Information Sciences Letters

Volume 9
Issue 1 *Jan. 2020*

Article 4

2020

Analysis of Different Clustering Algorithms for Accurate Knowledge Extraction from Popular DataSets

Shamveel Hussain Shah
Department of Computer Science, University of Engineering and Technology, Taxila, Pakistan, javed.iqbal@uettaxila.edu.pk

Muhammad Javed Iqbal
Department of Computer Science, University of Engineering and Technology, Taxila, Pakistan, javed.iqbal@uettaxila.edu.pk

Muhammad Bakhsh
Department of Computing, Abasyn University Peshawar and Pakistan Academy for Rural Development Peshawar, Pakistan, javed.iqbal@uettaxila.edu.pk

Amjad Iqbal
Department of Computer Science, HITEC University Taxila, Pakistan, javed.iqbal@uettaxila.edu.pk

Information Sciences Letters
An International Journal

http://dx.doi.org/10.18576/isl/090204

Analysis of Different Clustering Algorithms for Accurate Knowledge Extraction from Popular DataSets

Shamveel Hussain Shah [1], *Muhammad Javed Iqbal* [1,*], *Muhammad Bakhsh*[2] *and Amjad Iqbal*[3]

[1]Department of Computer Science, University of Engineering and Technology, Taxila, Pakistan
[2]Department of Computing, Abasyn University Peshawar and Pakistan Academy for Rural Development Peshawar, Pakistan
[3]Department of Computer Science, HITEC University Taxila, Pakistan

Received: 21 Nov. 2019, Revised: 22 Dec. 2019, Accepted: 24 Dec. 2019.
Published online: 1 Jan. 2020.

Abstract: Data mining is a method to mine valuable hidden knowledge, patterns and associations from massive and sparse datasets. This process proceeds through various techniques e.g. classification, clustering and association etc. Clustering is an important data mining technique which group similar data items together in a group. In this study comparison is performed with six different clustering techniques using six different datasets. Comparison was performed on the basis of different evaluation parameters. By overall results it is concluded that k-Mean algorithm is best, simplest, produced quality clusters and has high performance amongst all other five algorithms. Performance of EM algorithm is worst amongst all other five algorithms as it took more time to produce inaccurate results. Hierarchical algorithm is best on small datasets but on huge datasets it took more time. Performance of density based Clustered and Canopy algorithm is almost same with slight difference in results. We also compared our study results with existing results and proved that proposed results are quite reasonable and accurate. Our research analysis and results make better understanding for cluster researcher to improve existing techniques and also to analyze more techniques and to propose a new clustering technique.

Keywords: Data Mining, Clustering, k-Mean Clustering, Hierarchical Clustering, EM Clustering, Make Density Based Clustering, Farthest First Clustering, Canopy Clustering**.**

1 Introduction

Data mining is a method to mine valuable hidden knowledge, patterns and associations from massive and sparse datasets. This process proceeds through various techniques e.g. Classification, Clustering and association rules etc. In this research, only clustering techniques have been discussed and analyzed. Clustering is an important and primary data mining technique [1] which group similar data items together in a group [1-5]. It is mainly used for data analysis purpose and also in different data mining applications e.g. in pattern detection, text mining, web analysis, information retrieval, marketing and medical diagnostic etc. [6]. It is an important technique for extraction of correct and accurate results from sparse multidimensional datasets [7, 8].

In general, clustering techniques be can categorized into partition-based, hierarchical-based, density-based algorithms [9,10].Partitioned based clustering algorithm partition the data points into k parts, where each part denotes a cluster[1]. k-Mean is one of partition-based algorithm, where mean value of cluster objects represents midpoint of each cluster [11]. Hierarchical clustering technique divides the dataset by building a hierarchy or tree of clusters [11]. Density-based algorithms build clusters with respect to high density regions [10]. Canopy algorithm is mostly used as preprocessing step for k-Mean and Hierarchical algorithms [12]. It is simple and used to speed up clustering process for large datasets. Farthest First algorithm partitions a large dataset into k-clusters where each cluster has a center point and Farthest First algorithms tries to minimize the maximum distance from any point to center point [12]. EM algorithm is an iterative method frequently used to find log likelihood and to estimate parameters for statistical methods [12].

