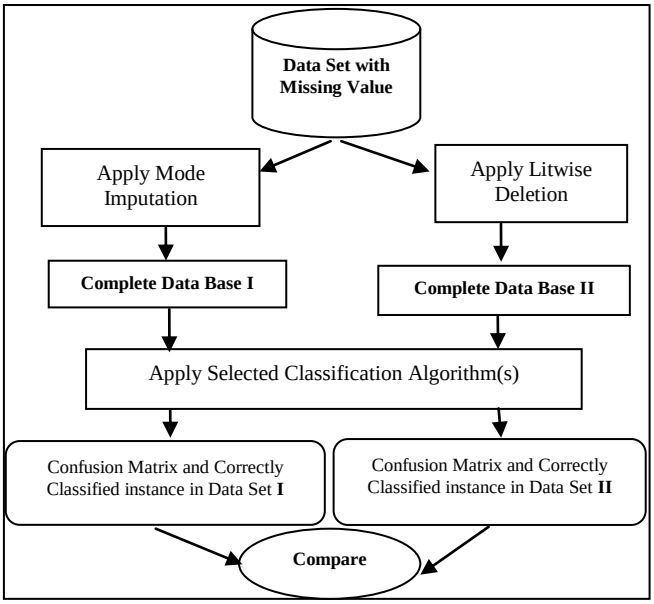

Fig. 1. Flow diagram for handling missing data.

3) Third step

Performing different classification algorithms on the research data set by using WEKA as one of important tools available for implementing data mining algorithms to train the base classifiers then the evaluation of the implemented classifiers is performed by using the testing data set.

C. Classification with WEKA

The classification algorithms in this research are implemented based on WEKA. Waikato Environment for Knowledge Analysis (WEKA) is an open source software written in JAVA, a collection of machine learning algorithms allows the researcher to mine his own data for trends and patterns. The algorithms can either be applied directly to a dataset or called from the researcher own JAVA code [22]. WEKA contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

The terrorism data base set of Egypt is splitting into two main sets: Training data set (with percentage split 66%), and Testing data set(with percentage split 34%) from the whole data set, and that is applied by using the default settings of WEKA.

IV. Experimental Results, Analysis, and Performance Evaluation Figures and Tables

In our experiment we applied different classification algorithms on the Terrorism data of Egypt from 1970 to 2013, by using two different approaches to handle the missing data instances, mode-imputation and litwise deletion methods with the help of WEKA Software. During experiment, pre-processed data set which consists from 869 data instances (records) is converted to .ARFF file to be used by WEKA. The classification algorithms results obtained according to two test options which are:

1) Evaluation on Test Split that divides the input data set into 66% for the training data and 34% for the test set.
2) 10 Fold Cross-Validation.

The results from the applied classification algorithms in the two approaches will be evaluated according to four performance measures which are defined bellow:

1) The Classification Accuracy: is the percentage number of correctly classified instances (the number of correct predictions from all predictions made)
2) Precision: is a measure of classifier exactness (used as a measure for the search effectiveness)
3) Recall: is a measure of classifier completeness.
4) F-Measure: also called F-Score, it conveys the balance between the precision and the recall.

A. Mode-Imputation Approach

In this approach we deal the missing data in our data set by using the mode and frequency distribution of the attributes to handle the missing data instances.

1) Evaluation of classification algorithms by test split

In case of dividing the input data set into 66% for the training data and the remaining 34% for testing the classifiers; the results are shown in Table I which provide a clear comparison among the selected classifiers according to accuracy, precision, recall, and F-measure which shows that:

TABLE I: Test Split & Accuracy Results of Mode-Imputation Approach

Acc. / Alg. Used	Accuracy (Correctly Classified Instances)	Incorrectly Classified Instances	Precision	Recall	F-Measure
NB	61.0169%	38.9831%	0.544	0.61	0.558
KNN	56.6102%	43.3898%	0.663	0.566	0.577
C4.5	56.6102%	43.3898%	0.320	0.566	0.409
ID3	32.2034%	8.1356%	0.796	0.798	0.796
SVM	67.1186%	32.8814%	0.623	0.671	0.633

From the accuracy point of view; SVM correctly classified about 67.1186% of the data; it means 180 items out of 295 in the 34% test split of the data SVM is outperformed NB which correctly classified about 61.0169% of the data. It is obvious that the accuracy of KNN and C4.5 are almost the same, and ID3 achieved lowest accuracy 32.2034% among the other classifiers although it has the highest precision, recall, and f-measure over KNN and SV classifiers. C4.5 classifier achieves lower Precision, recall, and f-measure values than other classifiers. The overall performance of NB is very near from KNN classifier.

