IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 7 Issue 10, October 2020
ISSN (Online) 2348 – 7968 | Impact Factor (2020) – 6.72
www.ijiset.com

Compare Clustering Algorithms of Weka Tool

1Dr. Fatima Dafallah Mohammed Elhassan and 2Dr. Yasir Mohamed Ahamed

1 Professor in King Khalid University, Kingdom of Saudi Arabia, King Khalid University, Faculty of Science and Arts, Majardh, Computer Science Department.

2Assistant Professor in Holy Quran University, The Republic of The Sudan, University of the Holy Quran and Islamic Sciences, Faculty of Computer Science and Information Technology, Computer Science Department

Abstract

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 object of other groups. Different clustering algorithms are present to form clusters. The weka tool, is used to compare different clustering algorithms

In this paper show the comparison of the different clustering algorithms of weka and find out which algorithm will be most suitable for the users, there is the comparison of four clustering algorithms. Are k-means clustering algorithm; COMWEB Algorithm, Canopy Algorithm and hierarchical clustering algorithm All the mentioned algorithms are explained and analyzed based on the certain evaluation parameters. These parameters are the number of clusters created, incorrectly clustered instances, time taken to build the model.

In all four algorithm result is generated on the basis of similar objects and time to create that clusters. The Best algorithm found his Canopy clustering. It is taking less time than other clustering algorithm to find similar clusters through weak tool for student dataset

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

1. INTRODUCTION

Data mining is the use of automated data analysis techniques to uncover previously undetected relationships among data items. Data mining often involves the analysis of data

stored in a data warehouse. one of the major data mining techniques are regression, classification and clustering.

Clustering is a technique of natural grouping of data objects which are unlabeled and it forms these grouping in such a way that data objects belonging to one cluster are not similar to the objects belonging to another cluster [1]. Clustering or cluster analysis is one of the most essential and important unsupervised learning technique. It covers the three well-known categories of cluster analysis namely partition, hierarchical and density-based clustering.

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 and we can work in weka easily without having the deep knowledge of data mining techniques. In this paper, there is the comparison of four clustering algorithms. The algorithms considered for an experiment in this study are k-means clustering algorithm; COMWEB Algorithm, Canopy Algorithm and hierarchical clustering algorithm. All the mentioned algorithms are explained and analyzed based on the certain evaluation parameters. These parameters are the number of clusters created, incorrectly clustered instances, time taken to build the model.

This paper organized with sex section, section 1 Introduction section 2 clustering algorithms and WEKA tool section 3 describes the basis for algorithm comparison , section 4 literature survey, Section 4 shows the experiment and results and section 5 concludes the paper.

2. Cluster analysis and WEKA TOOL

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including pattern recognition, image analysis, information retrieval, bioinformatics, data compression, computer graphics and machine learning.

Besides the term clustering, there are a number of terms with similar meanings, including automatic classification, numerical taxonomy, bornology, typological analysis, and community detection. The subtle differences are often in the use of the results: while in data

mining, the resulting groups are the matter of interest, in automatic classification the resulting discriminative power is of interest. [2]

WEKA TOOL

WEKA is graphical user interface (GUI), it's an open source software developed at Waikato University in New Zealand. It contains four applications; explorer, experimental, knowledge flow and the command line interface (CLI) and also contains tools for data pre-processing, classification, clustering, regression and visualization. The pre-processing is an important step that is used to extract and improve the quality of data. WEKA tool import dataset from a proper file like attribute relation file format which is the preferable one. Figure 2 and Figure 3 show the output of data pre-processor and model visualization in WEKA, respectively.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 [3].

Figure 1: The WEKA tool GUI

3. Cluster ALGORITHMS USING WEKA TOOL

Clustering is a task for which many algorithms have been proposed. No clustering technique is Universally applicable, and different techniques are in favour for different clustering purposes. So an understanding of both the clustering problem and the clustering technique is required to Apply a suitable method to a given problem. In the following, I describe general of a Clustering technique algorithms in weak tool show in figure2.