Generally clustering is learning without a teacher because for a space having n number of samples, no true class labels are available for each sample which makes it harder as

*Corresponding author e-mail: javed.iqbal@uettaxila.edu.pk

compared to supervised learning. Such situation arises issue that how do we recognize the significance of results, when there is no availability of answer labels. For that external and internal evaluation is required to be performed separately. There are also some other issues related to clustering that are discussed here in brief. In clustering many statistical data applications would not be enough, as new techniques are needed for the analysis of uncertain data in fast and more precise way. In large databases, without supervision clustering methods show less control to handle complex clustering tasks since data complexity can be increased by increasing number of dimensions of data. Algorithms required assistance of an expert to access the density and number of expected partitions. Compactness and data separation are the main problems of quality clustering. Efficiency in term of speed and to detect concept drift in accuracy are serious problems for data mining clustering. Many existing approaches lack accuracy in detecting and identifying outliers. To identify number of clusters is a difficult task when the number of class labels are unknown so, a thorough analysis of number of clusters is required to produce quality results. Otherwise, similar tuples can be divided into many tuples and diverse tuples can merge together. In hierarchal based approach this situation could be catastrophic, because if some tuples incorrectly merge with each other in a cluster then such action can't be reversed. All datasets do not contain same type of attributes e.g. categorical or nominal, but they also contain other type of attributes e.g. binary, ordinal etc. so there is a need to convert other type of attributes in categorical or nominal type to make calculation easier. In partitioned based approach many algorithms randomly select initial k clusters, so a comprehensive and precise overview of data is essential. Otherwise, empty clusters will be obtained after little iteration as a result of improper selection of initial clusters. Most of these issues related to datasets have been resolved in preprocessing stage to improve clustering results.

The purpose of this research is to analyze and evaluate some of the important data mining clustering techniques and to compare them on the basis of 'Time taken to build model', 'Correctly classified instances', 'Incorrectly classified instances', 'Root mean squared error' using different datasets. In study six different datasets are utilized to perform analysis of six different clustering techniques ('Canopy', 'EM', 'Farthest First', 'Hierarchical Clusterer', 'Make density based Clusterer', 'simple k-Means') using different evaluation parameters. Experimental results, key findings and analysis have been discussed in research to show which technique is better amongst others. Our analysis and results really make better understanding for researcher to improve existing techniques and to analyze more techniques and to propose a new clustering technique.

The organization of paper is as follows: Section 2 describes literature review, Section 3 explains proposed methodology and Section 4 elaborates results, analysis and findings. The last Section provides conclusion, recommendations and future work.

2 Literature Review

In literature study, many research articles related to our topic are reviewed. Some researchers tried to improve existing clustering techniques, some of them proposed new ones and others compared and analyzed the existing techniques. A brief summary of few recent articles is given in the following subsection.

Popat and et.al proposed that the aim of cluster analysis is to find similar patterns [1]. Authors surveyed different clustering techniques in their research. Authors divided techniques into these categories: Partitioned algorithms, Hierarchical algorithms, Density based. Authors performed comparison of different algorithms and by results showed that amongst other algorithms Hierarchical Clustering is better. In conclusion authors compared each clustering category with its pros and cons and also discussed the concept of Similarity measures the most important criteria for document clustering.

Chaudhari et al. discussed that clustering is a practice to put same type of data into clusters [2]. In the research authors analyzed three main clustering techniques: k-Means, Hierarchical-based clustering and Density-based clustering algorithm. Authors evaluated performance of each algorithm based on their ability to build class wise clusters correctly using a data mining tool Weka. After analyzing the results authors concluded that: The performance of k-Means technique is superior over Hierarchical-based clustering technique. Density-based clustering technique is not appropriate for data having great inconsistency in density. Hierarchical-based clustering technique is more sensitive for noisy data.

Kumar et al. presented that clustering finds an arrangement from the group of data having no labels [3]. Authors analyzed four key clustering techniques e.g. Partition-based techniques, Hierarchical-based techniques, Grid-based techniques and Density-based techniques. Authors also compared efficiency of these techniques based on their ability to build class wise clusters correctly. Authors concluded that: while using hierarchical-based technique, a process cannot be reversed once it's been completed. In case of partitioning technique, different statistical measures are used like mean, median and mode. Grid-based technique, construct grids of unlike sizes. Density-based technique is appropriate only for illogical shape data.

Chaudhary et al. presented study of three different Density-based Clustering methods including DBSCAN, DBCLASD and DENCLUE [4]. For comparison and evaluation of experimental results authors used six different evaluation parameters. By results authors concluded that DBSCAN method has lowest running time while DENCLUE has highest running time. Similarly, cluster quality of DBCLASD method is higher while cluster

quality of DENCLUE method is lowest. Their study provides help to find a suitable Density-based method for a certain situation among different situations.

Leela and et al. discussed that clustering is an unsupervised learning process which generates group of similar data objects[5]. Authors analyzed various type of clustering techniques e.g. fuzzy c-mean, k-Means, Subtractive clustering and Mountain clustering using a dataset of iris flowers. Authors compared algorithms on the bases of time complexity, accuracy and run time using MATLAB tool. Then authors proposed a new and better Y-mean method to improve experimental clustering results. Results showed that newly proposed Y-mean method performed better in comparison to other clustering methods having low execution time.

Baser et al. described that data mining is method to extract secreted information, valuable patterns and drifts from huge datasets[6]. To perform such type of tasks there, exist several data mining techniques e.g. classification, clustering, clustering, outlier analysis etc. In research, authors performed comparison of different clustering techniques specifically. Authors concluded that each clustering method has some advantages as well as disadvantages and is useful in certain situation. Currently no such method exists, which solely provide solution for each and every situation.

