Analysis of Clustering Algorithm of Weka Tool on Air Pollution Dataset

Richa Agrawal
SOIT, RGPV Bhopal M.P

Jitendra Agrawal
SOIT, RGPV Bhopal M.P

ABSTRACT
Data mining is the process of extracting knowledge from the huge amount of data. The data can be stored in databases and information repositories. Data mining task can be divided into two models descriptive and predictive model. In the Predictive model, we can predict the values from a different set of sample data, they are classified into three types such as classification, regression and time series. The descriptive model enables us to determine patterns in a sample data and sub-divided into clustering, summarization and association rules. Clustering creates a group of classes based on the patterns and relationship between the data. There is different types of clustering algorithms partition, density based algorithm. In this paper, algorithms are analyzing 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. This present the application's of data minning WEKA tool it provide the cluster's huge data set and clustering thet provide making hand in the optimizing in search engine.

Keywords
Data Mining, Clustering algorithms, K-mean, LVQ, SOM, cobweb, WEKA

1. INTRODUCTION
Clustering is one of the descriptive models used to cluster a set of objects into certain groups according to their relationships Clustering is a technique used in many fields such as image analysis, pattern recognition, statistical data analysis etc.Clustering is a division of data into groups of similar objects. Each cluster consists of various objects that are similar amongst them and dissimilar compared to objects of other groups. Different clustering algorithms are present to form clusters [1]. WEKA tool is used to compare different clustering algorithms. It is used because it provides a better interface to the user than compare to other data mining tools. In this paper, there is the comparison of partitioning and non partitioning based clustering algorithms. why chooses WEKA, because we can work in weka easily without having the deep knowledge of data mining techniques.Section 1 gives the introduction about clustering algorithms and WEKA tool. Section 2 defines literature survey. Section 3 describes the basis for algorithm comparison. Section 4 shows the results and section 5 concludes the paper.

2. WHAT IS CLUSTER ANALYSIS?
Cluster analysis[1] teams objects (observations, events) supported the {data} found within the data describing the objects or their relationships. The goal is that the objects in a very cluster are similar (or related) to 1 different and totally different from (or unrelated to) the objects in different teams. The larger the likeness (or homogeneity) among a gaggle, and therefore the larger the inequality between teams, the —better or a lot of distinct the clump. The definition of what constitutes a cluster isn't well outlined, and, in several applications clusters don't seem to be well separated from each other. yet, most cluster analysis seeks as a result, a crisp classification of the information into non-overlapping teams. to higher perceive the issue of deciding what constitutes a cluster, take into account figures 1a through 1b, that show fifteen points and 3 other ways that they'll be divided into clusters. If we tend to permit clusters to be nested, then the foremost cheap interpretation of the structure of those points is that there ar 2 clusters, every of that has 3 sub clusters. However, the apparent division of the 2 larger clusters into 3 sub clusters could merely be associate degree artefact of the human sensory system. Finally, it's going to not be unreasonable to mention that the points from four clusters. Thus, we tend to stress once more that the definition of what constitutes a cluster is general, and therefore the best definition depends on the kind of knowledge and therefore the desired results.

3. WEKA TOOL
WEKA is one of the users friendly and an open source software runs on any platform. WEKA tool was developed by the University of Waikato in New Zealand. In beginning WEKA tool was written in C language, Later the application has been rewritten in java language. WEKA provides the implementation of an algorithm which can be applied to a data set. It includes many algorithms for clustering, association rule mining, attribute selection and regression. WEKA has two file format ARFF(attribute relation file format) and CSV(comma separated values). WEKA helps us to learn more about the data from analyzing the output result [2].

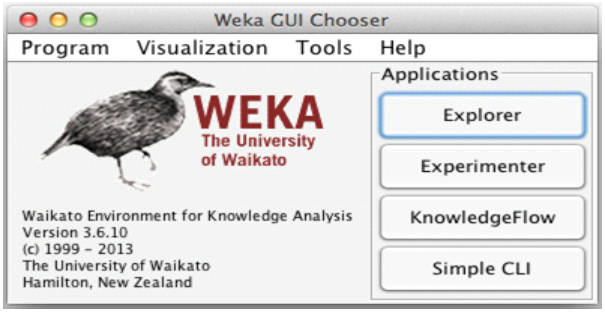

Figure 1: The WEKA tool GUI

Clustering is the main task of Data Mining. And it is done by the number of algorithms. The most commonly used algorithms in Clustering are Partitioning and Density based algorithms.

