Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 4, Issue 11, November -2017

IMPROVED FUZZY K-MEANS CLUSTER ALGORITHM TO ANALYSE WEATHER DATA IN COIMBATORE REGION

[1]B. Murugesakumar, [2]Dr.K.Anandakumar, [3]Dr. A. Bharathi

[1]*Research Scholar, Bharathiar University, Coimbatore, Tamilnadu*
[2]*Assistant Professor(Sl.Grade), Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Sathyamangalam, Tamilnadu*
[3]*Professor, Department of Information Technology, Bannari Amman Institute of Technology, Sathyamangalam, Tamilnadu.*

Abstract:- *Weather analysis has been playing its vital role in meteorology and become one of the most challengeable problems both scientifically and technologically all over the world from the recent years. This research paper carries historical weather data collected locally at South Indian textile city, Coimbatore that was analyzed for useful knowledge by applying data mining methods. Data includes last five years' period [2012-2017]. It had been tried to extract useful practical knowledge of weather data on monthly based historical analysis. This Proposed research work was done using data mining tool called Weka by examining changing patterns of weather parameters which includes maximum temperature, minimum temperature, wind speed and rainfall. After preprocessing of data and outlier analysis, improved fuzzy K-means clustering algorithm and Decision Tree techniques were applied. Two clusters were generated by using improved K-means Clustering algorithm with lowest and highest of mean parameters. The result obtained with smallest error (27%) was selected on test data set. While for the number of rules generated of the given tree was selected with minimum error of 20%. The results showed that for the given adequate set data, these techniques can be used for weather investigation and climate change studies.*

Keywords: *K-means, Apriori, Clustering Techniques, Weather data, Data set*

I. INTRODUCTION

To expect the weather by numerical means, meteorologists have developed impressive models that approximate the atmosphere by using mathematical equations to describe how atmospheric temperature, pressure, and moisture will change over time. The equations are programmed into a computer and data on the present atmospheric conditions are fed into the computer. The computer solves the equations to determine how the different atmospheric variables will change over the next few minutes. The computer repeats this procedure again and again using the output from one cycle as the input for the next cycle. For some desired time in the future (12, 24, 36, 48, 72 or 120 hours), the computer prints its calculated information. It then analyzes the data, drawing the lines for the projected position of the various pressure systems. The final computer-drawn forecast chart is called a prognostic chart. A forecaster uses the progs as a guide to predicting the weather. There are many atmospheric models that represent the atmosphere, with each one interpreting the atmosphere in a slightly different way. Weather forecasts made for 12 and 24 hours are naturally quite accurate. Forecasts made for two and three days are usually good. Beyond about five days, forecast accuracy falls off swiftly.

Data mining objectives is to provide accurate knowledge in the form of useful rules, techniques, visual graphs and models for the weather parameters over the datasets. This knowledge can be used to support the decision-making for various sectors. The goals for data analysis are those which involve weather variations that affect our daily runtime changes in min and max temperature, humidity level, rainfall chances and speed of wind. This knowledge can be utilized to support many important areas which are affected by climate change includes Agriculture, Water Resources, Vegetation and Tourism. This research work shows that human society is affected in different ways by weather affects.

II. Review of Literature

Many researchers have been conducted for stream clustering of data. Aggarwal et al. [1] proposed CluStream framework for clustering evolving data stream; CluStream uses the concept of pyramidal time frame in conjunction with micro clustering approach. However, the CluStream framework does not handle trajectory stream data.

Elnekave et al. [2] presented an incremental clustering algorithm for finding evolving groups of similar mobile objects in spatiotemporal data. In this algorithm, each trajectory is represented by set of minimal bounding box (MBB), the entire

overlapping between two trajectories. MMBs represent the similarity between them. The algorithm uses a developed version of incremental -mean algorithm to cluster moving object trajectories.

Jensen et al. [3] presented a disk-based algorithm for continuous clustering of moving object. The algorithm employs clustering features structure that can be updated incrementally. Moving object may be deleted from or inserted into a moving cluster during a period of time. Next, the approach merges and splits the clusters through monitoring their compactness.

