A Study Paper on Outlier Detection on Time Series Data

\[1]Dipannita Kar, [2] Mr. Haresh Chande, [3] Mr. Rajendra Gaikwad

[1] Research Scholar (M.E.), [2] Assistant Professor, [3] SCI/ ENGR-SD
[1,2] Department of Computer Engineering,
[1,2] HJD Institute of Technical Education and Research
Kera, Kutch, Gujarat Technological University, Gujarat, India
[3] EPSA -VRG-CGDD, ISRO
Gujarat, India

Abstract: Time series data are observations collected sequentially over time. Consider as example weather prediction. When observations are collected at frequent time intervals, large data sets in the form of time series are generated. Outlier detection is a primary step in many data mining. An outlier is a piece of data or observation that deviates from other observations, outliers is not noise. An outlier may be due to variability in the measurement. To identify & remove outliers is challenging task in data mining. There are many algorithms for outlier detection. There are various algorithms for outlier detection are proposed earlier, some of them are k-mean, Density based, EM etc. They are proposed to detect the outliers. In these algorithm the time is an important attribute of each dataset, and also it is important in the process of data mining for giving the more accurate and useful information. In this work, algorithms are analyzing by using WEKA tool. In our research work, the algorithms analyzed are K-Mean, Density based, EM, Cobweb.

Index Terms **- Time Series, Outliers, K-mean algorithm, Density based algorithm, EM, WEKA.**

I. INTRODUCTION

In recent years, data mining has been obtained a great deal of attention in the information industry and in society, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. The gained information and knowledge can be used for various applications which can range from market analysis to production control and science researches [1]. Data mining is used in classification, clustering, association rule discovery, outlier detection etc.

Mining temporal large data sets is an active area of research. Time series is a kind of temporal data. Time series mining is process of mining time series large data sets for their knowledge identification. It has also provided a research opportunity for detecting and predicting event-based knowledge. Events can be defined as real-world occurrences that unfold over space and time. Event prediction which is one of the main goals in time series mining is an important research topic [3].

Outliers refers to data points that are dissimilar to the remaining points in the data set. In the energy domain, clean data is required to train and develop accurate models for forecasting [2]. Outlier detection is a fundamental issue in data mining. The identification of outliers can lead to the discovery of useful knowledge and has a number of practical applications in area such as transactions record, stock price movement, sensor data, weather prediction etc. Outliers exist in all types of data, and the detection of outlier is a challenging task in data mining.

II. WEKA TOOL

WEKA is a data mining system developed by the University of Waikato in New Zealand that implements data mining algorithms. WEKA is a state-of-the-art facility for developing machine learning (ML) techniques and their application to real-world data mining problems. It is a collection of machine learning algorithms for data mining tasks. The algorithms are applied directly to a dataset. WEKA implements algorithms for data pre-processing, classification, regression, clustering, association rules; it also includes a visualization tools. The new machine learning schemes can also be developed with this package. WEKA is open source software issued under the GNU General Public License [4].

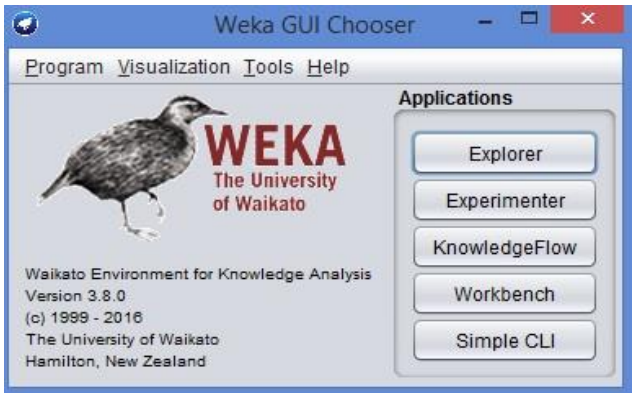


Figure 1 The WEKA tool GUI

The WEKA tool GUI consists of four buttons:

- Explorer: An environment for exploring data with WEKA.
- Experimenter: This is an environment for performing the experiments and conducting statistical tests between learning schemes.
- Knowledge Flow: This environment supports essentially the same functions as the Explorer but with a drag-and-drop interface. One advantage is that it supports incremental learning.
- Workbench: This is a unified graphical interface that combines the other three into one application. The workbench is highly configurable allowing the user to
- Simple CLI: Provides a simple command-line interface that allows direct execution of WEKA commands for operating systems that do not provide their own command line interface.