The GUI Chooser 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.
- 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.

4. LITERATURE SURVEY

There are various authors which have research various existing methods in this field and performed comparison of that clustering algorithms as well.

Raj Bala, Sunil Sikka and Juhi singh et. 2014, 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. After applying normalization to 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 [3].

Bharat Choudhari, Manan Parikh et. 2012, this paper analyze the three major clustering algorithms: K-Means, Hierarchical clustering and Density based clustering algorithm and compare the performance of these three major clustering algorithms on the aspect of correctly class wise cluster building ability of algorithm. After analyzing the results of testing the algorithms, obtain the following conclusions- the performance of K-Means algorithm is better than Hierarchical Clustering algorithm. All the algorithms have some ambiguity in some (noisy) data when clustered. Density based clustering algorithm is not suitable for data with high variance in density. K-Means algorithm is produces quality clusters when using huge dataset. Hierarchical clustering algorithm is more sensitive for noisy data [4].

Deepti V. Patange Dr. Pradeep K. Butey S. E. Tayde 2015, In this paper author presents different clustering techniques and their the comparison using Waikato Environment for Knowledge Analysis or in short, WEKA. After analyzing the results of testing the algorithms we can obtain the following conclusions: The performance of K-Means algorithm is better than EM, Density Based Clustering algorithm, all the algorithms have some ambiguity in some (noisy) data when clustered, K-means algorithm is much better than EM & Density Based algorithm in time to build model [5].

The above are some literatures that is used for the analysis study. Authors have performed the comparison on different clustering algorithms.

5. ANALYSIS OF VARIOUS ALGORITHMS USING WEKA TOOL

Data mining is field of computer science and information technology. As we know that there are lots of data that is available on the web. In that some information is relevant and some are not relevant so when we talk about the relevant data sets we have to mine that lots of data to get the relevant information from the raw data. For the analysis of various clustering algorithms we have taken air pollution dataset. And analysis is performed using four algorithms in the same dataset.

Dataset

A data set is a collection of any type of data. Most commonly a data set corresponds to the information of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable and every row corresponds to a given member of data set in question. There are various dataset like banking dataset, biological datasets, in which clustering can be performed. Here in this dissertation dataset of air pollution is taken [6].

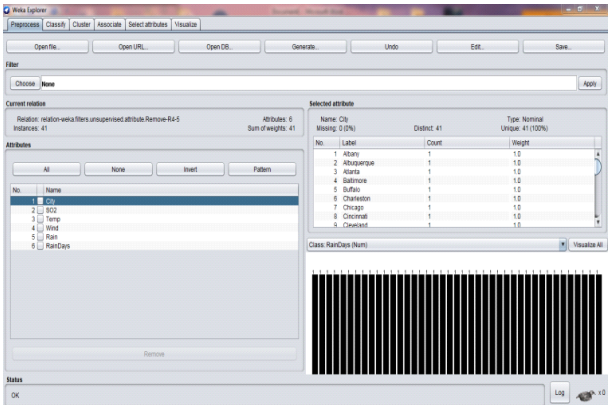

Figure 2: Exploring dataset

In this Research Work analysis of various outlier detection techniques were used those algorithms are:

I. K-Mean Algorithm.
II. LVQ Algorithm.
III. SOM Algorithm.
IV. Cobweb Algorithm

K-Mean Algorithm

K-means clustering algorithm is first proposed by Macqueen in 1967 which was uncomplicated, non-supervised learning clustering algorithm. K-mean is a portioning clustering algorithm. This technique is used to classify given data objects into different k clusters through the iterative method, which tends to converge to a local minimum. So the outcomes of generated clusters are dense and independent of each other [7].