Mishra et al. explored that clustering method divides dataset in to different clusters such that clusters have similarities [7]. Authors presented comparison between some common document clustering techniques. In particular, they compared: k-Means, Fuzzy c-means, Mountain and Subtractive clustering. From result authors concluded that k-Mean method is better than other methods. Authors also found that high dimensional data creates problems for algorithms to find relationship among variables of data. In future authors want to propose a novel method for clustering to improve accuracy and performance for high dimensional datasets.

Prabha et al. suggested that clustering process is a supportive task to retrieve accurate and effective information in efficient manner [8]. In research authors surveyed few clustering algorithms and analyzed working principle of algorithms with the help of data samples. Furthermore, authors executed experimentation on some UCI repository datasets in order to access the quality of each clustering algorithm. From result authors concluded that there is a need to improve level of clustering of few algorithms. However, in future authors will try to improve the quality of existing methods.

Sheshasayee et al. disclosed that in requirement engineering gathering of requirements from respective persons is very essential [9]. Issues arise when collected set of requirements are numerous and engineers can't focus on specific requirements. Author's research mainly focuses to gather effective requirements from the collected set of requirements using various clustering methods. In research authors used two clustering methods e.g. k-Means and Fuzzy c-means. Then performance of method has been evaluated based on output. Authors concluded that Fuzzy c-mean method is proficient for huge datasets and is useful for requirement engineering clustering.

Singh et al. explained that in data mining hidden and unknown links, patterns and associations are explored from huge datasets [10]. In article authors performed comparison between nine different clustering analysis techniques using Weka tool. Authors compared performance in terms of 'execution time', 'number of iterations', 'sum of squared error' and 'log likelihood'. Finally, from result authors made conclusion that k-Means method is simple in comparison to other methods and also performance of k-Mean is superior to Hierarchical-based method. Density-based method is not appropriate for data containing vast differences in density. Hierarchical-based method is more vulnerable to noisy data. Both k-Mean and density-based methods are superior to EM method.

N. Valarmathy et al. surveyed the application of various well known data mining clustering algorithms to traditional educational systems [11].The major idea of authors study is to make a detailed survey on different kind of clustering techniques (e.g. various hierarchical methods, partitioned methods and density-based methods) its advantages, disadvantages and its applications. Author also compared the performance of these algorithms using different metrics among which DBSCAN algorithm performed well in terms all the measure. Analysis show that author work can be used as a quick review to know about different clustering methods available in data mining. In future author will improve DBSCAN algorithm to produce improved results.

M. Z. Rodriguez et al implemented an organized evaluation of nine familiar clustering techniques accessible in R language using scattered data [12]. To tackle with possible data variation problems, Authors used datasets having numerous changeable characteristics. Moreover, authors also assessed the sensitivity of clustering techniques w.r.t their constraints conformation. From results authors concluded that while using the methods with default setting, particularly the spectral approach performed better. Also, the proposed study provides guidance to choose a better clustering method for analysis. In future extension of this work other algorithms could be compared and a comparable methodology could be used to solve semi-supervised classification problems.

Evaluation Parameters for Validation of Results

In the literature, different evaluation parameters have been utilized to validate the clustering results. These parameters include, accuracy, running time, precision, recall, F-measure, cluster shape, no. of clusters, complexity and handling outliers. Table 1 presents the analysis of literature with respect to the given parameters.

By concluding, it can be said that in literature many existing clustering techniques have been discussed, compared, analyzed and new clustering techniques have

been proposed. Many results discussed in the literature are good, but problem is that they have not been discussed in detail and in-depth. Also results of literature are in conflict to each other as according to some authors one algorithm is best but according to other authors some other algorithm is best e.g. according to literature [1], hierarchical clustering is best and according to literature [2] k-Mean clustering is best. So, it is difficult for readers to choose a single best clustering algorithm amongst others. Our study provides a solution to this problem by providing an in-depth analysis of most common and recent clustering techniques.

Proposed study produced detailed results based on different evaluation parameters using six different datasets. By producing comprehensive and accurate results we have to remove results conflicts and to suggest reader one best clustering algorithm amongst other clustering algorithms. To choose a best clustering technique amongst other clustering techniques will be a great advantage for data mining research community.