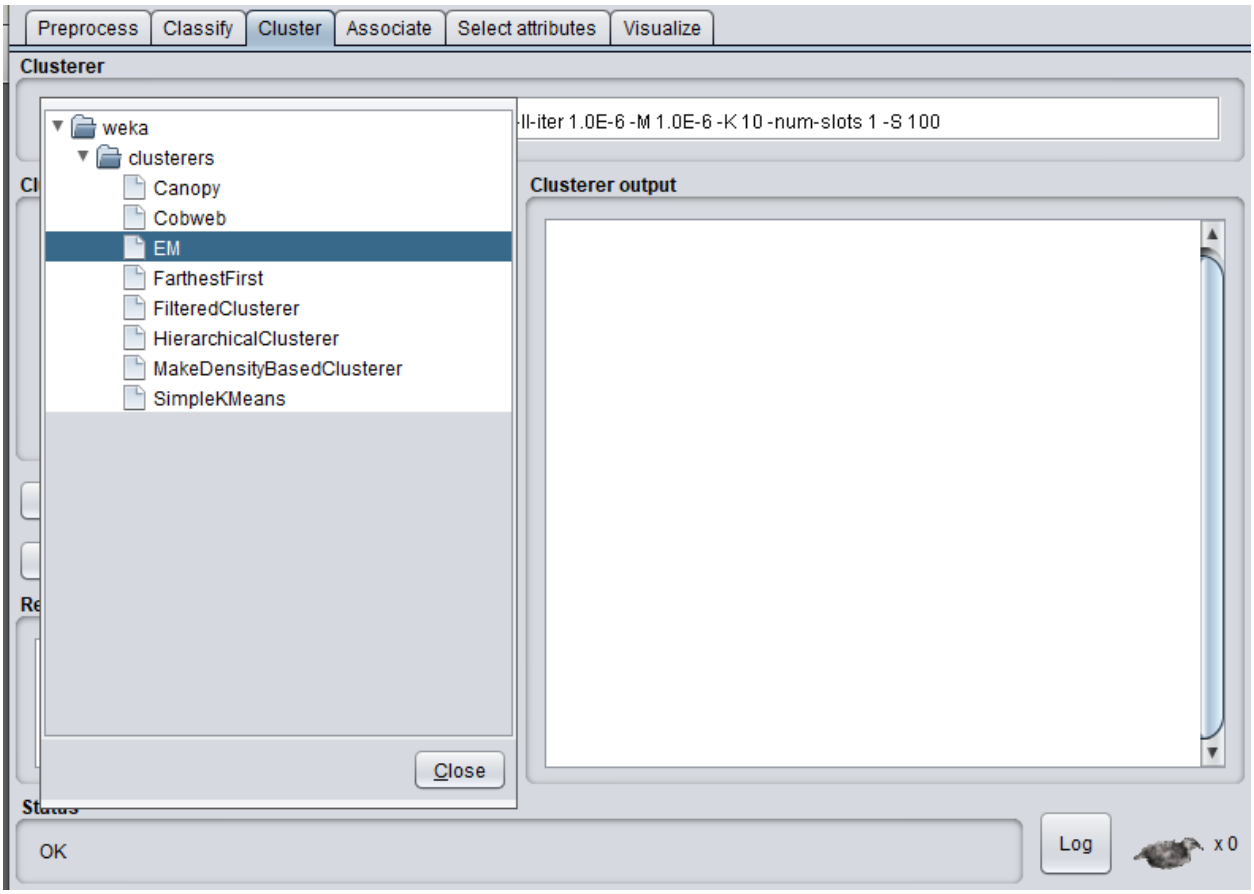

Figure 2: The clustering of WEKA tool

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 [4].

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 [5].

Canopy Algorithm

Canopy Clustering is a very simple, fast and surprisingly accurate method for grouping objects into clusters. All objects are represented as a point in a multidimensional feature space. The algorithm uses a fast approximate distance metric and two distance thresholds T1 > T2 for processing. The basic algorithm is to begin with a set of points and remove one at random. Create a Canopy containing this point and iterate through the remainder of the point set. At each point, if its distance from the first point is < T1, then add the point to the cluster. If, in addition, the distance is < T2, then remove the point from the set. This way points that are very close to the original will avoid all further processing. The algorithm loops until the initial set is empty, accumulating a set of Canopies, each containing one or more points. A given point may occur in more than one Canopy. Canopy Clustering is often used as an initial step in more rigorous clustering techniques, such as K-Means Clustering . By starting with an initial clustering the number of more expensive distance measurements can be significantly reduced by ignoring points outside of the initial canopies [6].

HierarchicalClusterer

Hierarchical clustering is mainly focuses on building of hierarchy of clusters, i.e. cluster tree and it is represented in a dendrogram [7]. It is either merging smaller clusters into larger clusters or splitting larger clusters into smaller ones. A clustering of the data items is obtained through cutting a dendrogram at a desired level [8]. A cluster tree is defined as "a tree showing a sequence of clustering with each clustering being a partition of the data set" [9]. Following are the general procedures for performing hierarchical clustering [10].

4. LITERATURE REVIEW

Many research studies have been done in educational data mining to predict the students' performance

In [11], in this paper author presents different clustering techniques and their comparison using Waikato Environment for Knowledge Analysis or in short, WEKA. After analysing 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.

In [12] analyse 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 analysing 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].

In [13] 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].

In [14], association, classification, clustering and outlier detection data mining techniques were applied to analyse 3,314 graduate student performance records over a fifteen-year period. The dataset was analysed using Rule Induction, Naïve Bayesian classifier, K-Means clustering algorithm followed by density-based and distance-based outlier detection methods. 18 attributes of the student dataset were considered, and only 6 attributes: matriculation GPA, gender, specialty of the students, the city of the student, the grade and the type of secondary school attended were selected for the data mining analysis. The remaining 12 attributes were dropped due to their large variances and because some of the attributes are personal information that did not provide useful knowledge.

In [15] the unsupervised clustering analysis performed, identified four unique clusters in the dataset using k-means algorithm. Data mining method was applied by to evaluate student data towards identifying the key attributes that influence the academic performance of students. This provides an opportunity for improving the quality of higher education.

In [16], Dall et al. discussed Clustering techniques and divide them into three major categories: Partitioned Clustering, Density based Clustering and Hierarchical Clustering which are further subdivided

According to [17] Lu et al., when the data is large and continuous in nature then the traditional approaches of mining are not applicable because the real time data is quickly changing and

requires Fast response as well. Random access to data stream is very expensive so, a single access to streaming data is provided. Also the storage needed is very large. Therefore, clustering and mining techniques for stream data are required.