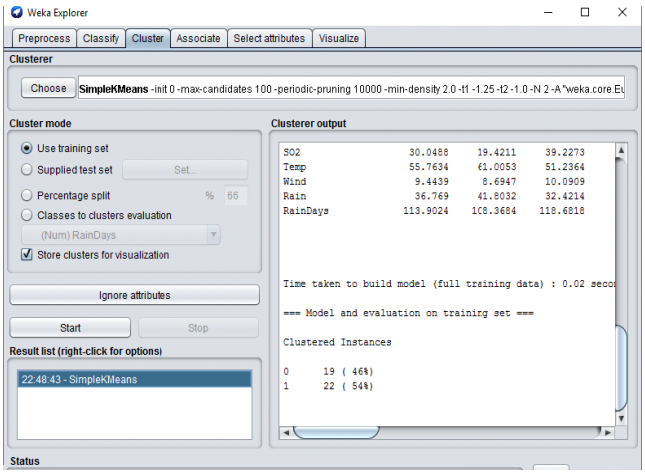

Figure 3: Applying k-mean algorithm

LVQ Algorithm

When classifying linearly separable data by learning vector

quantization (LVQ) or K-Means algorithm (KMA), we cannot necessarily obtain satisfactory classification results for bad selections of initial cluster centers and differences among the distributions of class data. To realize reliable classification, clustering based on multiple criteria for LVQ and KMA is proposed, and its performance is provided. To obtain suitable cluster centers, KMA with the split and merge procedure that is introduced to minimize the squared-error distortion. LVQ using those cluster centers as initial ones is applied to the data, and K clusters are produced [8].

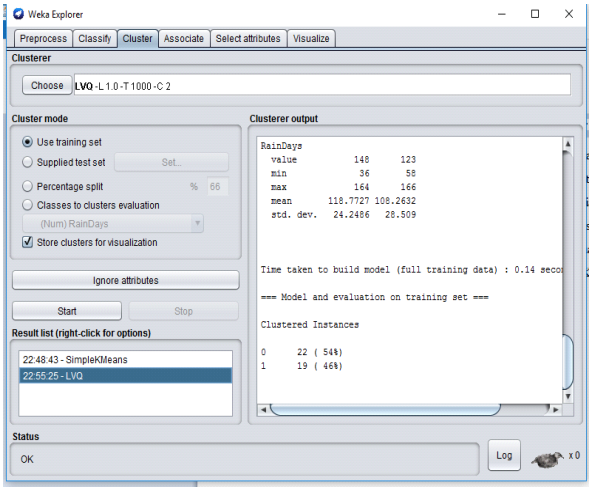

Figure 5: Applying LQV Algorithm

SOM Algorithm

The self-organizing map (SOM) is a tool used in exploratory phase of data mining. It projects input space on prototypes of a low-dimensional regular grid that can be effectively utilized to visualize and explore properties of the data. When the unit of SOM Cluster is large, to facilitate quantitative analysis of the map and the data, similar units need to be grouped, i.e., clustered. In this paper, different approaches to clustering of the SOM are considered. In particular, the use of hierarchical agglomerative clustering and portative clustering using - means are investigated .

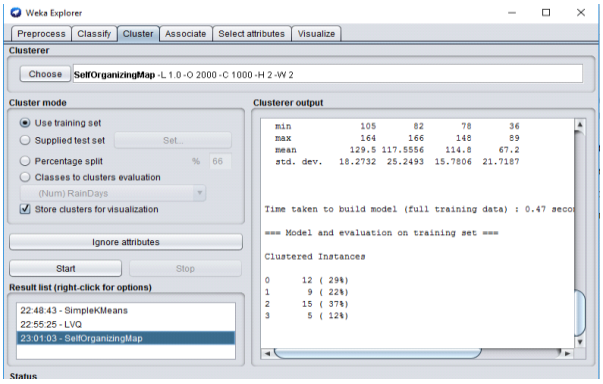

Figure 7: Applying SOM Algorithm

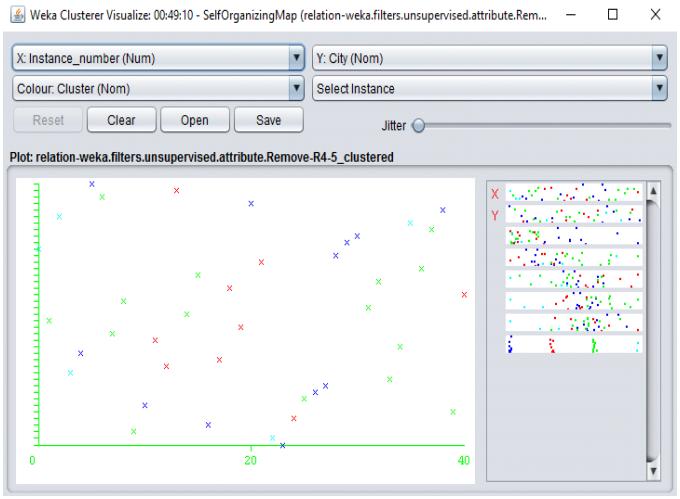

Figure8: Result of SOM in form of graph

Above diagram shows that SOM algorithm applying to data.