Li et al. [4] suggested the TCMM framework which consists of two parts: online micro clustering and offline macro clustering for incremental trajectories clustering. Online micro clustering stores statistical information of similar trajectory segments in cluster features (CF) data structure and updates CFs when new batch of segments is added. Similar CFs are merged to solve memory limitation issue. Offline clustering is implemented on the set of micro clusters based on density based clustering when user sends request to see the clustering results. Some studies use optimization strategies such as indexes or pruning to minimize search and enhance the efficiency of clustering.

Yu et al. [5] proposed ConTraClu algorithm to cluster continuous high speed trajectories data stream and discover moving pattern such as flock. The algorithm consists of online clustering of trajectory segments depending on density based approach and updating process of closed clusters depends on bi-Tree index.

Da Silva et al. [6] proposed an incremental algorithm for trajectory data stream. The algorithm uses a micro group structure to truck moving object and its evolution at consecutive time windows. Micro group describes the relationship among moving objects and evolves (merge or split) in the next time period.

Mao et al. [7] produce two-stage framework TSCluWin over sliding window model. During the first stage, sufficient summarized information of the micro clusters is stored and maintained continuously in EF data structure. During the second stage, a small number of micro clusters are produced depending on micro clusters. There also exist some different approaches but they deserve to be mentioned. Costa et al. [8] interpret trajectory as a discrete time signal and use Fourier transform to measure the similarity between two trajectories.

III.MATERIALS AND METHODS

Data Cleaning

In this stage, a consistent format for the data model was developed which took care of missing data, finding duplicated data, and weeding out of bad data. Finally, the cleaned data were transformed into a format suitable for data mining.

Data Selection

At this stage, data relevant to the analysis was decided on and retrieved from the dataset. The meteorological dataset had ten (10) attributes, their type and description is presented in Table 1, while an analysis of the numeric values are presented in Table 2. Due to the nature of the Cloud Form data where all the values are the same and the high percentage of missing values in the sunshine data both were not used in the analysis.

Data Mining Stage

The data mining stage was divided into three phases. At each phase all the algorithms were used to analyze the meteorological datasets. The testing method adopted for this research was percentage split that train on a percentage of the dataset, cross validate on it and test on the remaining percentage. Thereafter interesting patterns representing knowledge were identified.

Data Conversion

That is also known as the data association. This is selected form of data into a suitable data mining stage. Save the data files in comma-separated by value (CSV) format of file and data set was standardized to reduce the data scaling.

Data Mining Phase

This phase has divided into the three more stages. At individually stage, the algorithm for analyzing meteorological data sets is implemented. Then test methods are used in this study which is the percentage split of the data set for training, cross validation and testing of the remaining percentage. Subsequently, this recognizes the knowledge representation of interesting patterns.

Data pre-processing

The data preprocessing is the next step of data mining after collection of data. Challenges in temperature, rainfall and wind speed data; knowledge discovery process is facing poor data quality. Thus, the data is pre-processed to remove noise and unwanted data. Pretreatment means concentrating the removal of other unwanted variables from the data, while the data preprocessing includes these steps:

1. Data scrubbing: it's the stage where noise and irrelevant data is removed. Data cleaning procedures are implemented to fill out missing values and to eliminate noise in recognizing outliers and to correct data irregularities

2. Data integration: it's recognized as the data conversion; in this stage, the suitable form of data is converted for the procedure of data mining by reduction of data and construction of attributes.

