ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 7, July 2018

Comparison of K-Means Algorithm and Hierarchical Algorithm using Weka Tool

Saleena T.S [1], Dr. S.J.Sathish Aaron Joseph[2]

Asst Professor, Dept of Computer Science, SS College, Areekode, Malappuram Dt, Kerala, India[1]

Asst Professor and HOD of Computer Application, J J College of Arts and Science, Pudukkottai, Tamilnadu[2]

Abstract: Data mining is a process of Collecting useful information and patterns from huge data. Clustering is a process of partitioning a set of data or objects into a set of meaningful sub-classes, called clusters. In clustering, objects of the data set are grouped into clusters, such a way that each group are very different from each other and the objects in the same group are very similar to each other. In this paper analyses two major clustering algorithms: K-Means and Hierarchical. The performance of these two clustering algorithms is compared using the clustering toolkit Weka, which is a platform-independent open source toolkit.

Keywords: Data mining, Clustering, Clustering algorithms, K-means algorithms, Hierarchical clustering and Weka toolkit

I. INTRODUCTION

Data mining is a process of extraction of useful information and patterns from huge data. Data mining is a logical process that is used to search through large amount of data in order to find useful data. Clustering is a process of partitioning a set of data or objects into a set of meaningful sub-classes, called clusters. In clustering, objects of the data set are grouped into clusters, such a way that each group are very different from each other and the objects in the same group are very similar to each other. Unlike Classification, in which predefined set of classes are presented, but in Clustering there are no predefined set of classes which means that resulting clusters are not known before the execution of clustering algorithm. In this, these clusters are extracted from the dataset by grouping the objects in it.

II. K-MEANS CLUSTERING

K-means is a widely used partitioned clustering method in the industries. The K-means algorithm is the most commonly used partitioned clustering algorithm because it can be easily implemented and is the most efficient one in terms of the execution time. K-Means was first introduced by James Mac Queen in 1967.

K-means technique is used to classify the given data objects into k different 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. The algorithm consists of two separate phases. In the first phase user selects k centres randomly, where the value k is fixed in advance. To take each data object to the nearest centre. Several distance functions are considered to determine the distance between each data object and the cluster centre. When all the data objects are included in some clusters, the first step is completed and an early grouping is done. Then the second phase is to recalculate the average of the early formed clusters. This iterative process continues repeatedly until the criterion function becomes the minimum.

K-Means Algorithm Properties

- There are always K clusters.
- There is always at least one item in each cluster.
- The clusters are non-hierarchical and they do not overlap.
- Every member of a cluster is closer to its cluster than any other cluster because closeness does not always involve the 'center' of clusters.

Strengths of K-Mean

- Simple: - Easy to understand and to implement.
- Efficient: Time complexity: O(tkn), where n is the number of data points, k is the number of clusters, and t is the number of iterations.
- Since both k and t are small. k-Means is considered a linear algorithm.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 7, July 2018

Weaknesses of k-means

- The algorithm is only applicable if the mean is defined.
- The user needs to specify k.
- The algorithm is sensitive to outliers. Outliers are data points that are very far away from other data points.
- The k-means algorithm is not suitable for discovering clusters that are not hyperellipsoids (or hyper-spheres).

In Short, K Means is:-

- Despite weaknesses, k-means is still the most popular algorithm due to its simplicity, efficiency.
- No clear evidence that any other clustering algorithm performs better in general.
- Comparing different clustering algorithms is a difficult task. No one knows the correct clusters.

III. HIERARCHICAL CLUSTERING

Hierarchical methods are well known clustering technique that can be potentially very useful for various data mining tasks. A hierarchical clustering scheme produces a sequence of clustering in which each clustering is nestled into the next clustering in the sequence. Since hierarchical clustering is a greedy search algorithm based on a local search, the merging decision made early in the agglomerative process are not necessarily the right ones. One possible solution to this problem is to refine a clustering produced by the agglomerative hierarchical algorithm to potentially correct the mistakes made early in the agglomerative process. Hierarchical methods are commonly used for clustering in Data Mining. A hierarchical method creates a hierarchical decomposition of the given set of data objects. Here tree of clusters called as dendrogram is built. Every cluster node contains child clusters, sibling clusters partition the points covered by their common parent. In hierarchical clustering we allocate each item to a cluster such that if we have N items then we have N clusters. Find closest pair of clusters and merge them into single cluster. Calculate distance between new cluster and each of old clusters. We have to repeat these steps until all items are clustered into K no. of clusters. It is of two types, Agglomerative clustering and Divisive clustering

Agglomerative (bottom up): Agglomerative hierarchical clustering is a bottom-up clustering method where clusters have sub-clusters, which in turn have sub-clusters, etc. It starts by letting each object form its own cluster and iteratively merges cluster into larger and larger clusters, until all the objects are in a single cluster or certain termination condition is satisfied.