In [18] Fahad et al., The effectiveness of the candidate clustering algorithms is measured through a number of internal and external validity metrics, stability, and run time and scalability tests. Document Big volume of data or big data has its own deficiencies as it needs big storages and this volume makes operations such as analytical operations, process operations, retrieval operations, very difficult and hugely time consuming. To overcome these difficult problems big data is clustered in a compact format that is still a informative version of the entire data. DENCLUE, OptiGrid and BIRCH are suitable clustering Algorithms for dealing with large datasets, No clustering algorithm performs well for all the evaluation criteria, and future work should be dedicated to accordingly address the drawbacks of each clustering algorithm for handling big data.

In [19] examined the DM process in student's dataset using clustering and classification techniques. By using k-Means algorithms and decision tree, they analysed different factors that affect a student's learning behaviour and performance throughout his/her academic career in a higher educational institution.

In [6], the k-means clustering algorithm applied to predict the student performance using student's semester results. They divided the data into various sets of clusters and attained 60% accuracy results. This study relies on only academic marks with ignoring other important factors like social factors, which play an important role in academic performance for students

in [7], the data clustering algorithms of x-means used the final student's results to predict the graduation performance by focusing on the final year's marks in university as an affecting factor is not enough to predict student performance, which needs to study more factors to enhance the obtained results

in [22], the work conducted using a k-means clustering algorithm with the Elbow method to choose an appropriate number of clusters to analyze the relationship of academic marks and the gender type on the academic performance of students in the MCA course (a postgraduate level program in information technology). However, the gender factor cannot be considered an influencing factor since both genders can have common skills that have the same effect on academic performance

5. Experiments and Results

Data set

The data set name is Student_IT_GPA-WEKA .CSV this consist of the following 4 feature.[first year GPA, second year GPA ,third year GPA,class of degree] From University of Science & Technology Faculty of Engineering with 1329 instance see figure 3.

