Prediction of Terrorist Activities by Using Unsupervised Learning Techniques

Taimoor Hassan[1], Imran Sarwar Bajwa[2], Shoaib Hassan[1]

[1]College of Electrical and Mechanical Engineering, Lahore Garrison University, Lahore, Pakistan, [2]The Islamia University of Bahawalpur, Pakistan

Abstract

Terrorism now considered as a major threat to the world population. Terrorism is increasing day by day by different means. From the last decade terrorism rate is increasing exponentially. But there is no efficient way for prediction of terrorism activities. Our paper focuses on prediction of different terrorist attacks by using data mining techniques. In this paper we proposed prediction of attacks by using unsupervised clustering technique. We proposed a framework in which we do sentiments analysis of our data and then by using a combination of density based clustering and distance based clustering we assign classes to our data. Class labels help us to predict terrorism attacks. By research we come to know that combination of these two clustering techniques give accurate results. This proposed framework gives high level of accuracy and it is useful in prediction of attacks types. It gives us a way to deal with terrorism attacks in advance and makes our society peaceful.

Key words: Unsupervised learning; Distance Based Clustering; Density Based Clustering; Sentiments analysis

Corresponding author's email: *taimoorhassan9@yahoo.com*

INTRODUCTION

Terrorism means any activity that affects a common man. Terrorism includes hijacking, kidnapping, bomb blasting, murder and many other activities that create fear among society. Terrorist activities are increasing in overall world and terrorism rate is also increasing from last decade. Now a day terrorists are doing terrorism activities using different communication channels like through messaging or any social media. Detection of their dangerous activities is very necessary. Security departments should do overcome this threat. Security agencies should use some techniques for prediction of terrorist attacks types so that they can handle it in advance.

Data mining is a technique for extracting the knowledge from the data (Kriegel et al., 2011). It is a technique of analyzing the data and extract useful information from it. We use data mining for prediction of any information. We use different types of algorithms in data mining. These algorithms include clustering, classification, decision tree, adaboost etc. These algorithms can be used for extracting useful information from a bulk of data. But clustering is the most useful technique for predicting information. Clustering is unsupervised learning approach (Angelov et al., 2016). Unsupervised learning is a machine learning approach in which we have no training data. In unsupervised learning we have just a bulk of data which has no labels and we assign labels to data set points by using clustering technique (Bing Liu, 2010). Clustering (Glissman et al., 2011) gives labels to all data points and provides us information about classes of all data points. There are many types of clustering including distance based clustering, density based clustering, K-means clustering, K-medoids clustering etc. (http://www.start.umd.edu). But the most simplest and efficient clustering techniques are distance based and density based clustering. In this paper we have large database so a combination of distance based and density based clustering is useful for maintaining and proceeding large database (Angelov et al., 2016).

- Distance Based Clustering
- Density Based Clustering

Distance based clustering deals with distances of data points from the center of cluster. K means clustering is distance based clustering (Kriegel et al., 2011). In K means clustering center may or may not be a data point. In distance based clustering firstly we assume number of clusters which we want to make. After assuming clusters we assume center points of both clusters. Then we find the distance of every data point from these cluster points by using distance formulae (Godara and Yadav, 2013). After finding distances we check which data points have minimum distance from cluster center points. Data points that have minimum distance from any cluster

will assign to that cluster. After assigning all data points to some cluster then we again check mean value of newly form cluster. If new clusters have same mean values as previous clusters mean values then it means that we have made clusters successfully. If mean values are not same then we repeat the steps again and again until mean values of current and previous clusters become same. Distance based clustering method is outlier sensitive. Noise has also effect on distance based clustering. This type of clustering does not terminate at global optimum. It only terminates at local optimum. Implementation of this type of clustering is easy and simple. Density based clustering depends upon density of data points in a data set. In density based clustering we have two parameters. One parameter is radius of center point of data set and second parameter is minimum number of points that should come in the range of the radius. Density based clustering technique makes random shape of clusters. It has also ability of noise handling. Density based clustering depends upon total density connected points. Outliers can easily be detected by using distance based clustering. DBSCAN is also density based clustering (Godara and Yadav, 2013). In density based clustering clusters are equal to total denser areas that lie in data set. It is also efficient for detecting outliers.