Table 1: Comparative Analysis Of Previous Work Based On Different Evaluation Parameters

Sr. No	Authors	Accuracy	Running Time	Precision	Recall	F- Measure	Cluster Shape	No. of Clusters	Complexity	Outliers Handling
1	S. Popat et al.[1]	No	No	No	No	No	Yes	No	Yes	Yes
2	Chaudhari et al.[2]	Yes	Yes	No	No	No	No	Yes	No	No
3	A.Kumar et al.[3]	No	No	No	No	No	Yes	Yes	Yes	Yes
4	Chaudhary et al.[4]	No	Yes	No	No	No	Yes	No	Yes	Yes
5	V. Leela et al.[5]	Yes	Yes	No	No	No	No	Yes	No	No
6	P. Baser et al.[6]	No	No	No	No	No	Yes	No	Yes	Yes
7	H. Mishra et al.[7]	Yes	Yes	No	No	No	No	No	No	No
8	S. Prabha et al.[8]	Yes	No	Yes	Yes	Yes	No	No	No	No
9	Sheshasay eet al.[9]	No	Yes	Yes	Yes	Yes	No	No	No	No
10	P. Singh et al.[10]	Yes	Yes	No	No	No	No	Yes	No	No

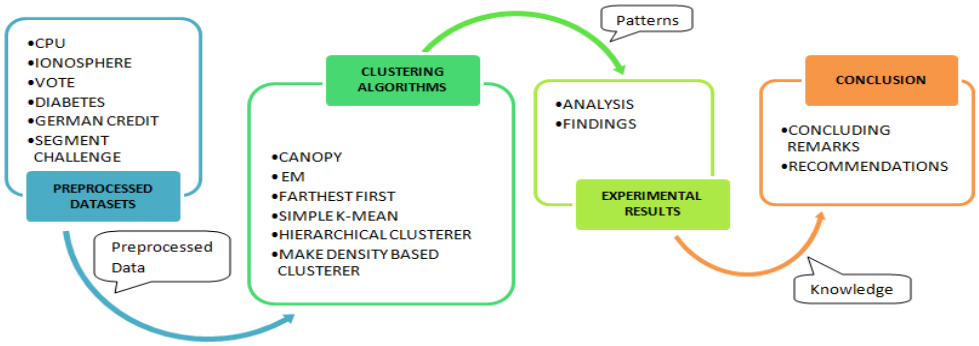

Figure 1: Flow Diagram of Proposed Methodology

3 Methodology

To carry out this research work, the following methodology has been adopted. This methodology comprises of various phases. In each phase, some specific tasks are performed. Figure 1 shows proposed methodology diagram of the system.

3.1 Proposed Datasets

Analysis was performed using Weka on six datasets form named as: "CPU, Ionosphere, Vote, Diabetes, German Credit, segment challenge" for clustering analysis. All datasets were preprocessed so we don't have to preprocess datasets again. All Datasets were of different sizes, areas and characteristics for better analysis and evaluation. The detail in term of total number of instances and total number of attributes of datasets is given below in Table 2.

3.2 Clustering Algorithms

Although there are numerous clustering algorithms but for analysis these clustering algorithms are used ('Canopy', 'EM', 'Farthest First', 'Hierarchical Clustered', 'Make density based Clustered', 'simple k-Means') to produce competitive results. These algorithms are used because they are main clustering algorithm and each one has its own pros and cons as discussed before. Clustering analysis has been applied through Weka tool, so all algorithms are available in Weka.

3.3 Evaluation Parameters

To evaluate our results following evaluation parameters has been used: 'Time taken to build model', 'No of Clusters', 'Cluster distribution', 'No of Iterations'. First parameter represents total time taken by algorithm to build model; Second parameter shows number of clusters while third parameter shows the percentage distribution of clusters built by the algorithm. Fourth parameter shows the number of iterations taken by algorithm to produce results. These evaluation parameters were chosen as they are most commonly used in literature and these evaluation parameters provide better understanding of results.

3.4 Environmental Setup

All experiments have been performed on Intel Core 2 Duo CPU with 2GB of RAM. Weka tool has been used for the analysis and for comparison of different clustering algorithms. Since Weka is a primary data mining application that contains different types of data mining algorithms. For study six datasets and six classification algorithms were chosen and results were compared on the basis of time taken to build model, number of clusters produced, cluster distribution and number of iteration values". For evaluation purpose, a test mode percentage split (20 %) has been used.

4 Results and Discussion

Table 3 shows different values of 'Time taken to build model', 'correctly classified instances', 'Incorrectly classified instances', 'Root mean squared error' for clustering algorithms against each dataset. Simple k-Mean algorithm produced better results in less time as compared to other algorithms. "Canopy", "Farthest First" and "Make density based clusterer" algorithms are taking almost same time but producing different results. Canopy algorithm makes more clusters as compare to other algorithms. For Small datasets Hierarchical clusterer algorithm produced better results in short time but for large datasets its performance is poor as its time starts increasing. EM Algorithm is worst of them all as it takes more time and not produced good results as compare to other algorithms.

