IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Volume-6, Issue-6 E-ISSN: 2347-2693

Performance Evaluation of Fuzzy C Mean Clustering on Social Media Data Set

Kothapalli Revathi[1] , Chalumuri Avinash[2]

[1,2] Dept. of CSE, Gayatri Vidya Parishad College of Engineering, Madhurawada, Komaddi, India

Available online at: www.ijcseonline.org

Accepted: 21/Jun/2018, Published: 30/Jun/2018

***Abstract*-** As we all know that in recent days social media play's a very prominent role for sharing human social behaviors and participation of multi users in the network. This social media greatly increased the user's interest in posting various updates of them which happened in and around the world. This social media will also create a facility to study and analyze the general behavior of human to process the large stream of data which is available on the social media database. Till now there are several primitive algorithms that are available in the literature regarding the clustering of user's interest on social media but they failed to achieve in reducing the time complexity. In this proposed application we for the first time have designed a novel fuzzy c means clustering algorithm for grouping related information of users. By implementing this proposed algorithm and comparing with several primitive algorithms, we can get best group result and also reduce error rate for generating cluster groups.

Keywords— Clustering, Social Media, Fuzzy C Means, Grouping Messages, Time Complexity.

I. INTRODUCTION

Now a day's social media plays a vital role in accessing to lot of users for a group of internet applications. In this social media the user creates a separate and individual expressions for the data exchange and this remains a unique mode of data influence[1].In recent study we came to know that a lot of individuals like industry specialists try to collect a lot of data for the analysis and behaviour calculation based on their extracted topic. For example twitter and face book occupies the most important position in the list of several social media site [3].Along with these social media sites, slashdot is one form of website that is used to focus mainly on the technology news where the stories can be written by the posted users or several editors and a lot more. The main difference between slashdot and other sites is the ability for the user to rate the community of tweets or messages into two types based upon a positive or negative rating[4]-[5].

As we all know that the exchange of data almost done in a digital manner all the time so it is a great time for the platforms to remain in a normal condition. As the sites are increasing its communication and interactions ,there is a lot of time delay for finding the exact information from these social media sites [6] [7]. In general for clustering of messages from these social media sites, literature experts try to use k-Means algorithm, in which the cluster centre is initially found and from that location all the other related messages used to be clustered which is a time taking process .And at the end of the clustering process, we come an conclusion that clustering process may take more time and the result may not be efficient. Also there are some more limitations like space complexity by using the method like random centroid selection. These random centroids are very poor generation process to form the cluster groups and which remains wastage of time for processing.

II. BACKGROUND WORK

In this proposed section we mainly discuss about the background work required for finding the performance of a social media to create an opportunity to study human social behavior to analyze large amount of data streams. Here we try to propose a fuzzy c means clustering algorithm for grouping related information and in this section we mainly discuss about the text mining and its various models for clustering and classification of data.

Main Motivation

Data mining is the process or method of extracting the valuable data from a large data source and it acquires information from a multiple disciplines like artificial networks, pattern analysis, pattern identification, ML Approach and so on. Generally the data mining process mainly involves a very keen observance to finish various tasks with a different set of algorithms. The several algorithms attempt to fit a analysis into two separate models like predictive and descriptive manner and they were assigned with a different parameters to achieve the required

output. The data mining algorithms are mainly classified into various types based on the observation of three things. Now let us discuss about those three things in detail as follows:

1. **Search Based Algorithm:** This is one of the type which mainly used to find out the search techniques to get the desired result.
2. **Preference Based:** This method is mainly used by the end users to achieve the preferences what they expect.
3. **Model Based:** This is the third model in which the clustering is done based on the model that suits the desired results.