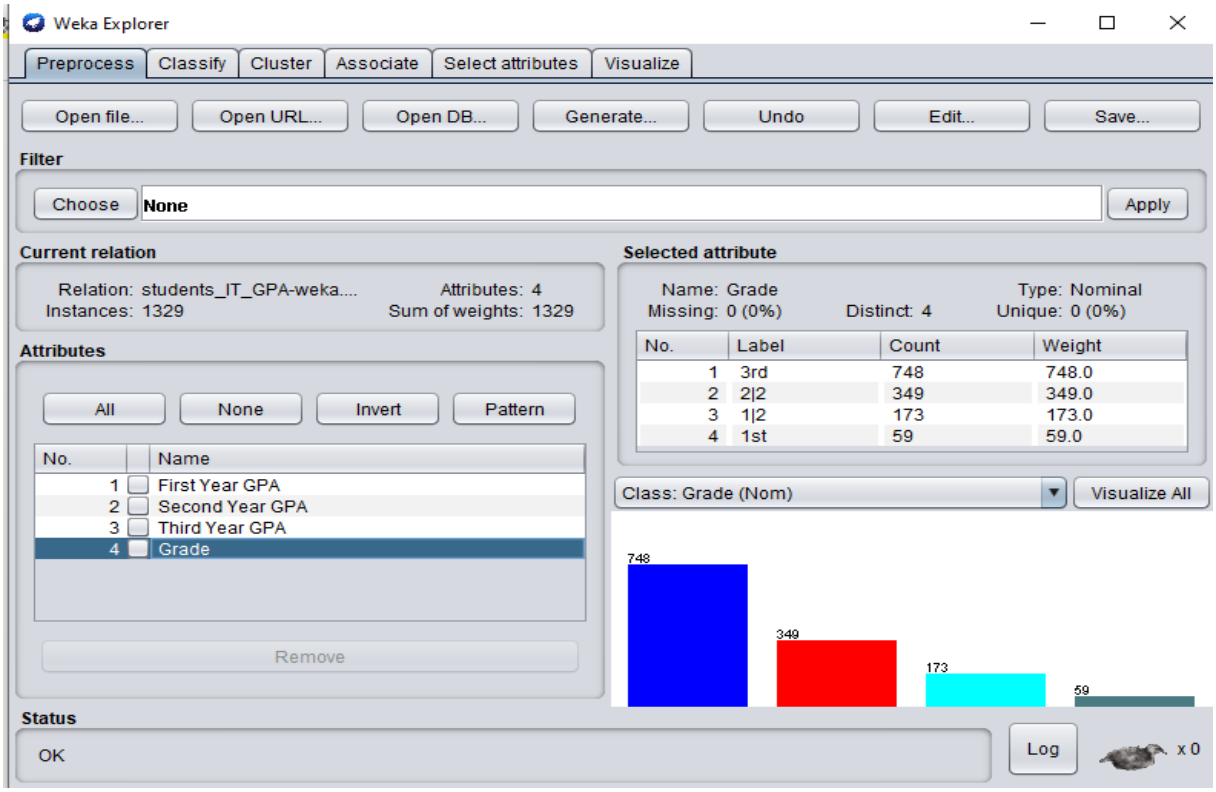

Figure.3.data set

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

I. K-Mean Algorithm.

II. COMWEB Algorithm.

III. Canopy Algorithm.

IV. HierarchicalClusterer

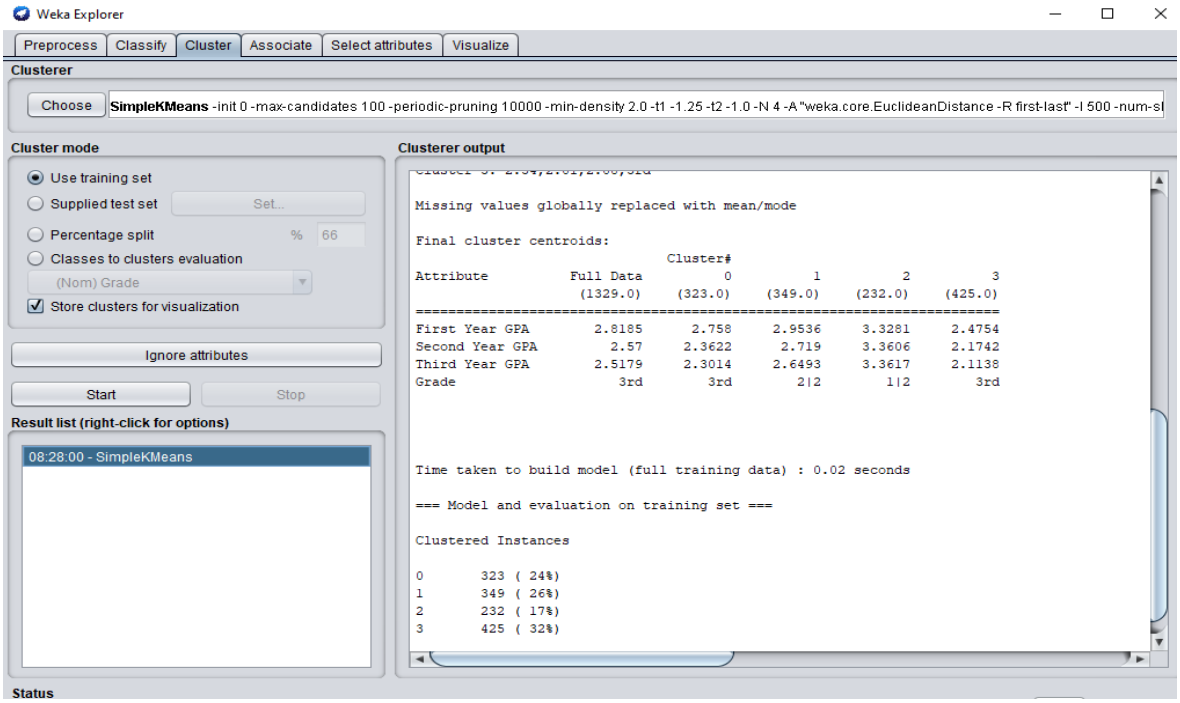

Figure.4. Applying k-mean algorithm