Table 2: Detailed Description of Experimental Datasets

Sr. No	Dataset	Total No of Instances	Total No of Attributes
1	CPU	209	7
2	Ionosphere	351	35
3	Vote	435	17
4	Diabetes	768	9
5	German Credit	1000	21
6	Segment Challenge	1500	20

Table 3. Experimental Results Using Different Clustering Algorithms on Mentioned Datasets

DataSet	Clustering Algorithms	Time Taken to Build Model	No of Clusters	Clusters Distribution	No of Iterations
CPU	Canopy	0.01 Sec	3	26 % 73 % 1 %	--
	EM	0.37 Sec	3	26 % 35 % 39 %	6
	Farthest First	0.02 Sec	2	94 % 6 %	--
	Hierarchical Clusterer	0.11 Sec	2	99 % 1 %	--
	Make Density Based Clusterer	0 .02 Sec	2	73 % 27 %	7
	k-Means	0.01 Sec	2	83 % 17 %	7
Ionosphere	Canopy	0.03 Sec	3	57 % 25 % 19%	--
	EM	1.04 Sec	3	51 % 7 % 42%	1
	Farthest First	0.02 Sec	2	98 % 2 %	--
	Hierarchical Clusterer	0.87 Sec	1	100%	--
	Make Density Based Clusterer	0.02 Sec	2	49 % 51 %	6
	k-Means	0.01 Sec	2	36 % 64 %	6
Vote	Canopy	0.01 Sec	4	36 % 34 % 21 % 9 %	--
	EM	1.73 Sec	4	38 % 31 % 17 % 14 %	35
	Farthest First	0.02 Sec	2	57 % 43 %	--
	Hierarchical Clusterer	1.33 sec	2	100 % 0 %	--
	Make Density Based Clusterer	0.02 Sec	2	55 % 45 %	3
	k-Means	0.01 Sec	2	55 % 45 %	3

Diabetes	Canopy	0.03 Sec	5	60 % 22 % 5 % 6 % 7 %	--
	EM	3.86 Sec	5	29 % 29 % 20 % 18 % 4 %	2
	Farthest First	0 Sec	2	75 % 25 %	--
	Hierarchical Clusterer	2.56	2	66 % 34 %	--
	Make Density Based Clusterer	0.03 Sec	2	66 % 34 %	3
	k-Means	0.02 Sec	2	66 % 34 %	3
German Credit	Canopy	0.13 Sec	6	24% 22 % 18 % 16 % 12 % 8 %	--
	EM	6.58 Sec	2	73 % 27 %	3
	Farthest First	0.01 Sec	2	84 % 16 %	--
	Hierarchical Clustered	3.03 Sec	2	58 % 42 %	--
	Make Density Based Clusterer	0.03 Sec	2	43 % 57 %	4
	k-Means	0.02 Sec	2	67 % 33 %	4
Segment Challenge	Canopy	0.03 Sec	4	48 % 14 % 23 % 15 %	--
	EM	8.62 Sec	3	22 % 35 % 42 %	2
	Farthest First	0.03 Sec	2	83 % 17 %	--
	Hierarchical Clusterer	4.44 Sec	2	53 % 47 %	--
	Make Density Based Clusterer	0.03 Sec	2	66 % 34 %	6
	k-Means	0.02 Sec	2	72 % 28 %	6

Figure 2 shows time taken to build model comparison for all algorithms for each dataset. It can be seen clearly that simple k-mean, Canopy, Farthest First and Make density based Clusterer algorithms took almost same time for each dataset. Hierarchical clusterer took more time as compare to other four algorithms. In general, Hierarchical algorithm produced good results for small datasets in less time but for large datasets it took more time. EM algorithm took more time for each dataset as compare to all other algorithms. By talking about performance of each algorithm then performance of EM algorithm is worst of them all. Performance of simple k-Mean algorithm is best amongst all of them. Then performance decreases from Make density-based Algorithm, Farthest First algorithm, Canopy

algorithm to Hierarchical clusterer algorithm.

Figure 3 shows accuracy of each algorithm with respect to each dataset. It can be seen clearly that accuracy of k-Mean Algorithm is high amongst all other algorithms. Then accuracy decreases in a specific pattern from Make density-based algorithm to EM algorithm. The accuracy of EM algorithm is lowest from all other algorithms.

hierarchical algorithm took more average time to build model as compared to literature results. But as current and literature results of two algorithms are almost same so it can be said that our results are reasonable.

Figure 5 shows comparison between current and literature results of average accuracy in percentage (%) of k-Mean and Make density-based algorithm. There is slight

