Terror Attack Identifier: Classify using KNN, SVM, Random Forest algorithm and alert through messages

Abhishek Barve[1], Manali Rahate[2], Ayesha Gaikwad[3], Priyanka Patil

[1]Assistant Professor, Vidyalankar Institute of Technology, Mumbai, India
[2,3,4] Students, Vidyalankar Institute of Technology, Mumbai, India

---***---

Abstract - *The system to prevent terrorist attacks that will relay emergency alerts at all phones is set to begin .This system could warn people of terrorist strikes by text messages by broadcasting it to all the people in the nearby location. With the popularity of social networks , mostly news providers used to split their news in various social networking sites and web blogs.*

Machine learning techniques will be used to train the data .In order to create the instances words from each short message were consider and bag-of-words approach was used to create feature vector .The data was trained using KNN(K-Nearest Neighbor), Support vector machine, Random forest machine learning techniques.

***Key Words*: Data Mining, Twitter, news, text analysis, terrorist attack, tweets.**

1. INTRODUCTION

Now-a-days in India, there are many news groups who share their news headlines as short messages in micro blogging services such as Twitter. Authors of these messages write about their life, share opinions on variety of topics and discuss current issues. Because of a free format of messages and an easy accessibility of micro blogging platforms, Internet users tend to shift from traditional communication tools (such as traditional blogs or mailing lists) to micro blogging services.

As more and more users post about products and services they use, or express their political and religious views, micro blogging web-sites become valuable sources of people's opinions and sentiments. We use a dataset formed of collected messages from Twitter. Twitter contains a very large number of very short messages created by the users of this micro blogging platform. The contents of the messages vary from personal thoughts to public statements. The short messages will be classified by the system into a group: war-terrorist-crime.

1.1 Objectives

To develop a system that will extract the live tweets from twitter site, will classify those tweets and display the news under its section that will help news seeker to keep track of news. For development of the proper system a perfect classifier has to be selected that can be done by comparing different classifier result on tweets provided.

1.2 Scope

Scope of this dissertation is to develop a system that will collect short messages from twitter social networking site. The collected twitter messages are used to train by using SVM, Random Forest and KNN data mining techniques and a classifier is built that will classify the messages (e.g. war-terrorist).The performance of each classification techniques is calculated that will be the effectiveness of the system. Thus precision and recall values are calculated to measure the performance of each classifier system. F_β was calculated to obtain a single value measurement. The results generated from all 3 classifiers is compared in order to find the classifier that provides high performance for most groups will be consider as the best classifier for classifying the messages extracted from twitter, so that users or analyst in specific field able to know about the news

1.3 Proposed system

We are using K-Nearest Neighbour data mining method for classifying twitter message into new group. This Chapter deals with the study which involves detail knowledge of twitter, Web Mining, data gathering techniques for tweets extraction, feature selection technique and detail of classification algorithms used for extraction.

2. Implementation

This shows how the system is implemented. For this first module extract the tweets from the trusted news channel that is the input for the system. The output module gives the result in the form of tweets classified in news group: war-terrorist-crime, economy business, health, sports development-government, politics, accident, entertainment, disaster-climate, education, society and international. For the classification KNN, SVM and Random Forest are used, twits are classified and analysis in done on the result drawn from all three algorithms is shown in order to find the best classifier for the twit's classification.

Data Gathering

The classification will be applied into the short messages-news of Twitter micro blog. Thus, twitter short messages are needed to be collected. Twitter API provides the ability of retrieving such short messages for a given user in XML file format. Each XML file could carry out 200 short messages at once.

Pre Processing

Once gathered the data, the features are need to extract from the short messages. The words are used as features. Thus the bag-of-words approach was used to extract the features. This will pool the words from all short messages and will create a document vector, containing words

Data Classification

Document Vector containing the tweets are passed as an input to KNN classifier where news will be classify into twelve groups. The training process was developed in order to recognize whether the selected short message belong to the group A, short messages will be classified manually as "Group A" or "other". 90 % data was used to train the system and 10 % were used to test the system.

Current Tweets And Classification

As we all know that news should be current if it is old then it is of no use taking that into consideration we will fetch the live tweets from Twitter website and then pre-processing will be done so that we can classify it into twelve news group.

Displaying The News Through Broadcast message

Classified news from classifier can be displayed through broadcast message so that the news seeker can directly assess the latest news from anywhere and news is classified so that news seeker can directly access the category.