WEKA Explorer window:

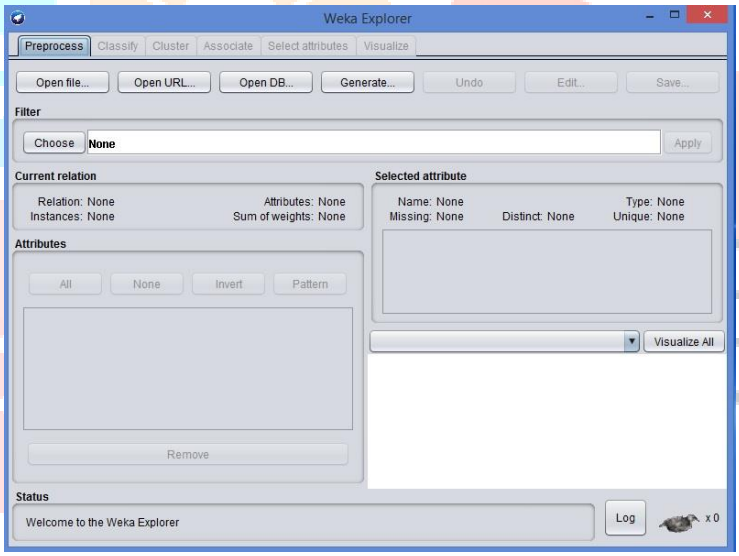


Figure 2 WEKA Explorer

III. LITERATURE SURVEY

There are various authors which have research various existing technique to detect the outliers in time series data.

A Short introduction to Data Mining and its Applications: In this paper, author present the short introduction of data mining and its applications. This paper introduces some models and solutions of data mining, especially emphasizes on the applications of the associated rules and solutions of data mining in e-commerce. Author can also present some useful data mining patterns. After that some problems of data mining can also discuss [1].

A Comparative Analysis of Clustering Algorithms: In this paper authors perform a comparative analysis of four clustering algorithms namely K-means algorithm, Hierarchical algorithm, Expectation and maximization algorithm and Density based algorithm. These algorithms are compared in terms of efficiency and accuracy, using WEKA tool. Comparison is performed on Bank dataset. After apply normalization only simple K-means clustering algorithms forms clusters with less time and more accuracy than other algorithms. In terms of time and accuracy K-means produces better results as compared to other algorithms [5].

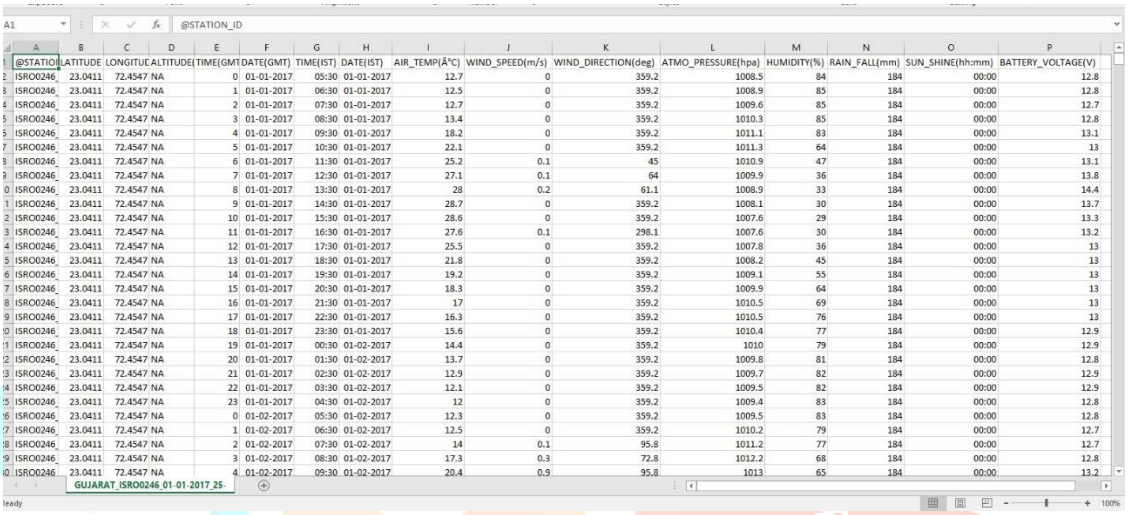
Analysis of Clustering Algorithm of Weka Tool on Air Pollution Dataset: In this paper, algorithms are analysing and comparing the various clustering algorithm by using WEKA tool to find out which algorithm will be more comfortable for the users for performing clustering algorithm. Authors have performed analysis with four clustering algorithms K-mean, LVQ, SOM and COBWEB. After comparison, the best algorithm found is K-mean algorithm. It takes less time than other clustering algorithm [6].