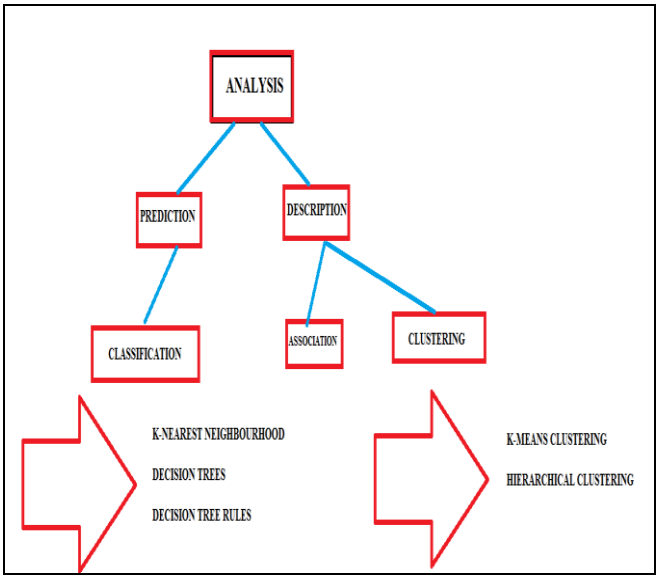


Fig 1. Demonstrates the various Data Mining Analysis Models

From the above figure 1, we can clearly get an idea that the data mining models are mainly classified into two types like prediction based and description based. In prediction based classification method comes with various algorithms that are used for classification of data, which was clearly shown in above figure. Now the other side of analysis, we have description model, where this model has two subparts like association rule mining and clustering as main methods.

The main models of data mining are as follows:

1. Predictive Based Model
2. Descriptive Based Model

1. Predictive Model

This model mainly designed in order to make a prediction about the values of the data using known results found from different data. Predictive data modeling may be made based on the use of historical data. It includes Classification, Regression, Time Series Analysis and Prediction [8]-[10].

2. Descriptive Model

This model mainly identifies patterns or relationships in data. Unlike the Predictive model a descriptive model [11] serves a way to explore the properties of data examined, not to predict new properties. It includes Clustering, Summarization, Association Rules and sequence discovery.

III. THE PROPOSED COMPARITIVE ANALYSIS OF K-MEANS AND FUZZY C-MEAN CLUSTERING ALGORITHM

In this section we mainly try to discuss and study the two algorithms like k-means and fuzzy c-mean clustering to perform clustering on twitter dataset using both the algorithms and calculate SSE(sum of square error).

1) K-MEANS ALGORITHM ON TWITTER DATA SET

Initially we try to discuss about the K-means algorithm by taking twitter data set into an account and try to cluster all the positive and negative data that is posted on the twitter by various users. Now let us discuss about this K-means algorithm in detail:

ALGORITHMIC PROCEDURE

Initially we try to take the twitter data set as input and then start the procedure.

Step 1: Read the twitter data set from the twitter server.

Step 2 : Now enter the number of clusters to be performed and try to choose a centroid randomly from that twitter dataset.

Step 3: Now try to find out each data point (d_i) from dataset and calculate the Manhattan distance from data point to centroids' (c_i).

Step 4: Now calculate the distance for each data point with that centroid.

$$\text{Distance} = (c_i - d_i)$$

Step 5: Now try to find out the very closet distance of each and every centroid from the input data point and form into clusters.

Step 6: Try to repeat the steps 3 and 4 until a new change occur in the centroids.

Step 7: Here once the step 6 is completed ,we will get a very new group of clustered data and which will show the total number of clusters that are formed after the calculation.

Step 8: In this step we try to find the manhattan distance and also try to calculate each cluster sum squared error by using following equation.

$$SSE=\sum_{i=1}^{n} dist\ (c_i, di)$$

2) FUZZY C-MEAN CLUSTERING ALGORITHM ON TWITTER DATA SET

In the above section we discussed about the K-means algorithm on twitter data set and try to find out the sum of square error which occurred by using K-means algorithm. Now in this section we try to find out the clustering based on Fuzzy C-Mean Algorithm on the same twitter data set and try to optimize the time redundancy.