1.K-Nearest Neighbour Classifier:

One of the various classifiers, 'KNN classifier' is a case based learning algorithm which is based on a distance or similarity function for various pairs of observation such as the Euclidean distance function. It is tried for many applications because of its effectiveness, non-parametric & easy to implementation properties. However, under this method, the classification time is very long & it is difficult to find optimal value of K. Generally, the best alternative of k to be chosen depends on the data. Also, the effect of noise on the classification is reduced by the larger values of k but make boundaries between classes less distinct. By using various heuristic techniques, a good 'k' can be selected. In order to overcome the above said drawback, modify traditional KNN with different K values for different classes rather than fixed value for all classes.

KNN algorithm is used to classify instances based on nearest training examples in the frame space. KNN algorithm is known as lazy learning algorithm in which function is approximated locally & computations are delayed until classification. A majority of instances is used for classification process. Object is classified into the particular class which has maximum number of nearest instances.

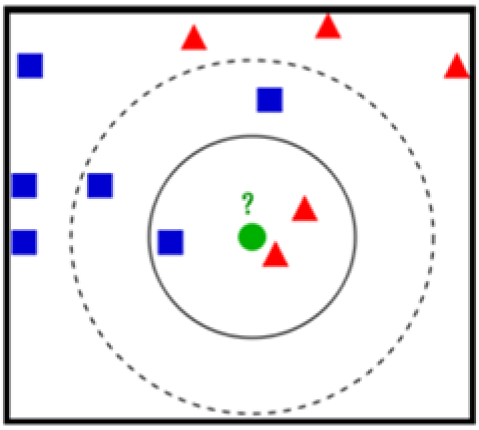

Fig -1: k Nearest neighbor

The test instance (green circle) should be classified either into blue square class or into red triangle class.

If k = 3 (solid line circle) test object (green circle) is classified into red triangle class because there are 2 triangle instances and only 1 square instance in the inner circle.

If k = 5 (dashed line circle) test object (green circle) is classified into blue square class because there are 3 blue square instances and only 2 red triangle instances in the inner circle.

Computation of distance function in KNN is based on distance between input test instance & training set instances. To compute the distance between instances, the distance/similarity function is important.

Euclidean distance: This is distance/similarity function is also called as the “Pythagorean theorem‟.

2.Support Vector Machines:

SVM tends to deal with high dimensional data sets. When the data were far from linear and the datasets are inseparable, Kernels are used to map the data into a high-dimensional space, where the new mapping is then linearly separable . SVM creates a hyper plane between data groups. It creates the hyper plane by maximizing the margin as given in figure 2. Margin is the distance from the hyper plane to the closest data points. SVM do not over generalize the problem as it minimizes both the error and the complexity as given in figure . SVM do not address to the local minimum of the error rate. This caused to increase the accuracy of SVM .

3.Random Forest :

A random forest adds an additional degree of randomness to bagging. Although each tree is constructed using a different bootstrap sample of the dataset, the method by which the classification trees ate built is improved. A random forest predictor is an ensemble of individual classification tree predictors. For each observation, each individual tree votes for one class and the forest predicts the class that has the plurality of votes. The user has to specify the number of randomly selected variables (mtry) to be searched through for the best split at each node.

Whilst a node is split using the best split among all variables in standard trees, in a random forest the node is split using the best among a subset of predictors randomly chosen at that node. The largest tree possible is grown and is not pruned. The root node of each tree in the forest contains a bootstrap sample from the original data as the training set. The observations that are not in the training set, are referred to as "out-of-bag" observations.

Since an individual tree is unpruned, the terminal nodes can contain only a small number of observations. The training data are run down each tree. If observations i and j both end up in the same terminal node, the similarity between i and j is increased by one. At the end of the forest construction, the similarities are symmetrized and divided by the number of trees. The similarity between an observation and itself is set to one. The similarities between objects form a matrix which is symmetric, and each entry lies in the unit interval [0,1]. Breiman defines the random forest as :

A random forest is a classifier consisting of action of tree-structured classifiers {h (x, e k), k = 1, ...} where {e k} are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x.

A summary of the random forest algorithm for classification is given below :

-Draw ntree bootstrap samples from the original data.

-For each of the bootstrap samples, grow an unpruned classification tree, with the following modification: at each node, rather than choosing the best split among all predictors, randomly sample mtry of the predictors and choose the best split from among those variables. Bagging can be thought of as the special case of the random forest obtained when mtry = p, the number of predictors.