IV.PROPOSED ALGORITHM

1. Collect weather data(govdata.in) after every one hour and save it in the original database.
2. After every two hours, previously stored data is converted into ordered data by using Weka tool. All type of data that is stored in database is finally stored in the modified "structural weather database".
3. "Structural weather database" is further divided into four sub-databases on the basis of weather separation.
4. Cluster centers are produced with the help of genetic algorithm after applying improved fuzzy K-Mean clustering algorithm to structural data.
5. Whenever any new data is added into the database then use improved fuzzy K-Means clustering to handle the new data addition.
6. Find the resulting clusters.
7. By using priority based algorithm, calculation of results can be done for different years (max 5 years).
8. By using threshold temperature value ranges, we can forecast the probable weather condition for a particular time period

Table 1: Attributes of Meteorological Dataset

Attribute	Type	Description
Year	Numerical	Year considered
Month	Numerical	Month considered
Wind speed	Numerical	Wind run in km
Evaporation	Numerical	Evaporation
Cloud Form	Numerical	The mean cloud amount
Radiation	Numerical	The amount of radiation
Sunshine	Numerical	The amount of sunshine
Min Temp	Numerical	The monthly Minimum Temperature
Rainfall	Numerical	Total monthly rainfall
Max Temp	Numerical	Maximum Temperature

Table 2:Tabular view for Coimbatore temperature and rainfall per month

	Temperature			Precipitation
Months	Normal	Warmest	Coldest	Normal
January	26.8°C	31.6°C	22.0°C	0
February	27.7°C	32.0°C	23.4°C	0
March	28.9°C	32.7°C	25.0°C	1
April	29.6°C	33.1°C	26.1°C	4
May	29.1°C	32.4°C	25.8°C	10
June	26.7°C	29.4°C	24.0°C	25
July	26.0°C	28.4°C	23.5°C	25
August	25.9°C	28.3°C	23.5°C	23
September	26.8°C	29.5°C	24.0°C	13
October	27.3°C	30.6°C	24.0°C	11
November	27.5°C	31.3°C	23.6°C	7
December	27.2°C	31.6°C	22.7°C	1