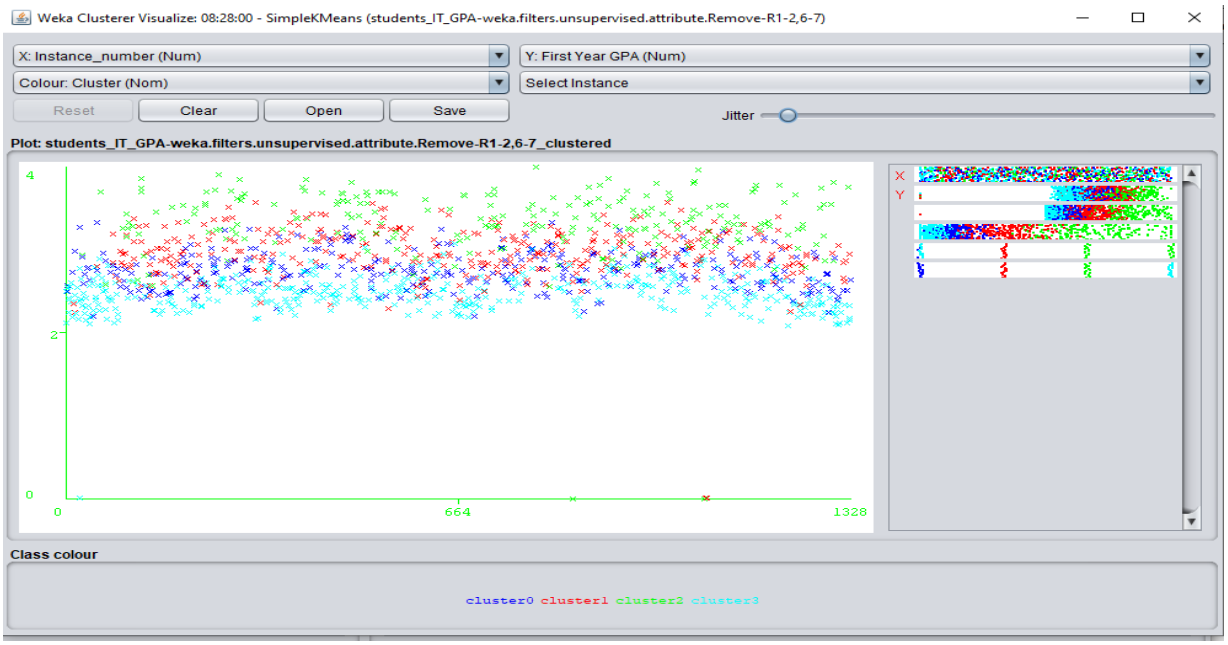

Figure.5. Visualize k-mean algorithm

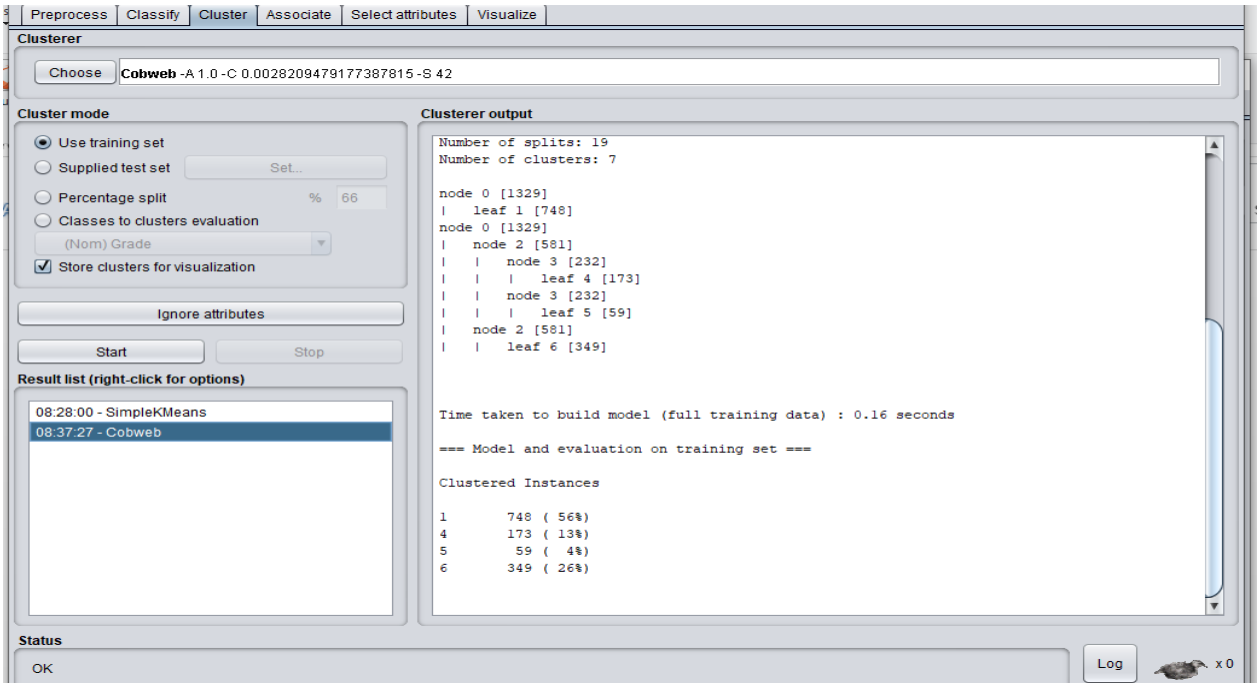

Figure.6. Applying COBWEB algorithm

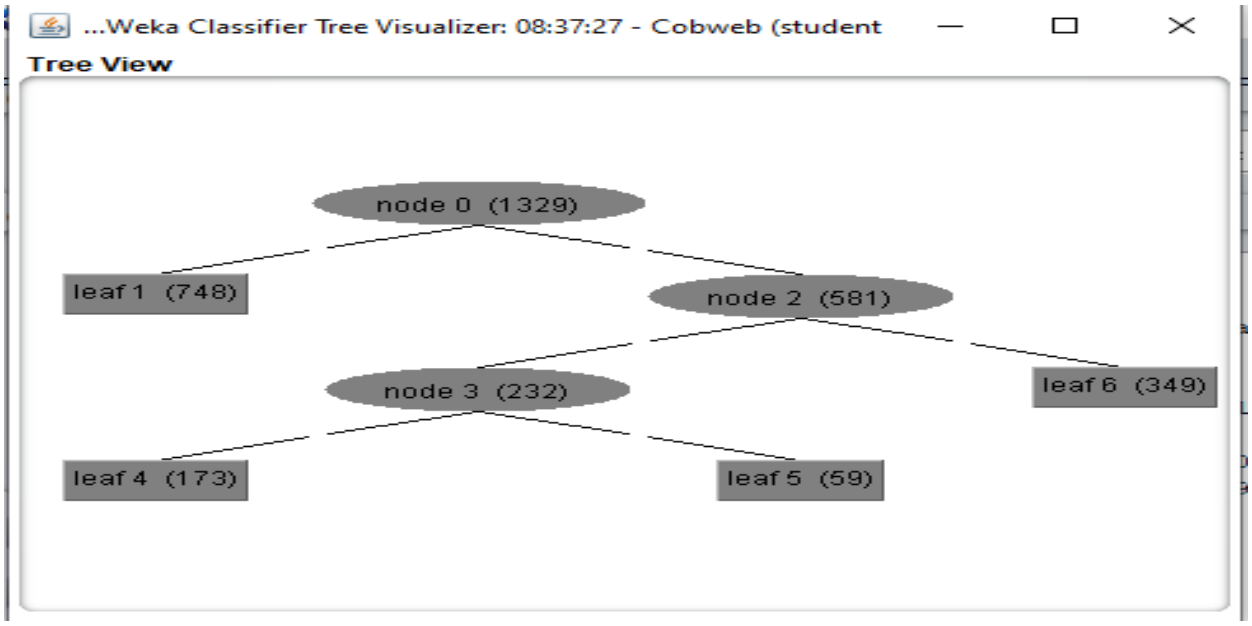

Figure.7. COBWEB tree structure

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.