Performance Analysis of Clustering Algorithms in detecting Outliers: This paper presents the analysis of K-means and K-Medians clustering algorithm in detecting outliers. The k-means clustering and k-medians clustering algorithm's performance in

detecting outliers are analysed here. K-means clustering clusters the similar data with the help of the mean value. K-medians is similar to k-means algorithm but median values are calculated there. After comparison, K-Means Clustering algorithm is taking more time to compute the outliers. So, in minimizing the errors, K-Medians Clustering algorithm is efficient enough than K-Means Clustering algorithm [9].

IV. DATASET

For performing the comparison analysis "Automatic Weather Station (AWS)" data has been used. AWS stations are capable of recording data about various weather related parameters like air-temperature, humidity, wind-speed, wind direction, atmospheric pressure, rainfall and sunshine. The data are recorded at the interval of one hour on 24 X 7 basis. These data can be obtained free of cost to registered users through Meteorological and Oceanographic Satellite Data Archival Centre (MOSDAC) [10]. In this analysis "Aws data" is used in .csv file format. Following table show the AWS data.



@STATION_ID	LATITUDE	LONGITUDE	ALTITUDE(	TIME(GMT)	DATE(GMT)	TIME(IST)	DATE(IST)	AIR_TEMP(Â°C)	WIND_SPEED(m/s)	WIND_DIRECTION(deg)	ATMO_PRESSURE(hpa)	HUMIDITY(%)	RAIN_FALL(mm)	SUN_SHINE(hh:mm)	BATTERY_VOLTAGE(V)
ISRO0246	23.0411	72.4547	NA	0	01-01-2017	05:30	01-01-2017	12.7	0	359.2	1008.5	84	184	00:00	12.8
ISRO0246	23.0411	72.4547	NA	1	01-01-2017	06:30	01-01-2017	12.5	0	359.2	1008.9	85	184	00:00	12.8
ISRO0246	23.0411	72.4547	NA	2	01-01-2017	07:30	01-01-2017	12.7	0	359.2	1009.6	85	184	00:00	12.7
ISRO0246	23.0411	72.4547	NA	3	01-01-2017	08:30	01-01-2017	13.4	0	359.2	1010.3	85	184	00:00	12.8
ISRO0246	23.0411	72.4547	NA	4	01-01-2017	09:30	01-01-2017	18.2	0	359.2	1011.1	83	184	00:00	13.1
ISRO0246	23.0411	72.4547	NA	5	01-01-2017	10:30	01-01-2017	22.1	0	359.2	1011.3	64	184	00:00	13
ISRO0246	23.0411	72.4547	NA	6	01-01-2017	11:30	01-01-2017	25.2	0.1	45	1010.9	47	184	00:00	13.1