It is obvious from Fig. 2 that a comparison is applied on our five classifiers due to precision, recall, and F-measure which shows that SVM has the highest accuracy could also consider with high precision, recall and F-measure results. The overall performance of NB is very near from KNN results. ID3 is out performed although it is not accurate.

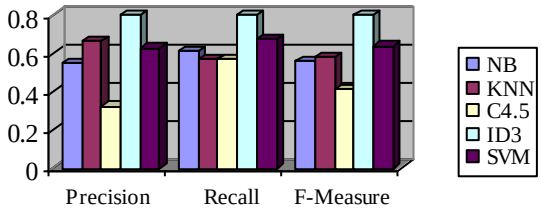

Fig. 2. Performance measures results of used classifiers.

2) Classification by using 10-fold cross validation

Similarly, Table II illustrates the accuracy and different performance measures of the classification algorithms used according to 10-fold cross validation of the input data set; SVM classifier correctly classified 580 data record out of 869 data records; this means that it successes to correctly classify about 66.7434% from the whole input data. KNN classifier is near from the SVM accuracy. ID3 classifier is not accurate hence it left about 56% of the input data unclassified (489 unclassified instances from 869), but it has higher precision, recall, and F-measure performance results.

TABLE II: 10 Fold Cross Validation & Accuracy Results of Mode-Imputation Approach

Acc. / Alg. Used	Accuracy (Correctly Classified Instances)	Incorrectly Classified Instances	Precision	Recall	F-Measure
NB	61.016%	38.983%	0.544	0.61	0.558
KNN	56.610%	43.389%	0.663	0.566	0.577
C4.5	56.610%	43.389%	0.320	0.566	0.409
ID3	32.203%	8.1356%	0.796	0.798	0.796
SVM	67.118%	32.881%	0.623	0.671	0.633

Fig. 3 shows the performance measures in case of using classification based on 10 fold cross validation where ID3 has higher precision, recall, and F-measure values than SVM, but it could not consider more accurate than SVM. KNN classifier performs well and very near from SVM especially in precision and F-measure results. NB classifier performs as KNN in most measures, and C4.5 performs badly than other classifiers in precision, recall, and F-measure.

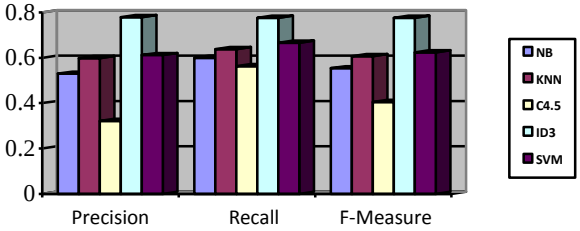

Fig. 3. Performance measures results of used classifier.

B. Litwise Deletion Approach

In this approach we deal with missing data instances in our real terrorism data set of EGYPT by using the Litwise deletion that does not affect the predicted attribute but caused a data dimension reduction that makes our real data more easier in the search space and reduce the time of pattern

discovering than imputation approach, then we entered our new data set as an input to WEKA software to be classified by the five classifiers and compare among them as explained in the following two subsections.

1) Evaluation of classification algorithms by test split

In letwise deletion approach, when the data is partitioned into two splits with percent 66 and 34 for testing the classifiers, a comparison between the used classifiers is made in Table III that shows; KNN is out performed the other classifiers in its accuracy especially SVM that proved success in the imputation approach where it classified successfully about 72.53% from the data into the correct class. ID3 has the lowest accuracy, because it leaves about 77.4648% of the test split instances without classification, it means it correctly classified 30 instances out of 142 instances in the test split. Other performance measures explained clearly in Fig. 4.

TABLE III: TEST SPLIT & ACCURACY RESULTS OF LITWISE-DELETION APPROACH

Acc. / Alg. Used	Accuracy (Correctly Classified Instances)	Incorrectly Classified Instances	Precision	Recall	F-Measure
NB	69.0141%	30.9859%	0.611	0.69	0.640
KNN	72.5352%	27.4648%	0.629	0.725	0.664
C4.5	56.3380%	43.6620%	0.317	0.563	0.406
ID3	21.1268%	1.4085%	0.939	0.938	0.937
SVM	71.8310%	28.1690%	0.629	0.718	0.666

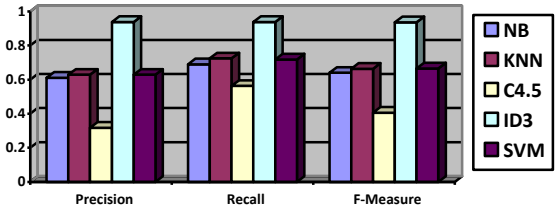

Fig. 4. Performance measures results of used classifiers.