Table 3: Coimbatore District Hourly Weather History & Observations

Time (IST)	Temp.	Heat Index	Dew Point	Humidity	Pressure	Visibility	Wind Dir	Wind Speed	Conditions
12:00 AM	26.0 °C	-	22.0 °C	78%	1014 hPa	4.0 km	East	9.3 km/h / 2.6 m/s	Haze
12:30 AM	26.0 °C	-	21.0 °C	74%	1014 hPa	4.0 km	East	9.3 km/h / 2.6 m/s	Haze
1:00 AM	26.0 °C	-	21.0 °C	74%	1014 hPa	4.0 km	East	3.7 km/h / 1.0 m/s	Haze
1:30 AM	25.0 °C	-	22.0 °C	83%	1014 hPa	4.0 km	ENE	3.7 km/h / 1.0 m/s	Haze
2:30 AM	25 °C	-	22 °C	77%	1012 hPa	4 km	NE	9.3 km/h /	Mist
2:30 AM	25.0 °C	-	22.0 °C	83%	1013 hPa	4.0 km	NE	9.3 km/h / 2.6 m/s	Haze
3:00 AM	25.0 °C	-	22.0 °C	83%	1013 hPa	4.0 km	ENE	9.3 km/h / 2.6 m/s	Scattered Clouds
3:30 AM	25.0 °C	-	23.0 °C	89%	1013 hPa	4.0 km	ENE	9.3 km/h / 2.6 m/s	Scattered Clouds
4:00 AM	25.0 °C	-	23.0 °C	89%	1013 hPa	3.5 km	ENE	13.0 km/h / 3.6 m/s	Scattered Clouds
4:30 AM	24.0 °C	-	22.0 °C	89%	1013 hPa	3.0 km	ENE	7.4 km/h / 2.1 m/s	Scattered Clouds
5:30 AM	24 °C	-	22 °C	89%	1012 hPa	4 km	NE	5.6 km/h /	Mist
5:30 AM	24.0 °C	-	23.0 °C	94%	1011 hPa	3.0 km	NE	5.6 km/h / 1.5 m/s	Partly Cloudy
7:00 AM	24.0 °C	-	23.0 °C	94%	1014 hPa	1.5 km	ENE	7.4 km/h / 2.1 m/s	Scattered Clouds
7:30 AM	24.0 °C	-	22.0 °C	89%	1015 hPa	2.0 km	ENE	5.6 km/h / 1.5 m/s	Scattered Clouds
8:00 AM	25.0 °C	-	22.0 °C	83%	1015 hPa	2.0 km	East	9.3 km/h / 2.6 m/s	Scattered Clouds
8:30 AM	25 °C	-	22 °C	81%	1014 hPa	4 km	East	9.3 km/h /	Mist
8:30 AM	25.0 °C	-	22.0 °C	83%	1016 hPa	2.0 km	East	9.3 km/h / 2.6 m/s	Scattered Clouds
9:00 AM	25.0 °C	-	22.0 °C	83%	1016 hPa	2.5 km	East	9.3 km/h / 2.6 m/s	Scattered Clouds
9:30 AM	27.0 °C	29.2 °C	22.0 °C	74%	1016 hPa	2.5 km	East	11.1 km/h / 3.1 m/s	Scattered Clouds
10:00 AM	28.0 °C	30.7 °C	22.0 °C	70%	1016 hPa	3.0 km	ENE	11.1 km/h / 3.1 m/s	Haze
10:30 AM	29.0 °C	32.0 °C	22.0 °C	66%	1015 hPa	4.0 km	East	14.8 km/h / 4.1 m/s	Haze
11:00 AM	29.0 °C	31.3 °C	21.0 °C	62%	1015 hPa	4.0 km	East	14.8 km/h / 4.1 m/s	Haze
11:30 AM	30 °C	-	20 °C	46%	1012 hPa	4 km	NE	11.1 km/h /	Haze
11:30 AM	30.0 °C	31.9 °C	20.0 °C	55%	1014 hPa	5.0 km	NE	11.1 km/h / 3.1 m/s	Haze
12:30 PM	31.0 °C	32.4 °C	19.0 °C	49%	1013 hPa	6.0 km	East	11.1 km/h / 3.1 m/s	Scattered Clouds
1:00 PM	32.0 °C	34.1 °C	20.0 °C	49%	1013 hPa	6.0 km	NE	11.1 km/h / 3.1 m/s	Scattered Clouds
1:30 PM	32.0 °C	33.5 °C	19.0 °C	46%	1012 hPa	8.0 km	East	9.3 km/h / 2.6 m/s	Scattered Clouds
2:00 PM	32.0 °C	33.5 °C	19.0 °C	46%	1011 hPa	8.0 km	ENE	9.3 km/h / 2.6 m/s	Scattered Clouds
2:30 PM	32 °C	-	19 °C	36%	1008 hPa	4 km	East	9.3 km/h /	Haze
2:30 PM	32.0 °C	33.5 °C	19.0 °C	46%	1011 hPa	8.0 km	East	9.3 km/h / 2.6 m/s	Scattered Clouds
3:00 PM	32.0 °C	33.5 °C	19.0 °C	46%	1010 hPa	8.0 km	NE	7.4 km/h / 2.1 m/s	Scattered Clouds

Time (IST)	Temp.	Heat Index	Dew Point	Humidity	Pressure	Visibility	Wind Dir	Wind Speed	Conditions
3:30 PM	32.0 °C	33.5 °C	19.0 °C	46%	1010 hPa	8.0 km	East	14.8 km/h / 4.1 m/s	Scattered Clouds
4:00 PM	32.0 °C	33.5 °C	19.0 °C	46%	1010 hPa	8.0 km	East	9.3 km/h / 2.6 m/s	Partly Cloudy
4:30 PM	32.0 °C	33.5 °C	19.0 °C	46%	1010 hPa	8.0 km	East	11.1 km/h / 3.1 m/s	Partly Cloudy
5:00 PM	32.0 °C	33.5 °C	19.0 °C	46%	1010 hPa	8.0 km	East	11.1 km/h / 3.1 m/s	Partly Cloudy

V.SYSTEM DESIGNED

Weather forecasting system provides us the information about future weather conditions for a particular region, locality over a specified time period. Weather directly depends upon the air molecules which can absorb high frequency sunrays. The air molecules data is collected by the system periodically after every one hour. The Weka tool uses these raw data to bound large information to find the "Structural weather Database". Then we have applied improved k-mean algorithm to this database. Then we have stored the resultant data in the main database. So accordingly, we classified the main database into four regions that are grouped according to the direction of wind flow over the year.