Sentiments analysis is a technique of textual mining. It is also called opinion mining. Sentiments analysis determines human behavior or attitude towards any subject or topic. Sentiments analysis is mostly used in social media for doing textual mining of data and extract useful information from it. As social media communication is increasing day by day so sentiments analysis is very useful in determining views of people related business, politics or any other thing. Sentiments analysis is very important for business success (Bing Liu, 2010; Glissman et al., 2011). Because it provides us complete information about market trends and market views about any business product. Now a day researchers are mostly using sentiments analysis technique for detecting human emotions. Sentiments analysis deals with natural language processing so it faces a lot of problems. In sentiments analysis we analyze all the data and determines positive and negative behavior in data. After analyzing positive and negative behavior we assign classes to them. This analysis gives us information that how many people have negative behavior about some activity and how many people have positive behavior about that activity (Nanda et al., 2004). We can use sentiments analysis both in supervised and in unsupervised learning. In unsupervised learning we analyze all the data and extract some defined phrases or words that are relevant to the behavior which we want to study. Those data points that have specific phrases or words will be in one class and those data points in which there is absence of specific words will be in second class. After analyzing we determine probabilities of occurrences of specific words and phrases in all data points (Bajwa et al., 2012). Then we also find the average of number of occurrences of these specific words and absences of these specific words. This average value provides us that either our data set has most positive behavior or most negative behavior (Mai et al., 2016).

MATERIALS AND METHODS

In this paper we proposed unsupervised clustering technique for predicting terrorist attacks. The proposed framework (Figure 1) performs sentiments analysis on our data for selecting specific words or phrases. After selection of specific words we use a combination of distance based and density based clustering technique (Bing Liu, 2010; Glissman et al., 2011) and make clusters of data points that represent some terrorist activity and those that do not any represent terrorist activity (Godara and Yadav, 2013; Kim et al., 2014). After making clusters we test our clusters by taking some test data. Our proposed framework determines the cluster of test data point. Our proposed framework consists of these following steps:

1. Data Collection
2. Data Preprocessing
3. Sentiments Analysis
4. Apply Clustering Techniques
5. Deployment and Testing the Framework

Data Collection

Data collection is first step of any data mining technique. Data collection means gathering the data from any source. There are many data bases that are publically available. We can use any publically database or global database for our proposed framework. We use global terrorism database in which there is record of all terrorist activities that occurred since 1970. There are more than 100000 cases in the global terrorism database. It includes all types of activities like bombing, kidnapping, street crime, murder, hijacking etc. This global database is necessary for sentiments analysis because it provides us the specific words or phrases that are common in any terrorist activity.

Data Preprocessing

Data preprocessing is a necessary step for doing any data mining algorithm. We said it preprocessing step because this step should be done before

implementing any data mining algorithm. Data preprocessing means data cleaning. When we collect data from any database then it is not completely clean. There are many factors like noise (outliers), inconsistency, incompleteness that make data dirty. We remove all these factors and make our data clean. Preprocessing step makes implementation of data mining algorithm easy. We can easily understand the data after preprocessing step. Data preprocessing step improves the performance of our whole proposed framework. There are many other steps that come in data preprocessing steps like data integration, data reduction and data reduction etc.