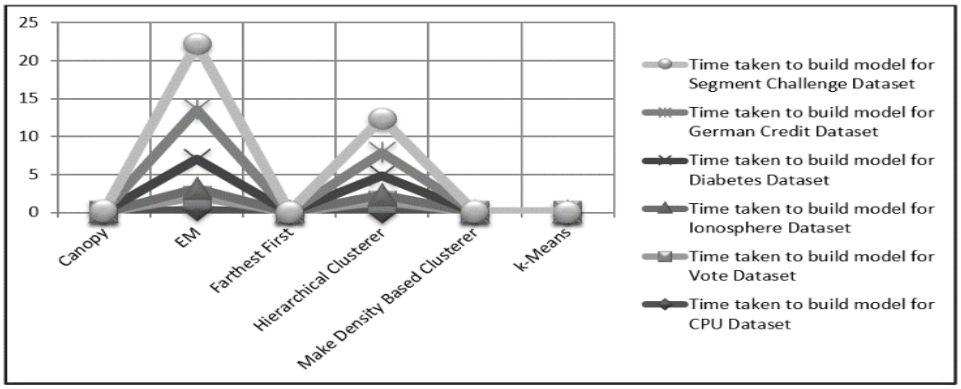

Figure 2: Running Time Comparison of Each Algorithm w.r.t Each Dataset

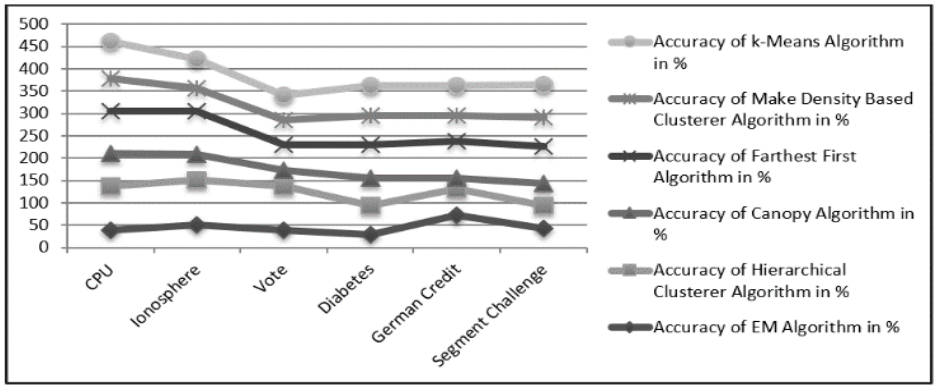

Figure 3: Accuracy Comparison of Each Algorithm w.r.t Each Dataset

Table 4 shows comparison between proposed results and amongst results given in literature. Comparison of results has been made in term of time taken to build model and in term of average accuracy of three algorithms: k-Mean, Hierarchical Clusterer and Make density-based Algorithm. These algorithms were compared because they are most important and primary clustering algorithms used for clustering. These algorithms and many other types of clustering algorithms were discussed in literature. Most studies used different datasets for analysis, but some datasets were common in our proposed study and in literature study e.g. CPU, ionosphere and Diabetes. So, in proposed study comparison of only common algorithms based on common datasets has been performed. Graphical representation of results has been shown below.

Figure 4 shows that current and literature results of average time taken to build model by k-Mean and Make density-based algorithm is almost same. But there is a difference between current and Literature results when it comes to Hierarchical cluster algorithm. In current results

difference between current and Literature results of all three algorithms. It could be because of different size or type of dataset or there could be any other reason. But as current and literature results of two algorithms are almost same expect minor differences in accuracy so it can be said that our results are reasonable and accurate.

4.1Findings from the Experiment and Results

From experiment and results it is concluded that k-Mean algorithm performance is best, and performance of EM algorithm is worst as compare to other algorithms. Hierarchical algorithm is best for small datasets. Performance of three algorithms "Farthest First, Make density based Clusterer and Canopy algorithm" is almost same with slight difference in results. For data with varying density, Make density-based cluster is not suitable. Overall conclusion can be made that k-Mean algorithm is simplest, produced quality clusters and has high performance amongst all other five algorithms.

Table 4. Comparative Analysis of Best Results

Algorithms	Average time taken to build model in Literature	Current Average time taken to build model	Average accuracy in Literature in %	Current Average accuracy in %
k-Means	0.3 sec	0.015 Sec	76	67.8
Hierarchical Clusterer	2.8 Sec	2.06 Sec	88	79
Make Density Based Clusterer	0.106	0.025 Sec	71	66