COMWEB 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 (concept) and is labeled 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 [10].

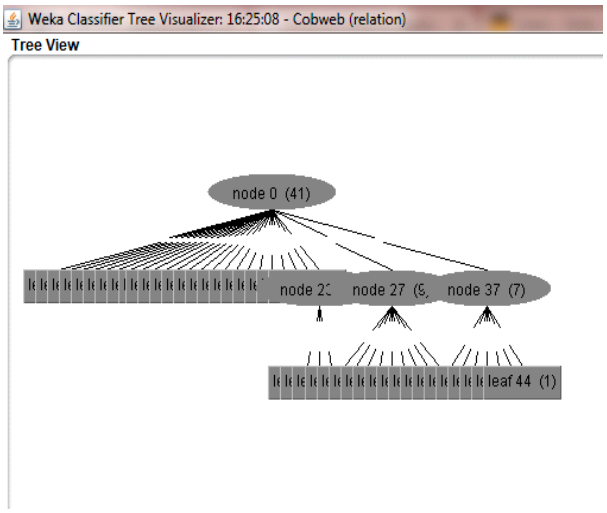

Figure 9: Cobweb tree structure

The above diagram shows the tree structure of the cluster when air pollution dataset is applied to through the weka tool and cobweb clustering algorithm.

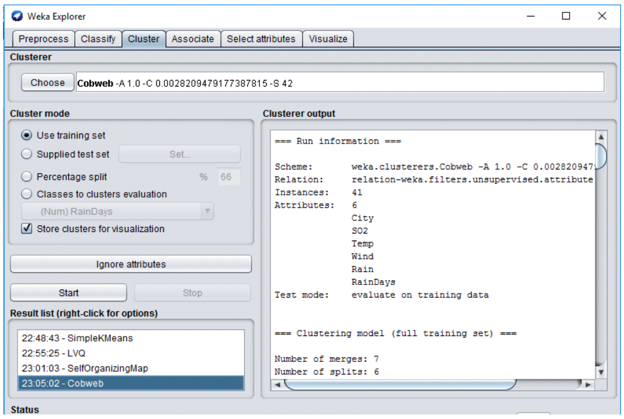

Figure 10: Visualize COBWEB Algorithm

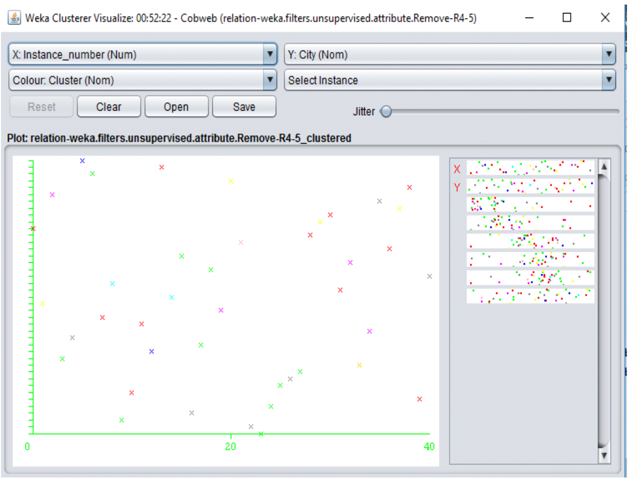

Figure11: Result of cobweb in form of graph

The above diagram shows the visualization of cobweb algorithm through weka tool. And generating results in term of time, no of merges, no of splits and number of clusters.