ISRO0246	23.0411	72.4547	NA	7	01-01-2017	12:30	01-01-2017	27.1	0.1	64	1009.9	36	184	00:00	13.8
ISRO0246	23.0411	72.4547	NA	8	01-01-2017	13:30	01-01-2017	28	0.2	61.1	1008.9	33	184	00:00	14.4
ISRO0246	23.0411	72.4547	NA	9	01-01-2017	14:30	01-01-2017	28.7	0	359.2	1008.1	30	184	00:00	13.7
ISRO0246	23.0411	72.4547	NA	10	01-01-2017	15:30	01-01-2017	28.6	0	359.2	1007.6	29	184	00:00	13.3
ISRO0246	23.0411	72.4547	NA	11	01-01-2017	16:30	01-01-2017	27.6	0.1	298.1	1007.6	30	184	00:00	13.2
ISRO0246	23.0411	72.4547	NA	12	01-01-2017	17:30	01-01-2017	25.5	0	359.2	1007.8	36	184	00:00	13
ISRO0246	23.0411	72.4547	NA	13	01-01-2017	18:30	01-01-2017	21.8	0	359.2	1008.2	45	184	00:00	13
ISRO0246	23.0411	72.4547	NA	14	01-01-2017	19:30	01-01-2017	19.2	0	359.2	1009.1	55	184	00:00	13
ISRO0246	23.0411	72.4547	NA	15	01-01-2017	20:30	01-01-2017	18.3	0	359.2	1009.9	64	184	00:00	13
ISRO0246	23.0411	72.4547	NA	16	01-01-2017	21:30	01-01-2017	17	0	359.2	1010.5	69	184	00:00	13
ISRO0246	23.0411	72.4547	NA	17	01-01-2017	22:30	01-01-2017	16.3	0	359.2	1010.5	76	184	00:00	13
ISRO0246	23.0411	72.4547	NA	18	01-01-2017	23:30	01-01-2017	15.6	0	359.2	1010.4	77	184	00:00	12.9
ISRO0246	23.0411	72.4547	NA	19	01-01-2017	00:30	01-02-2017	14.4	0	359.2	1010	79	184	00:00	12.9
ISRO0246	23.0411	72.4547	NA	20	01-01-2017	01:30	01-02-2017	13.7	0	359.2	1009.8	81	184	00:00	12.8
ISRO0246	23.0411	72.4547	NA	21	01-01-2017	02:30	01-02-2017	12.9	0	359.2	1009.7	82	184	00:00	12.9
ISRO0246	23.0411	72.4547	NA	22	01-01-2017	03:30	01-02-2017	12.1	0	359.2	1009.5	82	184	00:00	12.9
ISRO0246	23.0411	72.4547	NA	23	01-01-2017	04:30	01-02-2017	12	0	359.2	1009.4	83	184	00:00	12.8
ISRO0246	23.0411	72.4547	NA	0	01-02-2017	05:30	01-02-2017	12.3	0	359.2	1009.5	83	184	00:00	12.8
ISRO0246	23.0411	72.4547	NA	1	01-02-2017	06:30	01-02-2017	12.5	0	359.2	1010.2	79	184	00:00	12.7
ISRO0246	23.0411	72.4547	NA	2	01-02-2017	07:30	01-02-2017	14	0.1	95.8	1011.2	77	184	00:00	12.7
ISRO0246	23.0411	72.4547	NA	3	01-02-2017	08:30	01-02-2017	17.3	0.3	72.8	1012.2	68	184	00:00	12.8
ISRO0246	23.0411	72.4547	NA	4	01-02-2017	09:30	01-02-2017	20.4	0.9	95.8	1013	65	184	00:00	13.2