First region includes December, January, February; second region includes March, April; in the third region May, June, July; and in the last region August, September, October, November are included. First region is considered as winter region. Second and fourth are known as temperate region. Third is called summer region. When we have to search any data, then we can search it in its particular domain. In the k-mean algorithm we have organized data into clusters according to the region. Incremental improved k-mean algorithm is used for the addition of any new data and it makes the data fit into the proper cluster. The initial cluster center is generated by using the genetic algorithm. When the user enters data to the system, then it is compared with the previous set of data using the priority based algorithm. Multiple year data is stored in the database, now with the help of [(1/3-α), (α), (1/3)] we can use the order of priority. With the help of the record of statistics, we can conclude that weather basically depends on 2nd last year. So choose priority (1/3-α), (α) and (1/3) for the lastyear, 2nd last year (highest priority) and 3rd last year respectively. Prediction calculation has been done based on three years, where α is taken as a constant variable.

VI.PERFORMANCE ANALYSIS

The collected data sets are used in Weka to forecast analyse the weather data with the improved fuzzy K-means cluster algorithm.

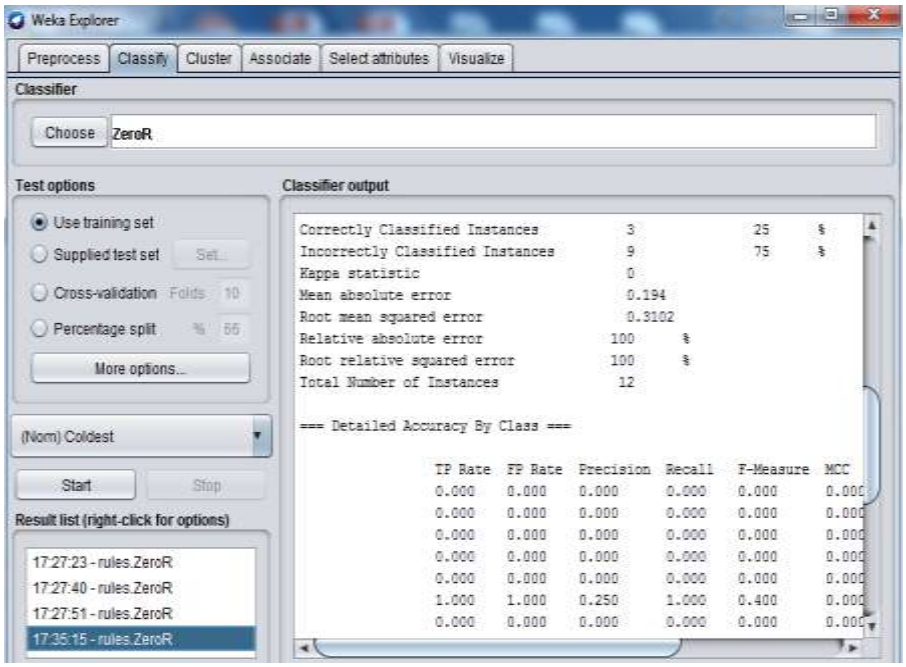

Figure 1: Classification of weather data

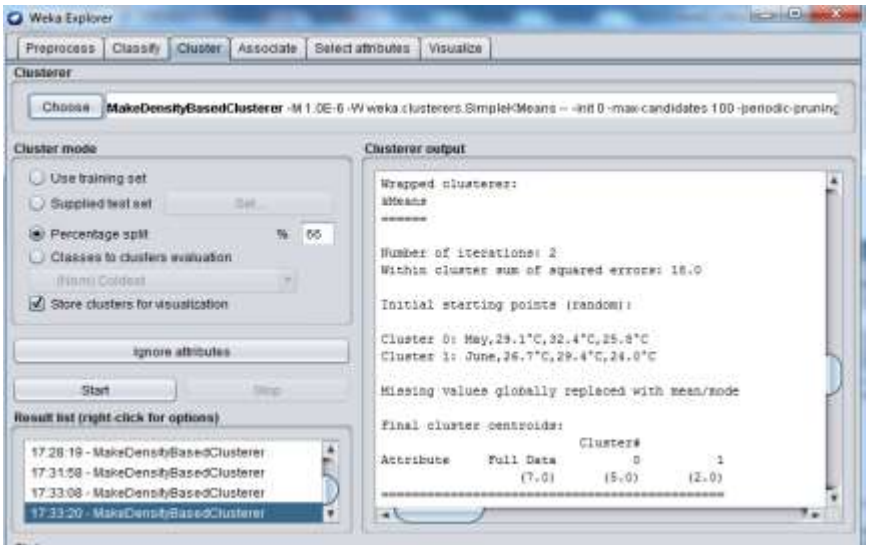

Figure 2: Clustering the data sets

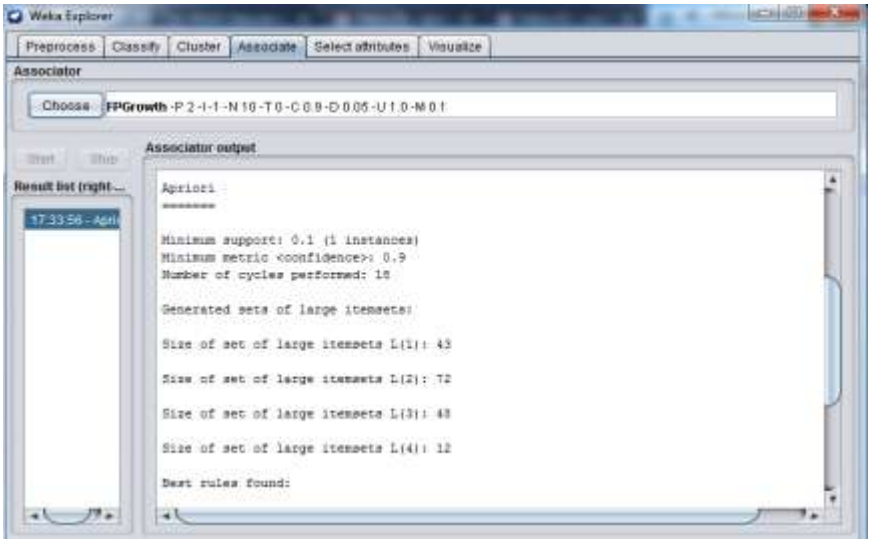

Figure 3: Associativity among the collected weather data

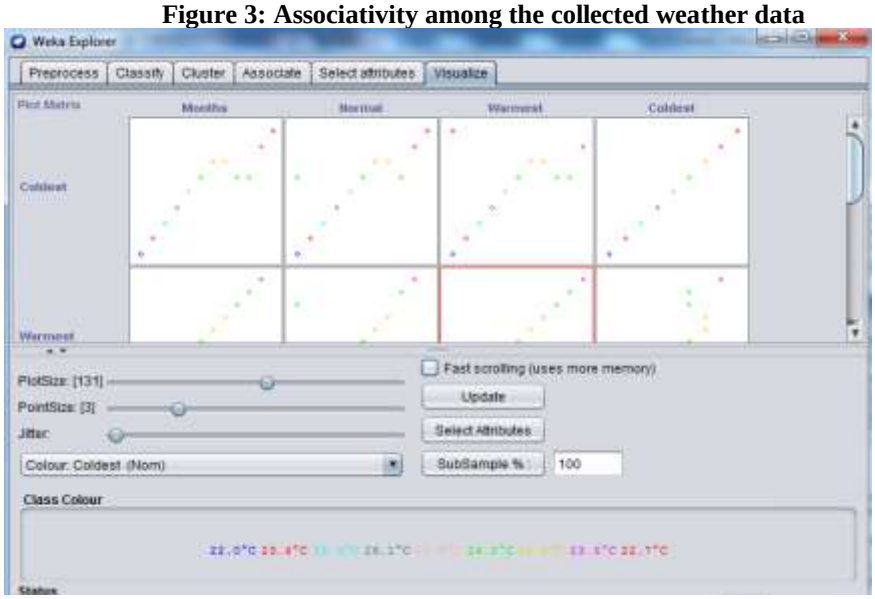

Figure 4: improved Fuzzy K-means cluster based weather data visualization

VII.CONCLUSION