ALGORITHMIC PROCEDURE

Initially we try to take the twitter data set as input and then start the procedure.

Input : A set of k clusters.

Step 1: In the step 1, we try to choose K-Data Objects randomly from a data set and we assume 'D' as the initial clusters.

Step 2: In this step 2,we try to find out the matched words between each data object d_i (1<=i<=n) and also find out each cluster Centre

$$Cj\ (1<=j<=k).$$

Step 3: In this step 3, we try to find out the SSE by using the formula like

$$SSE=1/w^2$$

Step 4: In this stage we will try to find out the total number of words that are available in the data point and we will find out the centroid and try to weight each and every data point and calculate the final twitter weights as follows:

$$Weight\ (W_i) = 1/dist(d_i,c_i)^2\ /\sum_{q=1}^{k} 1/dist(C_i,d_i)$$

Step 5: In this stage we try to find out the weight of each and every data point which is near to the centroids and this can be done once the step 4 is completed.

Step 6: In this stage we compute the distance of each and every data point along with cluster center by taking center

cj (1<=j<=k),

It is computed by taking the weight of data points d (di, cj) and assign the data object di to the nearest cluster.

```
Set cluster[i] = j;
Set w[i] = d (di, cj).
```

Step 7: For each cluster center j (1<=j<=k),

recalculate the centers;

Step 8: This step will be continues until the center is same for the current cluster and previous cluster.

Step 9: This is the last step where the resultant output is formed and the clustering result is displayed.

IV. IMPLEMENTATION PHASE

Implementation is the stage where the theoretical design is converted into programmatically manner. In this stage we will divide the application into several modules and then coded for deployment. We have implemented the proposed concept on java language under JSE platform .Here we used Java Swings and AWT as the front end technologies and back end we took twitter data set from google.The application is divided mainly into following 5 modules. They are as follows:

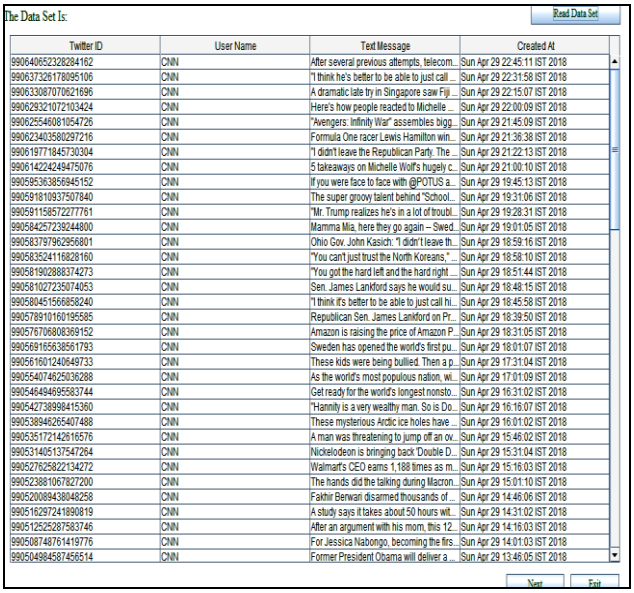
1) Admin Module
2) Load Data Set Module
3) K-Means Clustering Algorithm
4) Fuzzy C-Mean Algorithm
5) Comparative Analysis

1) Admin Module

Here the admin is one who has the valid login credentials in order to login into the application. Once he enter into the application ,he has the following operations like load input data set and try to calculate centroids, generate the clusters using various clustering algorithms and so on.