-Predict new data by aggregating the predictions of the ntree trees, i.e., majority votes for classification, average for regression.

-The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. Using a random selection of features to split each node yields error rates that compare to AdaBoost . An estimate of the error rate can be obtained, based on the training data, by the following :

-At each bootstrap iteration, predict the data that is not in the bootstrap sample, called "out-of-bag" data, using the tree which is grown with the bootstrap sample.

-Aggregate the out-of-bag predictions. On the average, each data point would be out-of-bag around 36% of the times, so aggregate these predictions. Calculate the error rate, and call it the "out-of-bag" estimate of error rate.

-The random forest performs well compared to several other popular classifiers. In addition, it is user friendly as it has only two parameters: (i) the number of variables in the random subset at each node, and (ii) the number of trees in the forest. The random forest is not usually very sensitive to the values of these parameters.

-Hardware And Software Requirements

-Hardware Requirement

Processor: Dual core processor

RAM: 2 GB of RAM.

Hard Disk: 40 GB +

Standard PS/2 Keyboard and Mouse.

Colour Monitor.

-Software Requirement

Operating System: Windows 7.

Web Browser: Mozilla Firefox 45.0.2.

Programming Language: Java

3. CONCLUSIONS

A system which is able to classify news headlines will be useful in various social researches. With the development of web technologies, people get involved in many social networks and web blogs. Twitter is a micro blog which allows many famous news suppliers to publish their news headlines. Twitter API supports user to retrieve available short messages. These retrieved files will be in XML file format and each file could retrieve maximum number of 200 short messages per once.

In order to apply machine learning, a proper feature set was required. The feature set was created by pooling the words and creating a document vector. This approach was named as bag-of-words approach. The frequency of each word was chosen as data. When considering all words together, it create huge dimension with 3569 instances. In order to reduce the dimension, a lover cut off frequency and an upper cut off frequency value was chosen. According to the current system, these values are respectively 10 and 38. The value was chosen as the frequency range which maximizes the accuracy. This caused to reduce the dimension up to 126.

There are 2 groups defined and each group was treated as separate binary classification problem as same short message could be belong into several groups. System was trained using KNN, SVM and Random Forest. The effectiveness of the training system can be measure using recall and precision values. Precision is the probability of retrieving relevant short messages. Recall is the probability of the relevancy of retrieved short messages. The harmonic measure (F-measure) was used to obtain a single value for recall and precision. The weighted F -measure (Fβ measure) was used as precision was needed to be emphasizing in current situation. The system provides best results for Accident, Development-Government, Climate-Disaster, Entertainment, Health, Education, Sports, War-Terrorist-Crime, Politics and Economy-business groups. KNN provides accuracy of 100% compared with SVM and Random Forest which provides the accuracy of 74.20 and 74.69 respectively

REFERENCES

[1] Inoshika Dilrukshi, Kasun De Zoysa, Amitha Caldera. "Twitter News Classification Using SVM", Computer Science & Education (ICCSE 2013 IEEE).

[2] Juan DU, Zhi an Yi. "A New KNN Categorization Algorithm for Harmful Information Filtering", 2012 IEEE

[3] Mateusz Budnik, Iwona Pozniak-Koszalka, Leszek Koszalka, "The Usage of the k-Nearest Neighbour Classifier with Classifier Ensemble", 12th International Conference on Computational Science and Its Applications, 2012 IEEE

[4] Mohammad Abdul Wajeed, T. Adilakshmi,"Semi-Supervised Text Classification Using Enhanced KNN".2011 IEEE

[5] Lijun Wang, Xiqing Zhao," Improved Knn Classification Algorithms Research In Text Categorization". 2012 IEEE

[6] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon," What is Twitter, a Social Network or a News Media?". International World Wide Web Conference Committee (IW3C2) , April 26–30, 2010

[7] https://dev.twitter.com/oauth

[8] Jaideep Srivastava, Prasanna Desikan, Vipin Kumar," Web Mining - Concepts, pp 400 – 402.

[9] S.Vidya, K.Banumathy, "Web Mining- Concepts and Application", International Journal of Computer Science and Information Technologies, Vol. 6 (4), 2015, 3266-3268.

[10] G.Kesavaraj , Dr.S.Sukumaran, "A Study On Classification Techniques in Data Mining", IEEE 4th ICCCNT - 2013 July 4 - 6, 2013.