6. RESULT ANALYSIS

The following table represents the analysis process of all four algorithms:

Table 1: Comparison result of different clustering algorithms

Algorithm	No. of Clusters	Cluster instance	No. of Iteration	Time Taken to build
LVQ	2	19(46%) 22(54%)	-	0.12s
SOM	4	12 (29%) 9 (22%) 15 (37%) 5 (12%)	-	0.33s
K-Means	2	19(46%) 22(54%)	7	0.02s
Cobweb	45	1 (2%) 1 (2%)	-	0.14s

		1 (2%) 1 (2%) 1 (2%) 1 (2%) 1 (2%) 1 (2%)		

7. CONCLUSION

Data mining is the branch of computer science and information technology. There are huge amount of data is spread all over the world and that data is in the form of raw data and raw data contains relevant information also. To get that relevant information mining process is used. From the huge amount of data some similar type of object creates a cluster. We have performed analysis with four clustering algorithms k-mean, LVQ, SOM, and COBWEB. In all four algorithm result is generated on the basis of similar objects and time to create that clusters. Best algorithm found is k-mean clustering. It is taking less time then other clustering algorithm to find similar clusters through weka tool for air pollution dataset.

8. REFERENCES

[1] Chauhan R, Kaur H, Alam M A, "Data Clustering Method for Discovering Clusters in Spatial Cancer Databases", International Journal of Computer Applications , (0975 – 8887) Vol.10– No.6, November 2010.

[2] Data Preprocessing in WEKA, Available at: http://facweb.cs.depaul.edu/mobasher/classes/ect584/weka/preprocess.html.

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

[4] Bharat Choudhari, Manan Parikh et., "A Comparative Study on Role of Data Mining Techniques in Education: A Review" , International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) Web Site: www.ijettcs.org Email: editor@ijettcs.org Volume 3, Issue 3, May – June 2014 ISSN 2278-6856.

[5] 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.

[6] https://archive.ics.uci.edu/ml/datasets/Air+Quality

[7] Z. Huang."Extensions to the k-means algorithm for clustering large data sets with categorical values". Data Mining and Knowledge Discovery,2:283–304, 1998.

[8] http://www.cs.bham.ac.uk/~jxb/NN/l18.pdf

[9] Marie Cottrell, "Some Other Applications of the SOM algorithm : how to use the Kohonen algorithm for forecasting", 2002.

[10] William Iba and Pat Langley. "Cobweb models of categorization and probabilistic concept formation". In Emmanuel M. Pothos and Andy J. Wills,. Formal approaches in categorization. Cambridge: Cambridge University Press. pp. 253–273. ISBN 9780521190480.

[11] Introduction to Weka, Available at: http://transact.dl.sourceforge.net/sourceforge/weka/WekaManual-3.6.0.pdf

[12] Kohonen, T. (1995) : Self-Organizing Maps, Springer Series in Information Sciences Vol 30, Springer.

[13] Kaski, S. (1997) : Data Exploration Using Self-Organizing Maps, Acta Polytechnica Scandinavia, 82.

[14] http://www.cs.ccsu.edu/~markov/ccsu_courses/datamining-ex3.html

[15] Sanjoy Dasgupta ―Performance guarantees for hierarchical clustering Department of Computer Science and Engineering University of California, San Diego.

[16] Ali, MA, Karmakar, GC & Dooley, LS 2008 'Review on Fuzzy Clustering Algorithms'. IETECH Journal of Advanced Computations, vol. 2, no. 3, pp. 169 – 181.

[17] Suganya, R & Shanthi, R 2012 'Fuzzy C- Means Algorithm - A Review'. Int. J. of Scientific and Research Publications, vol. 2, no. 11, pp. 1-3.

[18] Bora, DJ & Gupta, AK 2014 'A Comparative study Between Fuzzy Clustering Algorithm and Hard Clustering Algorithm'. Int. J. of Computer Trends and Technology, vol. 10, no. 2, pp. 108-113.

[19] Glenn Fung, "A Comprehensive Overview of Basic Clustering Algorithms", 2002.

[20] Ossama Abu Abbas., "Comparisons Between of Data Clustering algorithms", The International Arab Journal of Information Technology, Vol. 5, No. 3, 2008.

[21] Madjid Khalilian, Norwati Mustapha, MD Nasir Suliman, MD Ali Mamat, "K-Means Based Clustering Algorithm ", International multi conference of Enginnrs and Computer Scientists, 2010.

[22] Rui Xu, Wunsch, D., II, Dept. of Electr. & Comput. Eng., Univ. of Missouri-Rolla, Rolla, MO, USA, "Survey of clustering algorithms", IEEE Transaction on Neural Networks, 2005.

[23] HE Ling WU Ling-da, CAI Yi-chao(College of Information System & Management ,National University of Defense Technology, Changsha Hunan 410073,China), "Survey of Clustering Algorithms in Data Mining", 2007.