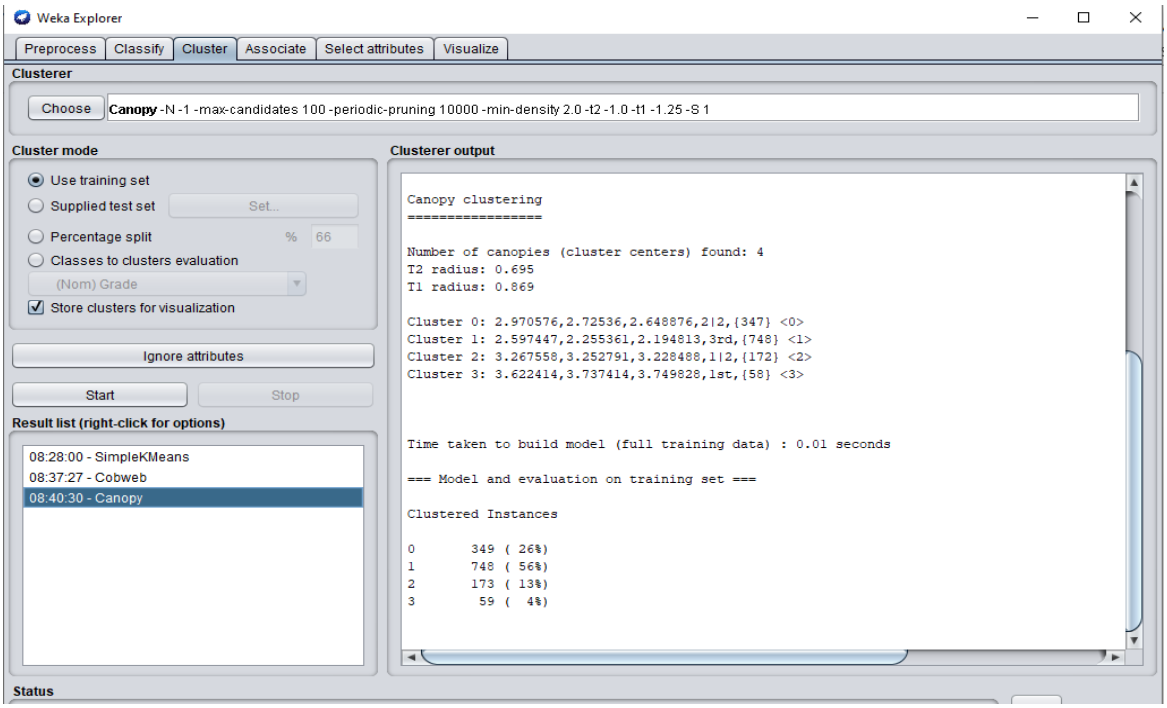

Figure.8. Applying CANOPY algorithm

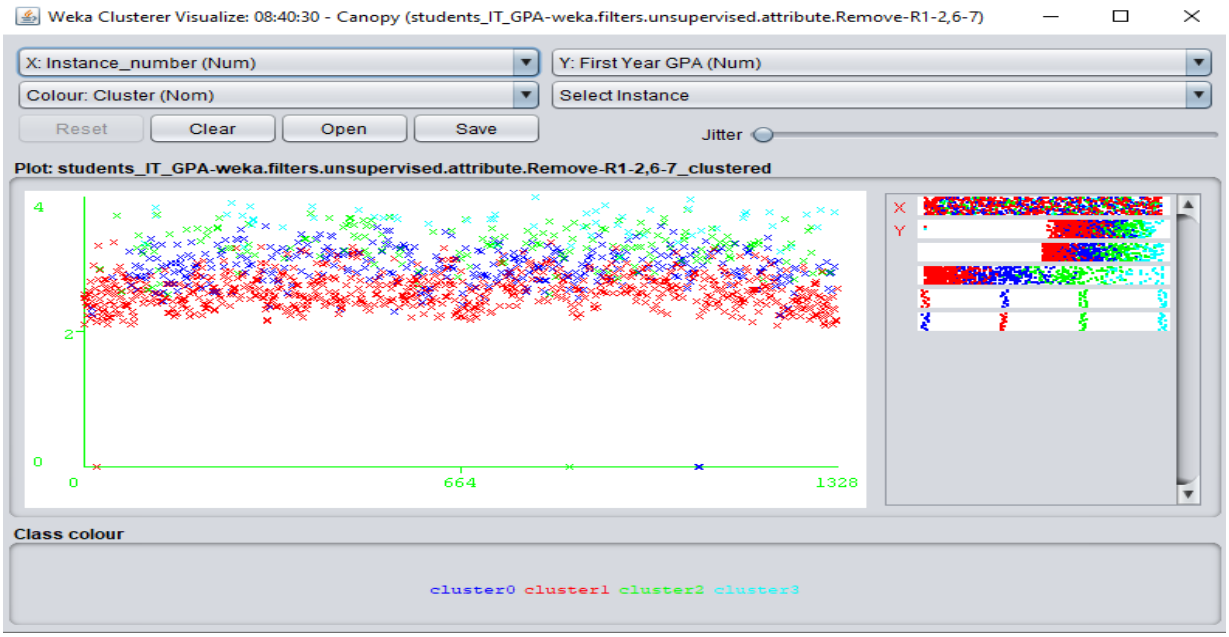

Figure.9. Visualize CANOPY algorithm

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

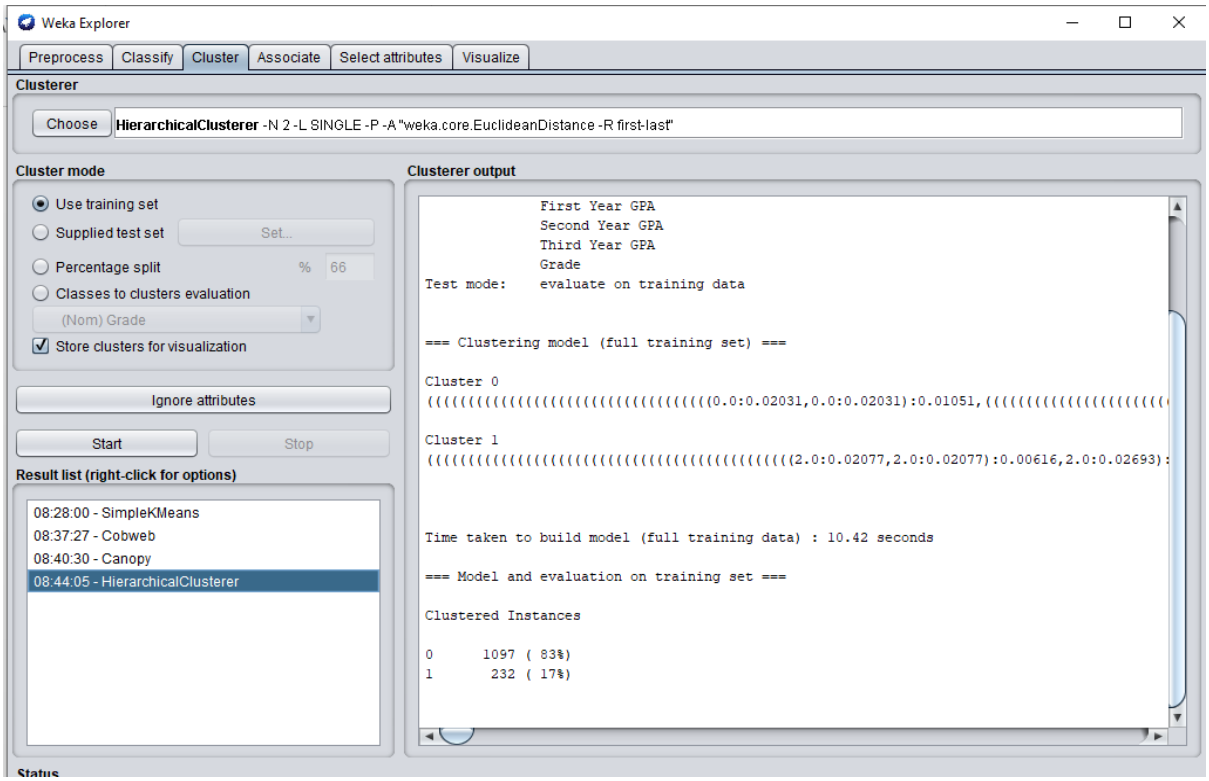

Figure.10. Applying HierarchicalClusterer algorithm

6. RESULT ANALYSIS

The following table represents the analysis process of all four algorithms

Table 1: Comparison result of different clustering algorithms

clustering algorithm	No. of Clusters	Cluster instance	Time taken to build model
K-Mean	4	323 (24%) 349 (26%) 232 (17%) 425 (32%)	0.02 seconds
COMWEB	7	748 (56%) 173 (13%) 59 (4%) 349 (26%)	0.16 seconds
Canopy	4	349 (26%) 748 (56%) 173 (13%) 59 (4%)	0.01 seconds
Hierarchical	2	1097 (83%) 232 (17%)	10.42 seconds

In table 1, the canopy clustering algorithm it is taking less time than other clustering algorithm. But Hierarchical clustering algorithm it is taking more time than other clustering algorithm to find similar clusters through weka tool for student dataset.