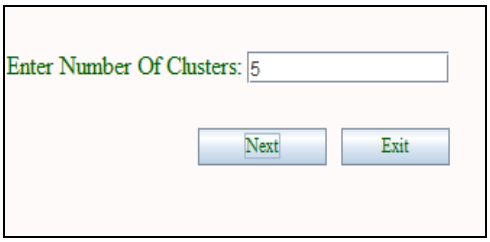
2) Load DataSet Module

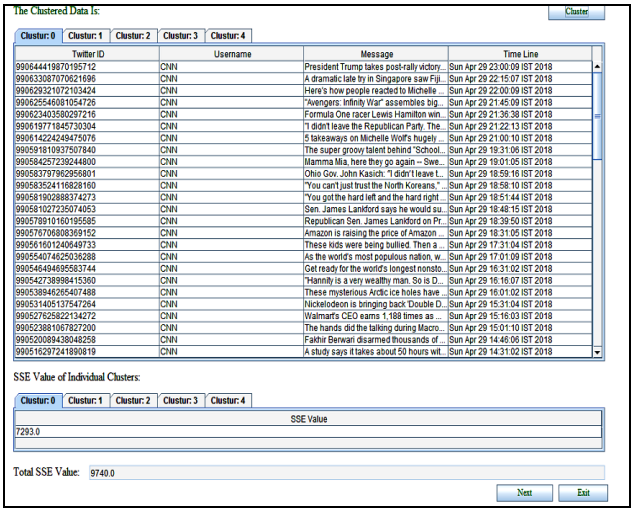
Here the user choose twitter data as the input input data set and once the twitter data set is loaded all the values are inserted into the window. The twitter data set mainly contains the fields like Twitter Id,Username and Text Messages along with created data and time. Here the Text messages are nothing but posts which may contain both positive, negative or neutral topics about certain topic. So here the input data set is loaded with all these basic fields.



3) K-Means Clustering Algorithm

Here in this module we try to find out the clustering of twitter data set by specifying the number of clusters we wish to divide.Initailly the application will ask for number of clusters and once the cluster value is given then the window will application will divide into appropriate number of clusters.As we have given 5 clusters as input for K-Means ,so the algorithm divide the data set into 5 cliusters.

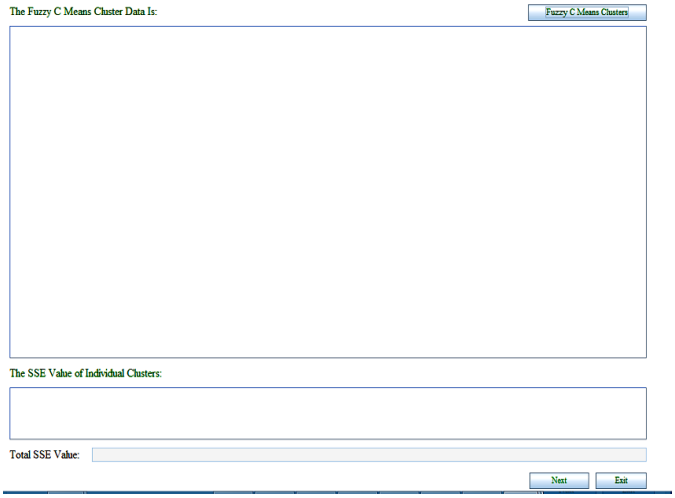




In the above window we can clearly see clusters with numbers from 0,1...4.And the related information is clustered based on the type of messages what the users posted in the data set.

4) Fuzzy C-Mean Clustering Algorithm

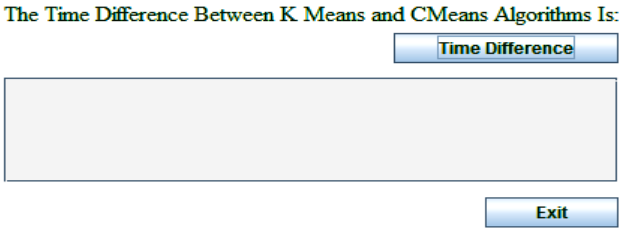
Here in this module we try to find out the clustering of twitter data set by using Fuzzy C mean algorithm and finally we calculate the SSE for the resultant output.



In the above window we can clearly see a button to divide into clusters and also can observe the SSE value for the clustered data.