Figure 3 AWS data

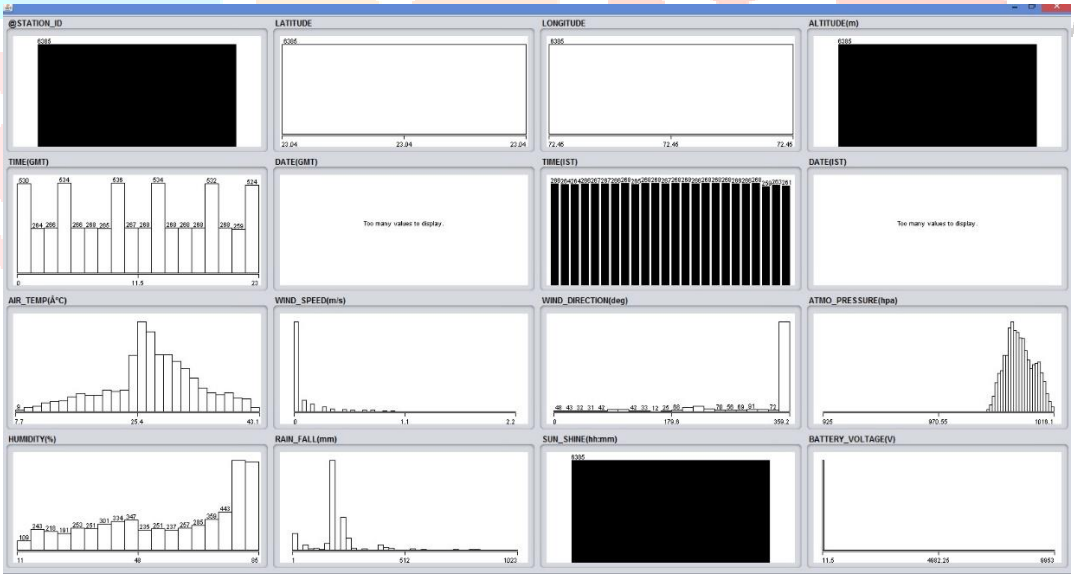


Figure 4 Exploring Dataset

V. ANALYSIS OF VARIOUS ALGORITHM USING WEKA TOOL

In this research work various algorithms are used to detect outliers. Those algorithms are:

I. K-Mean Algorithm
II. Density based Algorithm
III. EM Algorithm
IV. Cobweb Algorithm.

1. K-Mean Algorithm

K-mean clustering generates a specific number of disjoint clusters. The basic objective of this paper is to detect outliers or abnormal temperature reading in weather forecast using k-means clustering technique. To achieve this objective, k-mean data clustering technique is applied to find out the abnormal temperature based on input data.

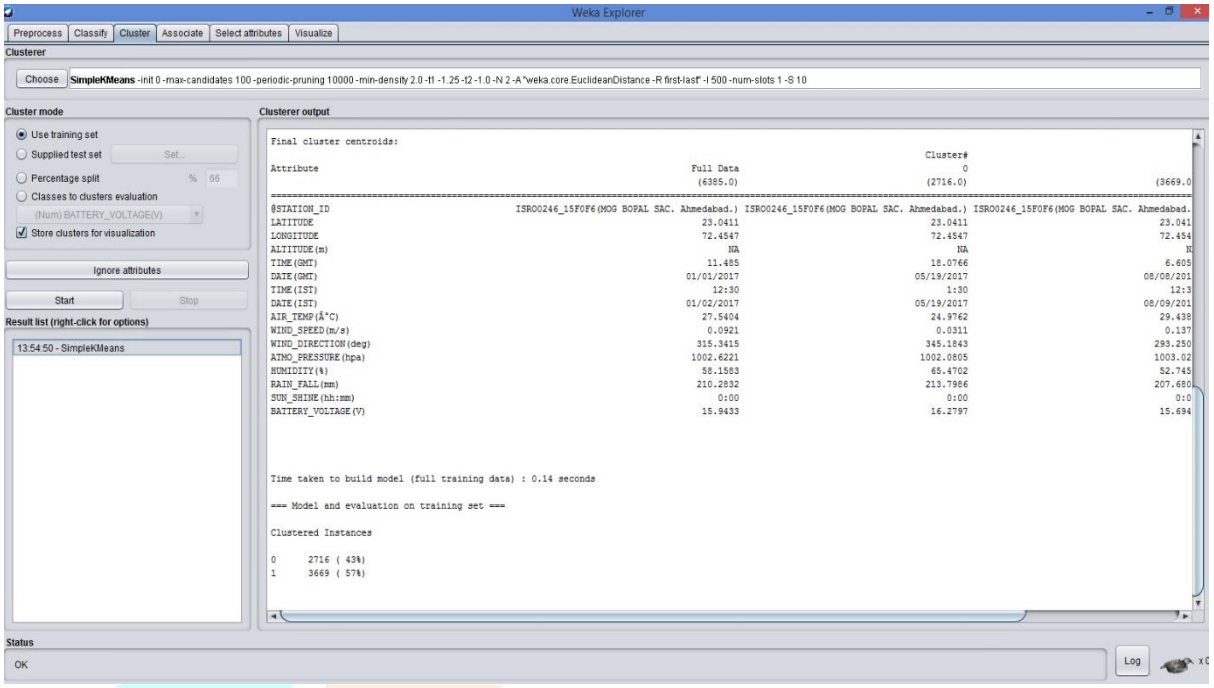


Figure 5 Applying K-Mean Algorithm

2. Density Based Algorithm

Density based algorithms typically regard clusters as dense regions of objects in the data space that are separated by regions of low density. They find and separate regions of high density from low density regions. It connects core objects and their neighbourhoods to form dense regions as clusters. Clusters are formed as maximum sets of density connected points and can detect noise and used when outliers are encountered [5].