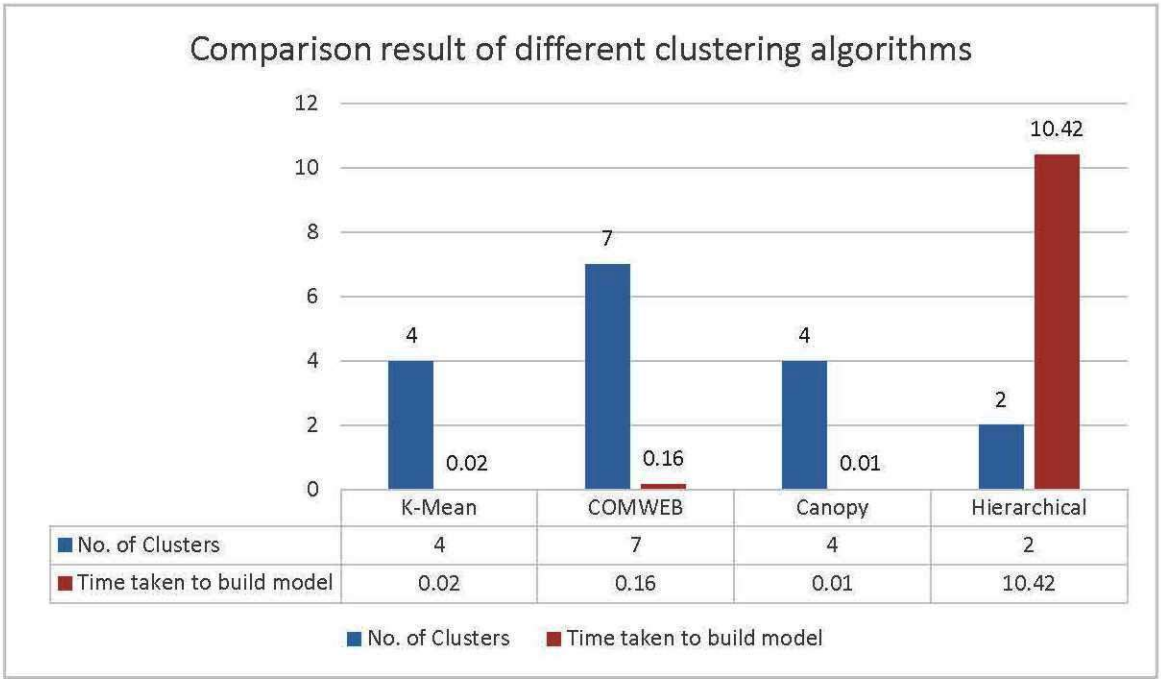

Figure.11.comparison of clustering algorithm

In Figure.11 shows the performance accuracy of the four clustering based on different clustering metrics. These metrics are; (no of clustering), (time to build model), these metrics shows that canopy clustering algorithm performs better than other clustering.

7. CONCLUSION

We are using data mining techniques in mainly in the medical, banking, insurances, education etc. Before start working in the with the data mining models, it is very necessary to knowledge of available algorithms from the huge amount of data some similar type of object creates a cluster. We have performed analysis with four clustering algorithms k-mean, Canopy, **H**ierarchical, 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 Canopy clustering. It is taking less time than other clustering algorithm to find similar clusters through weak tool for student dataset.

8. REFERENCE

1. Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012", Springer Nature America, Inc, 2014
2. https://en.wikipedia.org/wiki/Cluster_analysis,22102020:8:58

3. Yasir M.A, Fatima D. M. ,2020, predict the grad of student using classification algorithms, International Journal of Science, Environment and Technology, Vol. 9, No 2, 2020, 75 – 89 .
4. 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.
5. 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
6. https://mahout.apache.org/docs/latest/algorithms/clustering/canopy/2210202:10:05
7. Zhao Y., Karypis G., "Evaluation of hierarchical clustering algorithms for document datasets", the eleventh international conference on Information and knowledge management,2002, pp. 515-524.
8. Y. Leung, J. Zhang and Z. Xu, "Clustering by Space-Space Filtering", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no.12, pp. 1396-1410, 2000
9. Jain A.K., Murty M.N., Flynn P.J., "Data clustering: A Review", ACM Computing Surveys, Vol. 31, No. 3, September 1999
10. Aarya Vardhan Reddy, Paakaala Sai Saran Macha, Kumara Saketh Mudigonda, "Evaluation of Clustering Algorithms on Absenteeism at Work Dataset", IJSRD - International Journal for Scientific Research & Development| Vol. 6, Issue 06, 2018 ISSN (online): 2321-0613 , pp. 337-342
11. 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.
12. 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.
13. 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.
14. Tair, M.M.A., El-Halees, A.M., 2012. Mining educational data to improve students' performance: a case study. Int. J. Inf. Commun. Technol. Res. 2 (2), 140e146.
15. Al-Radaideh, Q., Al-Shawakfa, E., Al-Najjar, M., 2006. Mining student data using decision trees. In: Paper Presented at the the 2006 International Arab Conference on Information Technology. ACIT'2006.
16. A. Fahad, N. Alshatri, Z. Tari, A. Alamri, A. Zomaya, I.Khalil, S. Foufou, A. Bouras , A survey of ClusteringAlgorithms for Big Data: Taxonomy and Empirical Analysis,IEEE 2014,267 - 279
17. Y. Hong Lu, Y. Huang, Mining Data Streams using Clustering,Proceedings of the Fourth International Conference on Machiene Learning and Cybernetics, Guangzhou, 18-21August, 2005, IEEE 2005,pp. 2079 - 2083

18. M. A. Dalal, N D Harale, A survey on Clustering in data mining, International Conference and Workshop on Emerging Trends in Technology, TCET, Mumbai, India, ACM 2011, pp.559-562 .
19. A. R. Chordiya and S. B. Bagal, "Comparative Research of Clustering Algorithms for Prediction of Academic Performance of Students." .
20. J. Manoharan, S. H. Ganesh, M. L. P. Felciah, and A. K. S. Banu, "Discovering students' academic performance based on GPA using K-means clustering algorithm," in Proceedings - 2014 World Congress on Computing and Communication Technologies, WCCCT 2014, pp. 200–202.
21. R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "Analyzing undergraduate students' performance using educational data mining," Comput. Educ., vol. 113, pp. 177–194, Oct. 2017.
22. T. Devasia, T. P. Vinushree, and V. Hegde," Prediction of students performance using Educational Data Mining," in Proceedings of 2016 International Conference on Data Mining and Advanced Computing, SAPIENCE 2016, 2016.
23. Confluence Veranstaltung 6. 2016 Noida et al., Proceedings of the 2016 6th International Conference Cloud System and Big Data Engineering (Confluence) 14 -15 January 2016, Amity University, Uttar Pradesh, Noida, India. IEEE, 2016.
24. S. Lailiyah, E. Yulsilviana, and R. Andrea, "Clustering analysis of learning style on anggana high school student," TELKOMNIKA (Telecommunication Comput. Electron. Control., vol. 17, no. 3, p. 1409, Jun. 2019.
25. N.Valarmathy and S.Krishnaveni "Performance Evaluation and Comparison of Clustering Algorithms used in Educational Data Mining." vol. 7, Apr. 2019.