It is obvious that ID3 has highest values in precision, recall, and F-measure than other classifier as in mode-imputation approach, although it performs badly in the accuracy. C4.5 has lowest precision, recall, and F-measure results. KNN and SVM performance measures are almost the same as they perform effectively in the first approach.

2) Classification by using 10-fold cross validation

TABLE IV: 10 FOLD CROSS VALIDATION & ACCURACY RESULTS OF LITWISE-DELETION APPROACH

Acc. / Alg. Used	Accuracy (Correctly Classified Instances)	Incorrectly Classified Instances	Precision	Recall	F-Measure
NB	69.0284%	30.0716%	0.628	0.699	0.655
KNN	73.0310%	26.9690%	0.682	0.730	0.699
C4.5	56.5632%	43.4368%	0.528	0.566	0.421
ID3	26.0143%	1.9093%	0.924	0.932	0.928
SVM	75.4177%	24.5823%	0.699	0.754	0.716

The results of our experiment based on using 10-fold cross validation represented in Table IV where SVM is more accurate than other classifiers; it classified about 75.41% from the whole data into the correct class. It is obvious that KNN is nearly has the same accuracy as SVM. ID3 has lower accuracy than all other classifiers where it could not classify more than 26% of the data into the correct class.

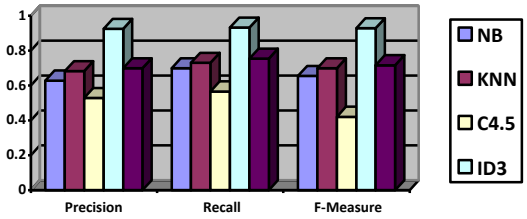

Fig. 5. Performance measures results of used classifiers.

We can notice from Fig. 5. That ID3 performs highly in precision, recall, and F-measures although it is not accurate, KNN, and SVM are almost the same in their results, NB precision, recall, and F-measures are very near from KNN c1assifier. C4.5 performs badly precision, recall, and F-measure.

V. CONCLUSIONS

A data mining classification ensemble approach is introduced in this paper research for the classification and prediction of the terrorist groups in Egypt from 1970 to 2013, the data used in our experimental study is based on real data represented by Global terrorism Database (GTD) from National Consortium for the study of terrorism and Responses of Terrorism (START). To achieve the goal of this research; two different approaches are implemented to handle the missing data namely; Mode-Imputation, and Litwise-Deletion as well as provide a detailed comparative study of the used classification algorithms by using WEKA software and evaluate the obtained results via two different test options which are**;** evaluation on test split of the input data set into 66% for the training data and 34% for the test set, the other option is 10 fold cross-validation during the experiments.

Five main classification algorithms are used in our study, those classification algorithms are: Naïve Bayes, K-Nearest Neighbour, Tree Induction C4.5, Iterative Dichotomiser, and Support Vector Machine. Those classification algorithms in are compared and evaluated according to four performance measures namely; classification accuracy, precision, recall, and F-measure.

The experiment conducted during the mode-imputation approach, in case of test split of the input data with splits 66% for training data, and 34% for testing data showed that SVM is more accurate than other classifiers especially NB, and KNN, the overall performance of NB and KNN is almost the same. ID3 has the lowest accuracy, but it performs well in other measures. In 10 fold cross validation case; KNN classifier is near from the SVM accuracy, precision, and F-measure. ID3 classifier is not accurate, NB classifier performs as KNN in most measures, and C4.5 performs badly than other classifiers in precision, recall, and f-measure.

The experiment conducted during Litwise deletion approach, in case of test split showed that KNN is out performed the other classifiers in its accuracy especially SVM that proved success in the mode imputation approach. C4.5 has lowest precision, recall, and F-measure results. KNN and SVM perform almost the same in precision, recall, and F-measure as they perform effectively in the first approach. In

10 fold cross validation case; SVM is more accurate than other classifiers. KNN, and SVM are almost the same in their results, NB precision, recall, and F-measures are very near from KNN c1assifier. C4.5 has the lowest precision, recall, and F-measure in contrast with ID3 which has highest results in precision, recall, and F-measure although it is not accurate.