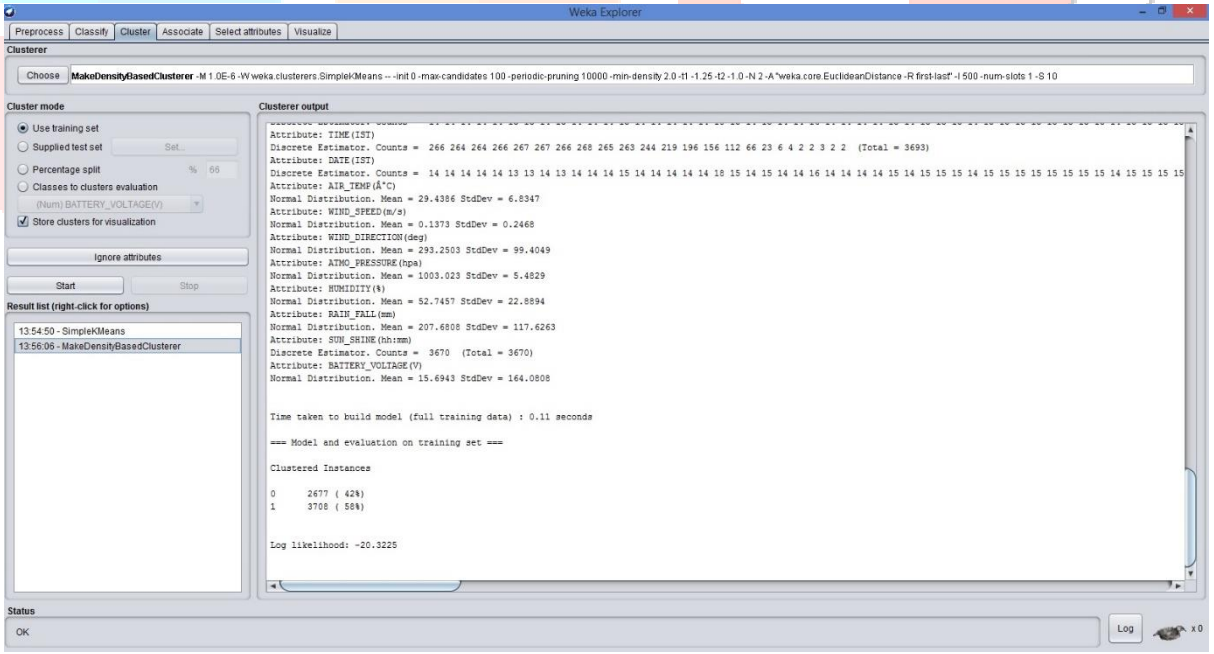


Figure 6 Applying Density based Algorithm

3. Expectation-maximization (EM) Algorithm

The EM algorithm is used to find (local) maximum likelihood parameters of a statistical model in cases where the equations cannot be solved directly. Typically, these models involve hidden variables in addition to unknown parameters and known data observations. That is, either missing values exist among the data, or the model can be formulated more simply by assuming the existence of further unobserved data points.