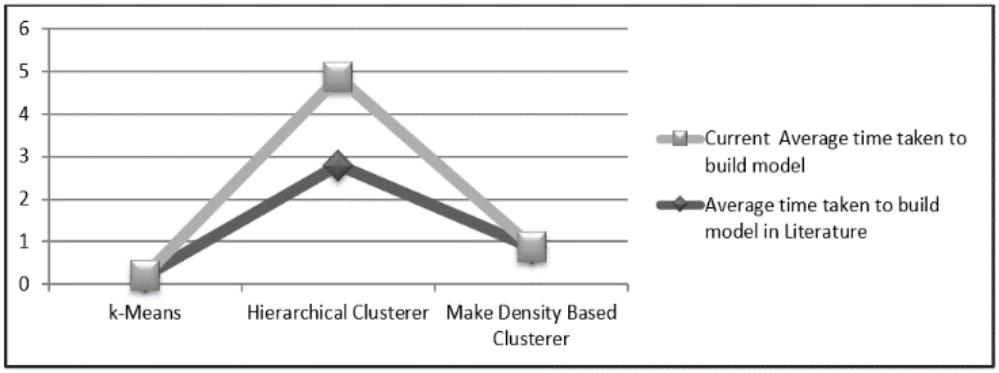

Figure 4: Average Running Time Comparison of Our and Existing Results

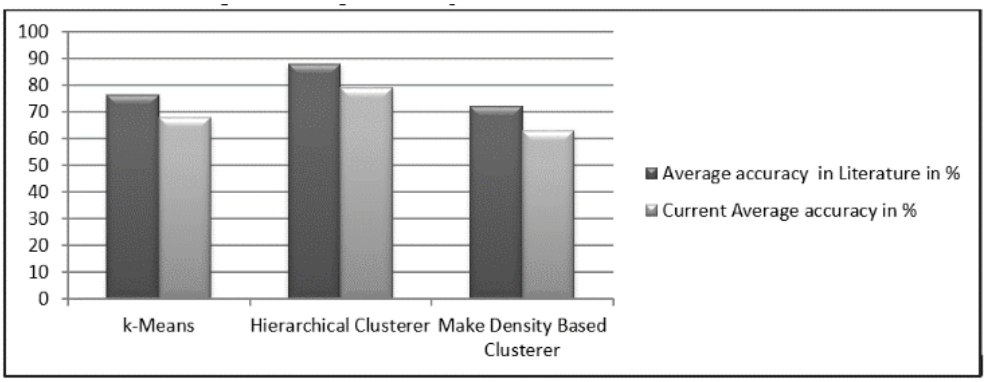

Figure 5: Average Accuracy Comparison of Our and Existing Results

5 Conclusion

In this study, comparison of six different clustering algorithms which includes ('Canopy', 'EM', 'Farthest First', 'Hierarchical Clusterer', 'Make density based Clusterer', 'simple k-Means') has been performed using six different datasets. Algorithms are compared on the basis of time taken to build model, number of clusters produced, cluster distribution and number of iteration values. From results it is concluded that performance of k-Mean algorithm is best amongst all other five algorithms as it produced accurate results in a short time. Performance of EM algorithm is worst amongst all other five algorithms as it took more time to produce inaccurate results. Hierarchical algorithm is sensitive to size of data. It is best for small datasets but on huge datasets it takes more time as compare to other algorithms.

Performance of three algorithms "Farthest Fist, Make density based Clusterer and Canopy algorithm" is almost same with slight difference in results. For data with varying density, Make density-based cluster is not suitable. Overall conclusion can be made that k-Mean algorithm is simplest, produced quality clusters and has high performance amongst all other five algorithms. We also compared our study results with existing results and proved that our results are quite reasonable and accurate as there was slight difference between them. Our proposed analysis and results make better understanding for cluster researcher to improve existing techniques and also to analyze more techniques and to propose a new clustering technique.

In Future, comparison and analysis of other clustering techniques will be performed and results will be compared with current results for better understanding and analysis.