With the help of proposed improved fuzzy K-mean clustering algorithm and weka tool, we have introduced a new technique to forecast the weather data in Coimbatore region. This technique is also suitable for the dynamic environment where the weather conditions change frequently. The fuzzy algorithm was used by us to find out or to make guess of initial cluster center. It has given us more suitable results. This technique does not give completely accurate results; it just forecasts the probable results. In the upcoming work, it can be extended to any huge data sets with varied parameters for effective analysis and correct prediction. This method can used to some other air pollution databases of different regions for weather forecasting.

REFERENCES

1. C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for clustering evolving data streams," in Proceedings of the 29th international conference on Very large data bases (VLDB Endowment '03), vol. 29, Elsevier, 2003.
2. S. Elnekave, M. Last, and O. Maimon, "Incremental clustering of mobile objects," in Proceedings of the Workshops in Conjunction with the 23rd International Conference on Data Engineering (ICDE '07), pp. 585–592, April 2007. View at Publisher · View at Google Scholar · View at Scopus
3. C. S. Jensen, D. Lin, and B. C. Ooi, "Continuous clustering of moving objects," IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 9, pp. 1161–1173, 2007. View at Publisher · View at Google Scholar · View at Scopus
4. Y. Yu, Q. Wang, and X. Wang, "Continuous clustering trajectory stream of moving objects," China Communications, vol. 10, no. 9, Article ID 6623510, pp. 120–129, 2013. View at Publisher · View at Google Scholar · View at Scopus
5. Y. Yanwei, Q. Wang, X. Wang, H. Wang, and J. He, "Online clustering for trajectory data stream of moving objects," Computer Science and Information Systems, vol. 10, no. 3, pp. 1293–1317, 2013. View at Google Scholar