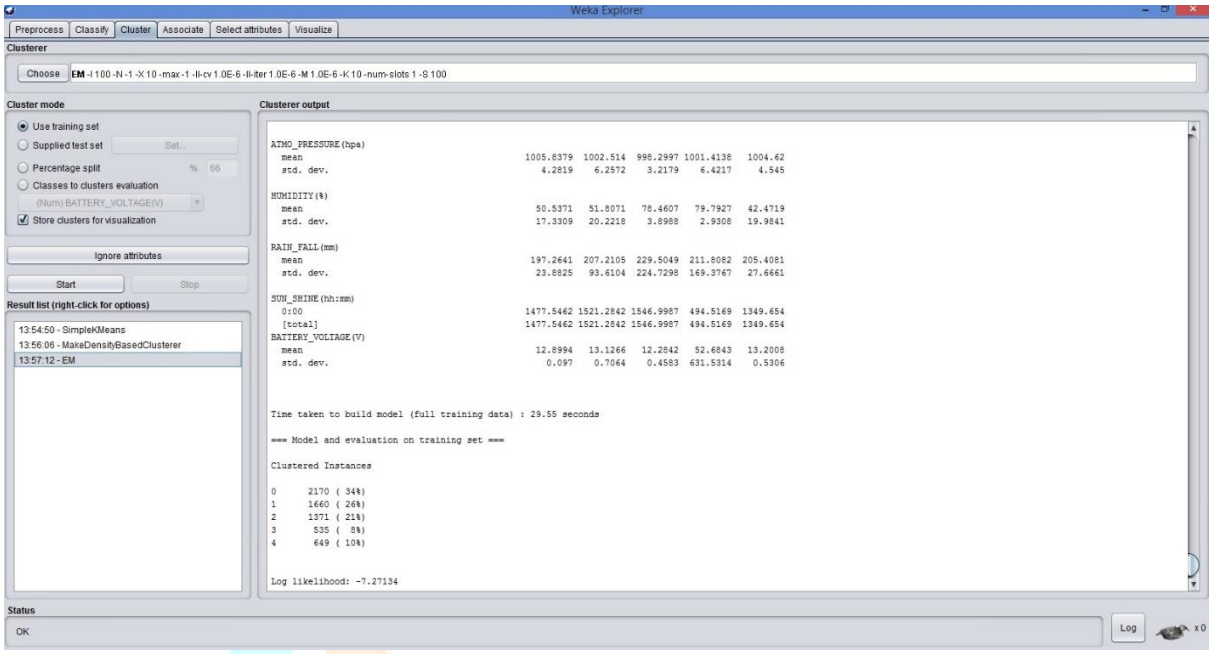


Figure 7 Applying EM Algorithm

4. Cobweb Algorithm

COBWEB is an incremental system for hierarchical conceptual clustering. COBWEB was invented by Professor Douglas H. Fisher, currently at Vanderbilt University. COBWEB incrementally organizes observations into a classification tree. Each node in a classification tree represents a class and is labelled by a probabilistic concept that summarizes the attribute-value distributions of objects classified under the node. This classification tree can be used to predict missing attributes or the class of a new object.