References

[1] K. Popat, Review and Comparative Study of Clustering Techniques, *International Journal of Computer Science and Information Technologies (IJCSIT)*., **5(1),** 805-812 (2014).
[2] B. Chaudhari, A Comparative Study of clustering algorithms Using weka tools, *International Journal of Application or Innovation in Engineering & Management (IJAIEM)*., **1(2)**, 154-158 (2012).
[3] K. A. Kumar, *A Comparative Study & Performance Evaluation of Different Clustering Techniques in Data Mining,* in Proc. ACEIT., 139-142, (2016).
[4] S. chaudhary, Comparative Study of Various Clustering Techniques in Data Mining, *Jaipur International Journal of Converging Technologies and Management (IJCTM)*., **1(1)**, 1-4 (2015).
[5] V. Leela, Comparative Study of Clustering Techniques in Iris DataSets, *World Applied Sciences Journal (Data Mining and Soft Computing Techniques)*., **29**, 24-29 (2014).
[6] P. Baser, A Comparative Analysis of Various Clustering Techniques used for Very Large Datasets, *International Journal of Computer Science & Communication Networks*., **3(4)**, 271-275 (2017).
[7] H. Mishra, A Comparative Study of Data Clustering Techniques, *International Research Journal of Engineering and Technology (IRJET)*., **4(5)**, 1392-1398 (2017).
[8] M. S. Prabha, Analysis of Different Clustering Techniques in Data and Text Mining, *International Journal of Computer Science Engineering (IJCSE)*., **3(2)**, 107-116 (2014).
[9] A. Sheshasayee, Comparative Analysis of Clustering Techniques for Requirements Clustering, *Middle-East Journal of Scientific Research*., **21(7)**, 1097-1102 (2014).
[10] P. Singh, Performance Analysis Of Clustering Algorithms In Data Mining In Weka, *International Journal of Advances in Engineering & Technology*., **7(6)**, 1866-1873 (2015).
[11] N. Valarmathy, Performance Evaluation and Comparison of Clustering Algorithms used in Educational Data Mining, *International Journal of Recent Technology and Engineering (IJRTE)*., **7(6S5)**, 103-113 (2019).
[12] Rodriguez, Clustering algorithms: A comparative approach, *PLOS ONE*., **14(1)**, e0210236 (2019).
[13] K. DeFreitas, Comparative Performance Analysis Of Clustering Techniques In Educational Data Mining, *International Journal on Computer Science and Information Systems*.,**10(2)**, 65-78 (2015).
[14] S. R. Pande, Data Clustering Using Data Mining Techniques, *International Journal of Advanced Research in Computer and Communication Engineering*., **1(8)**, 494-499 (2012).
[15] A. Joshi, A Review: Comparative Study of Various Clustering Techniques in Data Mining, *International Journal of Advanced Research in Computer Science and Software Engineering*., **3(2)**, 55-57 (2013).
[16] A. J. Patil, Comparative Study of Different Clustering Algorithms, *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*., **3(7)**, 10490-10497 (2014).
[17] D. Sharma, A Study of Data Mining Clustering Techniques, *International Journal of Advanced Research in Computer Science and Software Engineering*., **4(3)**, 490-494 (2014).
[18] D. Patel, A Comparative Study of Clustering Data Mining: Techniques and Research Challenges, *IJLTEMAS*., **3(9)**, 67-70 (2014).
[19] A.Bharathi, A Survey on Crime Data Analysis of Data Mining Using Clustering Techniques, *Internatioal Journal of Advance Research in Computer Science and Management Studies*., **2(8)**, 9-13 (2014).
[20] G. Singhal, A Comparative Study of Data Clustering Algorithms, *International Journal of Computer Applications*, **83(15)**, 41-46 (2013).
[21] Patil, Comparative Study of Different Clustering Algorithms, *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*., **3(7)**, 10490-10497 (2014).
[22] Sharma, Comparative Analysis of Various Clustering Algorithms Using WEKA, *International Research Journal of Engineering and Technology (IRJET)*, **2(4)**, 107-112, (2015).
[23] DoĞRu, Comparison of clustering techniques for traffic accident detection, *Turkish Journal of Electrical Engineering & Computer Sciences*., **23**, 2124-2137 (2015).
[24] H. Rana, A Study of Web Log Analysis Using Clustering Techniques, *International Journal of Innovative Research in Computer and Communication Engineering*., **1(4)**, 925-929 (2013).
[25] R. bala, A Comparative Analysis of Clustering Algorithms, *International Journal of Computer Applications*., **100(15)**, 35-39 (2014).
[26] Ali, Implementation and Analysis of Clustering Techniques Applied on Pocket Switched Network, *International Journal of Distributed Sensor Networks*., **11(9)**, 239591 (2015).
[27] R. Pallavi, Analysis of Clustering Technique in Marketing Sector, *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*., **5(2)**, 209-211 (2017).
[28] I. Daniel, Vector versus Tree Model Representation in Document Clustering, *Romanian Journal Of Information Science And Technology*., **16(1)**, 22 (2013).
[29] P. H. Ahmad, Performance Evaluation of Clustering Algorithm Using Different datasets, *International Journal of Advance Research in Computer Science and Management Studies*., **3**, 167-173 (2015).
[30] K. Sivaraman, Clustering analysis in data mining, *International Journal of Pure and Applied Mathematics*., **119**, 9639-9649 (2018).

Shamveel Hussain Shah did his MS Computer Science Degree from University of Engineering and Technology Taxila, Pakistan. He did his BS (Hons) Computer Science from University of Wah, Wah Cantt Taxila, Pakistan. His research interests include Machine Learning, Dep Learning, Pattern Recognition, Big Data Mining and Data Science.

Muhammad Javed Iqbal received his PhD Computer Science/Information Technology degree from Universiti Teknologi PETRONAS, Malaysia in February 2015. He did his M.Sc. Computer Science degree in 2001 from University of Agriculture, Faisalabad, Pakistan and MS/M.Phil Computer Science in 2008 from International Islamic University Islamabad, Pakistan. Presently. He is HEC approved PhD Supervisor and working as an Assistant Professor Computer Science Department, University of Engineering and Technology Taxila, Pakistan. After completion of his doctoral studies, he has been actively involved in research. He has more than twenty-five international journals and conferences. His research interests include Machine Learning, Data Science, Pattern Recognition and Big Data Mining and Analytics.

Muhammad Bakhsh received his M.Sc. & MS in computer science from International Islamic University, Islamabad in 2002 and 2006, and a Ph.D., also in computer science, from the Allama Iqbal Open University, Islamabad in 2017. He is working as Senior Research Associate at Pakistan Academy for Rural Development, Peshawar Pakistan. He has 13 years of experience in research and training. He has published over 16 research papers in peer reviewed international journals. He also served as a member of technical committees for international and national conferences and journals. His research interests include web accessibility, m-learning and data science.

Amjad Iqbal received his M.I.T (Master of Information Technology) degree from Virtual University of Pakistan in April 2014. He did his bachelor's degree from the Bahauddin Zikariya University, Multan, Pakistan in 2010. He is currently doing MS Computer Science (MSCS) at the department of computer science HITEC University Taxila, Pakistan. His research interests include Machine Learning, Data Science, Statistical Analyst, Programming, Image Processing, and Pattern Recognition etc.