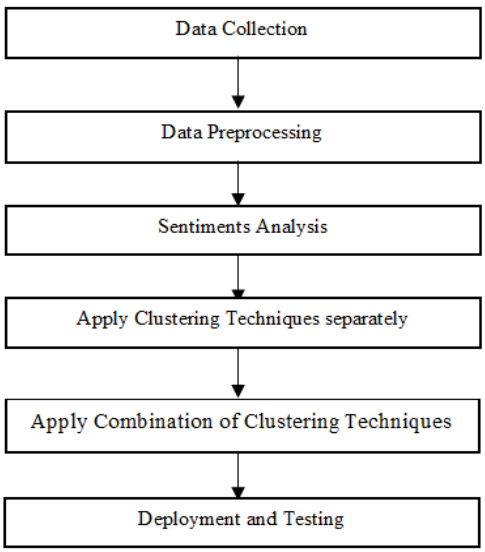

Figure 1: Proposed Framework

Sentiments Analysis and Attributes Selection

Sentiments analysis means human behaviors against any subject. Here our subject is terrorism so we do sentiments analysis of our database and check the behavior of people. We do sentiments analysis on our whole database and determine which specific words and phrases are mostly used in terrorist activities. Each terrorist message has some words or phrase that show his terrorist activity. The main purpose of sentiments analysis is to select all the words that represent terrorist activity and take next decisions on the basis of these words. These specific words may be bomb blast, kidnap, murder, hijacking, crime and many other that represent some terrorist activity. After sentiments analysis we get a set of words that represent terrorist activities. This set of words is helpful for future steps of clustering. On the basis of these specific words it is easy to make two clusters of our dataset. One cluster contains messages of all criminals and other cluster contains normal messages. So sentiments analysis is necessary for executing our proposed framework successfully. After sentiments analysis we select attributes of data points. Firstly we select 20 attributes. But after reduction process we select only 6 attributes of data that are most important and have high impact on data set. Words that use in messages are major attributes for clustering the data set. Because words reflect the behavior of a person so any criminal activity can easily be predicted by words. Selected attributes are given below:

- Words of messages (Show positive and negative behavior of people)
- Message owner's country
- Area of the country
- Message length
- Message time
- Message owner's name

Apply Clustering Algorithms Separately

In our proposed framework we suggest two clustering techniques that are distance based and density based clustering. Firstly we apply these two algorithms separately and finding their accuracies. Then we apply the combination of these two algorithms for achieving better accuracy. When we apply distance based clustering on our dataset then assumes two clusters in first step. We assume that one cluster is terrorist activity cluster and other cluster is terrorist activity free cluster. We assume centers of these two clusters. After assuming center points we find distances of all data points from these two center points by some distance formulae. Data points that have terrorist activity like words have less distance as compared to those data points that have no terrorist words. We check other attributes values also. After finding all the distances we assign all data points to some cluster. As each data point has two distance values so we assign data points to a cluster from which they have minimum distance value. After assigning all data points to some cluster we gain two clusters. Then we repeat this whole process again and again until we have done clustering successfully. After clustering we see there are two clusters in which one cluster has complete record of all terrorist messages and other cluster has all those messages that are normal. So distance based clustering is good for clustering the data.

Second technique which we apply on our data set is density based clustering. It is simplest type of clustering. It deals with denser areas of data set. All density connected points are in one cluster in density based clustering technique. When we apply

this clustering technique on our data set then firstly we assume our center points and define minimum connected points and radius of that center points. One center point is terrorist activity message data point and other center point is normal message data point. Then we check how many data points are in the radius range of our center points. If data points are equal or greater then defined minimum points then it will form a cluster otherwise cluster will not form. We see that the data points that have same attributes values as terrorist activity data point have then these points are in the range of radius of terrorism activity center data point. And we assign these points to that cluster. And those data points that have different attribute values are far away from terrorist activity center data point so they are not in the range of radius of terrorism activity center data point. These points are in the range of normal message center data point so we assign normal center data point cluster to those values. We repeat this process until we get two denser areas that are separable from one another. One area is representing terrorist activities messages cluster and other is representing normal messages cluster.

After applying both clustering separately we apply combination of these two clustering techniques on our data set. Combination of these two clustering techniques gives us more accurate results as compared to implementing them separately. We choose a combination of distance based and density based clustering because they are simple, easy to understand and gives more accuracy then using density and density based clustering separately. The main reason of using combination of these two algorithms is that it can easily handle large database (Angelov et al., 2016). Distance based clustering gives accurate clustering but it is sensitive to outliers. Because distance based algorithm is based on local optimum. We resolve this problem by applying combination of distance based and density based algorithm. Distance based clustering gives us accurate results and density based algorithm handle noise and outliers efficiently because it is based on global optimum. When we use combination of both techniques then we apply distance based clustering first. Distance based clustering gives us accurate clusters of terrorist activity messages and normal messages. After implementing distance based clustering when we apply density based clustering on our data set then we use those center points that we got from implementation of distance based clustering. We check density connected points through density based clustering. Density based clustering make accurate clusters of terrorist activity data and normal data.