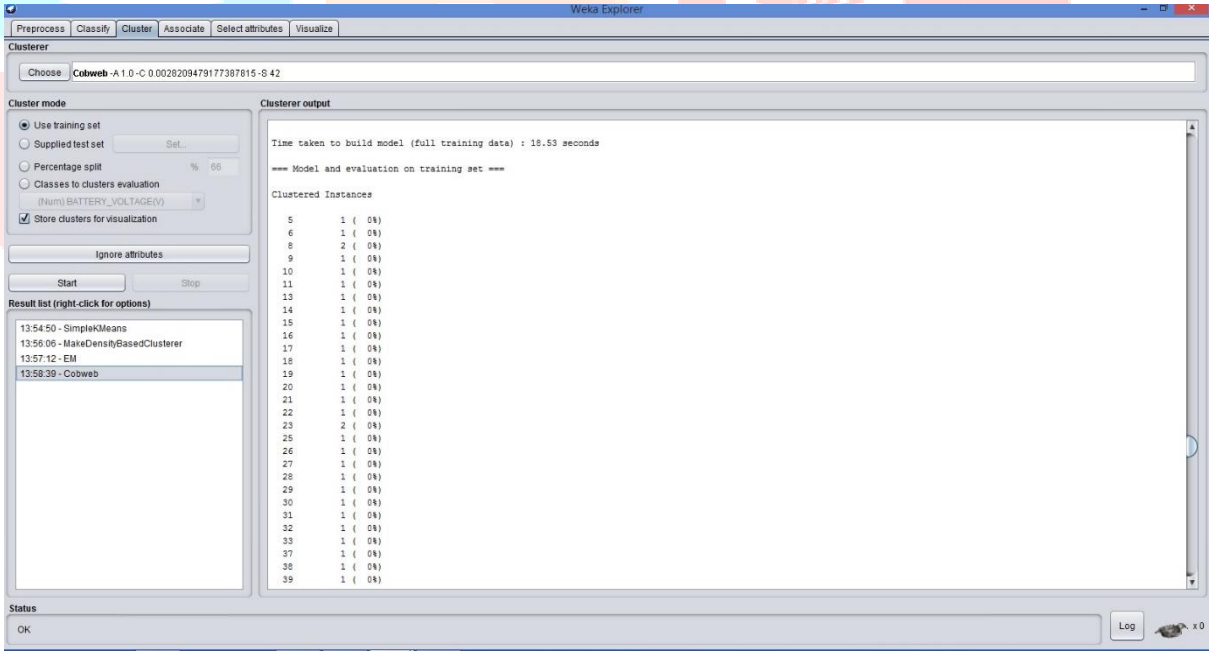


Figure 8 Applying Cobweb Algorithm

VI. RESULT ANALYSIS

The following table represents the analysis result of all algorithm:

Table 1 Comparison result of different clustering algorithm

Algorithm	No. of clusters	Cluster instance	No. of iteration	Time taken to build
K-Mean	2	2716 (43%) 3669 (57%)	14	0.16s
Density Based	2	2677 (42%) 3708 (58%)	14	0.08s
EM	5	2170 (34%) 1660 (26%) 1371 (21%) 535 (8%) 649 (10%)	3	29.95s
Cobweb	8257	1 (0%) 1 (0%) . . . 1 (0%)	-	18.48s

VII. CONCLUSION

In this paper, comparative study has been performed on the k-means, Density based, EM and Cobweb algorithm. Comparison is performed on AWS data using WEKA tool. Comparative results are shown in the form of table. The comparative study is performed on the basis of time. The best algorithm is Density Based algorithm. It takes less time than other algorithms.

REFERENCES

[1] Zhang Haiyang, "A Short Introduction to Data Mining and its Applications", Institute of Electrical and Electronics Engineers (IEEE), 2011

[2] Hermine N. Akouemo and Richard J. Povineli, "Time Series Outlier Detection and Imputation", Institute of Electrical and Electronics Engineers (IEEE), 2014

[3] Soheila Mehr Molaei and Mohammad Reza Keyvanpour, "An Analytical Review for Event Prediction System on Time Series" 2nd International Conference on Pattern Recognition and Image Analysis (IPRIA 2015) March 11-12, 2015

[4] Introduction to WEKA available at: http://people.sabanciuniv.edu/berrin/cs512/lectures/WEKA/WEKA%20Explorer%20Tutorial-REFERENCE.pdf

[5] Raj Bala, Sunil Sikka and Juhi Singh, "A Comparative Analysis of Clustering Algorithms", International Journal of Computer Applications (0975 – 8887) Volume 100 – No.15, August 2014

[6] Richa Agrawal and Jitendra Agrawal, "Analysis of Clustering Algorithm of WEKA Tool on Air Pollution Dataset", International Journal of Computer Applications (0975 – 8887) Volume 168 – No.13, June 2017

[7] http://www.cs.utexas.edu/users/ml/tutorials/Weka-tut/sld020.html

[8] Deepti V. Patange Dr. Pradeep K. Butey S. E. Tayde, "Analytical Study of Clustering Algorithms by Using Weka", National Conference on "Advanced Technologies in Computing and Networking"-ATCON-2015 Special Issue of International Journal of Electronics, Communication & Soft Computing Science and Engineering, ISSN: 2277-9477.

[9] Sairam, Manikandan, Sowndarya, "Performance Analysis of Clustering Algorithms in Detecting Outliers", International Journal of Computer Science and Information Technologies, Vol. 2 (1), 2011, 486-488

[10] Dipak Maroo and Markand Oza, "Assessment of ISRO Automatic Weather Station (AWS) Data", SAC/EPSA/ADVG/DWD/AWS-02, June 2014