VI. Future Work

For future research, there is a plan to further combine the used classification algorithms with genetic algorithms, and neural networks to improve the performance of classifiers, or make hybridization between different classifiers. Another direction for advanced research is to make a hybridization of SVM with one of the heuristic algorithms and evaluate their prediction performance.

Some researchers could perform a modification of this research by using different methods for handling missing data instances, and make a comparison. Others could use different test options to test the performance of the classification algorithms.

References

[1] I. Rizwan, A. Masrah, A. M. Aida, H. Payam, and K. Nasim, "An experimental study of classification algorithms for crime prediction," *Indian Journal of Science and Technology*, vol.6, March 2013.

[2] T. A. Tulips and R. Kumudha, "A survey on classification and rule extraction techniques for datamining," *IOSR Journal of computer Engineering (IOSR-JCE)*, vol. 8, Jan.-Feb. 2013.

[3] R. Kalpana and K. L. Bansal, "A comparative study of data mining tools," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 4, 2014.

[4] S. S. Prasad, M. Sonali, and S. Sonali, "Border security up gradation using data mining," *International Journal of Soft Computing and Engineering*, vol. 4, March 2014.

[5] O. Osemengbe and P. S. O. Uddin, "Data mining: An active solution for crime investigation," *IJCST*, vol. 5, 2014.

[6] The fight against terrorism: An application area with plenty of scope, *The Indian Magazine*, October 2014.

[7] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed., Morgan Kaufmann, 2011.

[8] H. Jantan, A. R. Hamdan and Z. A. Othman, "Classification for talent management using decision tree induction techniques," 2009.

[9] G. Faryral, B. H. Wasi, and Q. Usman, "Terrorist group prediction using data classification," presented at the International Conferences of Artificial Intelligence and Pattern Recognition, Malaysia, 2014.

[10] S. Ozekes and O. Osman, "Classification and prediction in data mining with neural networks," *Journal of Electrical and Electronics Engineering*, vol. 3, no. 1, pp. 707-712, 2003.

[11] D. Hongbo, "Data mining techniques and applications: An introduction," *Cenage Learning EMEA*, 2010.

[12] T. M. Mitchel, *Machine Learning*, McGraw-Hill Science/Engineering/Math, March 1, 1997.

[13] V. Batchu and D. J. Aravindhar, "A classification based dependent approach for suppressing data," *IJCA Proceedings on Wireless Information Networks & Business Information Systems (WINBIS 2012)*, 2011.

[14] S. Neelamegam and E. Ramaraj, "Classification algorithm in data mining: An overview," *International Journal of P2P Network Trends and Technology (IJPTT)*, vol. 4, p. 369, Sep. 2013.

[15] C. B. Sohini and M. Z. Shaikh, "A comprehensive and relative study of detecting deformed identity crime with different classifier algorithms and multilayer mining algorithm," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 3, 2014.

[16] A. Cufoglu, M. Lohi, and K. Madani. "A comparative study of selected classifiers with classification accuracy in user profiling," 2008.

[17] D. Chen and Z. Liu, "An optimized algorithm of decision tree based on rough sets model," in *Proc. International Conference on Electrical and Control Engineering*, 2010.

[18] A. Sachan and D. Roy," TGPM: Terrorist group prediction model of counter terrorism," *International Journal of Computer Applications*, vol. 44, no. 10, April 2012.

[19] P. H. Pilley and S. S. Sikchi, "Review of group prediction model for counter terrorism using CLOPE algorithm," *International Journal of Advance Research in Computer Science and Management Studies*, vol. 2, issue I, ISSN: 2321-7782, January 2014.

[20] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, San Francisco, CA: Morgan Kaufmann Publishers, 2006.

[21] V. Minakshi and R. Gimpy, "Missing value imputation in multi attribute data set," *International Journal of Computer Science and Information Technology*, vol. 5, no. 4, 2014.

[22] N. K. Petra. (2009). Classification in WEKA. Department of Knowledge Technologies. [Online]. Available: http://www.pdfdrive.net/classification-in-weka-e390376.html

Ghada M. Tolan earned her B.S. in operations research and decision support, at the Faculty of Computers and Information, Cairo University in 2001. Then she received her master degree in operations research in 2007, she is currently a PhD student. Her employment experience includes FCI institution since 2001; she is employed with the Department of Operations Research as a lecturer assistant. Her research interests include modeling and simulation, data mining and soft computing.