Deployment and Testing

In this paper we proposed a framework that predicts terrorist activity successfully. After proposing we deploy our framework and test either our proposed framework is working properly or not. For testing purpose we take a test sample and assign label to it. We assign label to test sample according to its attribute values. We check similarity between attribute values of test sample with our proposed clusters. Test sample assign to that cluster from which it has minimum distance and same attribute values. Accurate assignment of test sample gives us accuracy of our proposed framework.

RESULTS

Our proposed framework gives us accurate results of clustering. Our proposed framework improves the accuracy level of prediction of terrorist activities. It separate terrorist activity messages from normal messages. For showing results of our proposed framework we can use Rapid Miner 5.3 or Matlab. After implementing it we achieve two classes. Label of one class is terrorist class and other is normal class. Our proposed framework gives three results. These results are given below.

1. Results of implementing distance based clustering
2. Results of implementing density based clustering
3. Results of implementing combination of distance based and distance based clustering (Gives high accuracy and resolve problem of outliers)

Distance based clustering gives 46 % accuracy while density based clustering gives 47% accuracy (Kim et al., 2014). Density based clustering gives more accuracy then distance based clustering. When we use combination of density based and distance based clustering then we increase accuracy up to 50%. We use combination of these two algorithms because density based clustering gives best result for large database and distance based clustering takes less time for making clusters(Kim et al., 2014; Nanda et al., 2014). Furthermore density based clustering is more robust then distance based clustering (Bajwa et al., 2012). Our proposed framework improves accuracy of clustering the big data.

CONCLUSION

Terrorist attacks prediction by using unsupervised learning gives us a new path for extracting information from big data. Our proposed framework combines two main clustering techniques and gives best results. Data mining is a popular field widely

used in biomedical and for security purposes. Our unsupervised clustering based model makes proper discrimination between normal messages and terrorist messages. It scans all the items of large data set and determines terrorist attacks. Our proposed model also gives comfort to security agencies and helps them to deal with uncertain conditions in advance. It increases trust of people on security agencies.

REFERENCES

- Kriegel HP, Kröger P, Sander J and Zimek A. (2011). Density based clustering. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1(3):231-240.

- Angelov PP, Gu X, Gutierrez G, Iglesias JA and Sanchis A. (2016). Autonomous data density based clustering method. In The bi-annual IEEE World Congress on Computational Intelligence (IEEE WCCI).

- Bing Liu. (2010). Sentiment Analysis and Subjectivity, appear in Handbook of Natural Language Processing, Second Edition, (editors: N. Indurkhya and F. J. Damerau),

- Glissman S, Almaden, San Jose, Terrizzano I. Lelescu A, Sanz J. Systematic Web Data Mining with Business Architecture to Enhance Business Assessment Services, Annual SRII Global Conference (SRII), 2011

- Godara S, Yadav R. Performance Analyis of Clustering Algorithms for Character Recognition Using WEKA Tool, International Journal of Advanced Computer and Mathematical Sciences

- Kim Y, Shim K, Kim MS and Lee JS. (2014). DBCURE-MR: an efficient density-based clustering algorithm for large data using MapReduce. Information Systems 42:15-35.

- Nanda SJ, Raman R, Vijay S and Bhardwaj A. (2014). A new density based clustering algorithm for Binary Data sets. In High Performance Computing and Applications (ICHPCA), 2014 International Conference on (pp. 1-6). IEEE.

- Bajwa IS, Lee M and Bordbar B. (2012). Resolving syntactic ambiguities in natural language specification of constraints. In Computational Linguistics and Intelligent Text Processing. Springer Berlin Heidelberg 7181:178-187

- Mai ST, Assent I and Le A. (2016). Anytime OPTICS: An Efficient Approach for Hierarchical Density-Based Clustering. In Database Systems for Advanced Applications. Springer International Publishing 9642:164-179.