6. T. L. C. Da Silva, K. Zeitouni, and J. A. F. De Macedo, "Online clustering of trajectory data stream," in Proceedings of the 17th IEEE International Conference on Mobile Data Management (IEEE MDM '16), pp. 112–121, June 2016. View at Publisher · View at Google Scholar · View at Scopus
7. T. L. C. da Silva, T. L. Coelho, K. Zeitouni, J. A. F. de Macêdo, and M. A. Casanova, "CUTiS: optimized online ClUstering of Trajectory data Stream," in Proceedings of the 20th International Database Engineering & Applications Symposium, pp. 296–301, ACM, 2016.
8. G. Costa, G. Manco, and E. Masciari, "Dealing with trajectory streams by clustering and mathematical transforms," Journal of Intelligent Information Systems, vol. 42, no. 1, pp. 155–177, 2014.
9. Kaur, M. (2013). Big Data and Methodology-A review. International Journal of Advanced Research in Computer Science and Software Engineering,
10. Sabia, S. K. (2014). Applications of Big Data. International Journal on Advanced Computer Theory and Engineering (IJACTE).
11. R.Rajeshkanna and Dr A.Saradha "Cluster Based Load Balancing Techniques to Improve the Lifetime of Mobile Adhoc Networks", International Journal of Trend in Research and Development (IJTRD), ISSN: 2394-9333, Volume-2 | Issue-5 , October 2015.
12. Yuvraj S. Sase, P. A. (2014). Big Data Implementation Using Hadoop and Grid Computing. International Journal of Innovative Research in Science, Engineering and Technology , 6.
13. Harshawardhan S. Bhosale, P. D. (2014). A review paper on Big data and Hadoop. International Journal of Scientific and Research Publications ,
14. C.Chandhini, MeganaL.P , P. A (2013). Grid Computing-A Next Level Challenge with Big Data. International Journal of Scientific & Engineering Research.