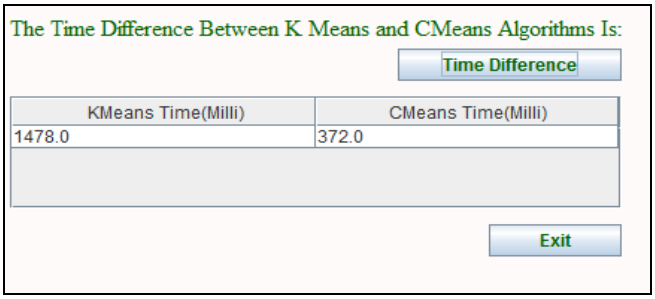
5) Comparative Analysis Module

This is the last module in which we can find out the comparative analysis of two clustering algorithms and we can find out the time difference for both the algorithms.



V. EXPERIMENTAL RESULTS

To show the performance of our proposed algorithm we mainly developed the application in java platform .Here we used JSE as programming interface with Java Swings and AWT as front end design. At the end we try to find out the performance in terms of time optimization between two proposed algorithms as follows"



From the above window we can clearly identify that K-means take 1478.0 Milli seconds of time for clustering the twitter data set and the same data set by using K-Means algorithm took only 372.0 milli seconds.Hence we can show the performance of Fuzzy C-mean is more efficient that compared with primitive K-means algorithm for clustering the data.

VI. CONCLUSION

In this paper, we for the first time have proposed an efficient clustering algorithm for reduce the time complexity and space complexity. This proposed thesis mainly proposed the optimized fuzzy means clustering algorithm for getting better cluster result in data set. By implementing this process we can easily find out similar data object in data set by calculating weight of each data object to centroids. The calculation of weight of data object will repeat until the no changes occur in the centroids. By applying this process we can reduce number of iteration compared to existing algorithm of k means. So that each data point from each cluster center in each iteration due to which running time of algorithm is saved. By implementing proposed system we can efficiently improve speed of the clustering and accuracy by reducing the computational complexity of standard k-means algorithm.

VII. REFERENCES

[1] J´erˆome Kunegis, Andreas Lommatzsch, and Christian Bauckhage. The slashdot zoo: mining a social network with negative edges. In Proceedings of the 18th international conference on World wide web, pages 741–750. ACM, 2009.

[2] Andreas M Kaplan and Michael Haenlein. Users of the world, unite! the challenges and opportunities of social media. Business horizons,53(1):59–68, 2010.

[3] Bogdan Batrinca and Philip C Treleaven. Social media analytics: a survey of techniques, tools and platforms. AI & SOCIETY, 30(1):89–116, 2015.

[4] Claudio Cioffi-Revilla. Computational social science. Wiley Interdisciplinary
Reviews: Computational Statistics, 2(3):259–271, 2010.

[5] David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi.

[6] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. ACM,2010.

[7] Harry Halpin, Valentin Robu, Hana Shepherd The Complex Dynamics of Collaborative Tagging, Proceedings 6th International Conference on the World Wide Web (WWW'07), Banff, Canada, pp. 211-220, ACM Press, 2007.

[8] Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. Everyone's an influencer: quantifying influence on twitter. In Proceedings of the fourth ACM international conference on Web search and data mining, pages 65–74. ACM, 2011.

[9] Valentin Robu, Harry Halpin, Hana Shepherd Emergence of consensus and shared vocabularies in collaborative tagging systems, ACM Transactions on the Web (TWEB), Vol. 3(4), article 14, ACM Press, September 2009.

[10] Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. Psychological Review, 95, 163–182.

[11] Ando, T. (2010), Bayesian Model Selection and Statistical Modeling,

[12] D. Pedreschi, S. Ruggieri, and F. Turini, "Discrimination-Aware Data Mining," Proc. 14th ACM Int'l Conf. Knowledge Discovery and Data Mining (KDD '08), pp. 560-568, 2008.