This algorithm produces sequence of clustering of decreasing number of clusters at each step. The clusters produced at each step results from the previous step, by merging two clusters into one. The clusters are merged by computing the distance between each pair of clusters. For n samples, agglomerative algorithms begin with n clusters and each cluster contains a single sample or a point. Then two clusters will merge so that the similarity between them is the closest until the number of clusters becomes 1 or as specified by the user.

Algorithm:

1. Start with n clusters, and a single sample indicates one cluster.
2. Find the most similar clusters Ci and Cj then merge them into one cluster.
3. Repeat step 2 until the number of cluster becomes one or as specified by the user. The distances between each pair of clusters are computed to choose two clusters that have more opportunity to merge. There are several ways to calculate the distances between the clusters Ci and Cj.

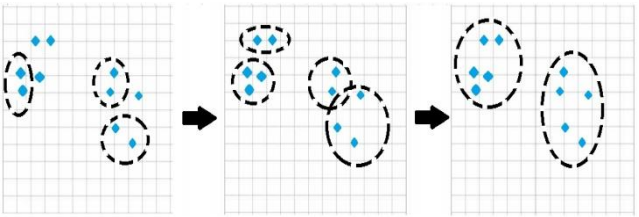

Figure1: Agglomerative Hierarchical Clustering

Divisive (top down):A top-down clustering method and is less commonly used. It works in a similar way to agglomerative clustering but in the opposite direction. This method starts with a single cluster containing all objects, and then successively splits resulting clusters until only clusters of individual objects remain. Divisive clustering: It is a top-down clustering method which works in a similar way to agglomerative clustering but in the opposite direction. This method starts with a single cluster containing all objects and then successively splits resulting clusters until only clusters of individual objects remain.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 7, July 2018

Algorithm:

1. Start with one cluster that contains all samples.
2. Calculate diameter of each cluster. Diameter is the maximal distance between samples in the cluster. Choose one cluster C having maximal diameter of all clusters to split.
3. Find the most dissimilar sample x from cluster C. Let x depart from the original cluster C to form a new independent cluster N (now cluster C does not include sample x). Assign all members of cluster C to Mc.
4. Repeat step 6 until members of cluster C and N do not change.
5. Calculate similarities from each member of Mc to cluster C and N, and let the member owning the highest similarities in Mc move to its similar cluster C or N. Update members of C and N.
6. Repeat the step 2, 3, 4, 5 until the number of clusters becomes the number of samples or as specified by the user.

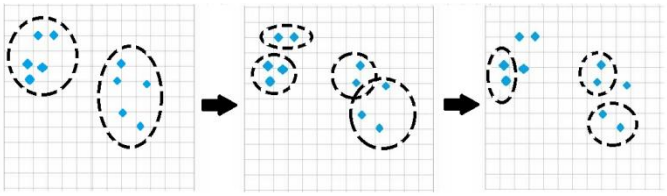

Figure2: Divisive Hierarchical Clustering

Advantages of Hierarchical Clustering

The advantages of the hierarchical clustering algorithms are the reason this algorithm was chosen for discussion. These advantages include,

- Easy of handling of any forms of similarity or distance.
- Consequently applicability to any attributes types.
- Small clusters are obtained which is easier to analyze and understand.
- Number of clusters is not fixed at the beginning. Hence, user has the flexibility of choosing the clusters dynamically.
- Conceptually Simple.
- Theoretical properties are well understood.
- When Clusters are merged /split, the decision is permanent => the number of different alternatives that need to be examined is reduced.

Weakness of Hierarchical Clustering

- Merging /splitting of clusters is permanent => Erroneous decisions are impossible to correct later.
- If objects are grouped incorrectly at the initial stages, they cannot be relocated at later stages.
- The results vary based on the distance metrics used.
- Divisive methods can be computational hard.
- Methods are not (necessarily) scalable for large datasets.
- Needs a termination/readout condition.
- The final mode in both Agglomerative and Divisive is of no use.

IV. WEKA TOOLKIT

Weka is considered as a landmark system in the history of the data mining among machine learning research communities[9]. The toolkit has gained widespread adoption and survived for an extended period of time. The toolkit is developed at the University of Waikato, New Zealand. The acronym stands for Waikato Environment for Knowledge Analysis. Weka is platform-independent open source toolkit.

Weka is freely available on the World-Wide Web and accompanies a new text on data mining which documents and fully explains all the algorithms it contains. Applications written using the Weka class libraries can be run on any computer with a Web browsing capability; this allows users to apply machine learning techniques to their own data regardless of computer platform.

The primary learning methods in Weka are “classifiers”, and they induce a rule set or decision tree that models the data. Weka also includes algorithms for learning association rules and clustering data. All implementations have a uniform command-line interface. A common evaluation module measures the relative performance of several learning algorithms over a given data set.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 7, July 2018

V. COMPARISON

Above section involves the study of each of the two techniques introduced previously using Weka Clustering Tool on a set of data consists of 10 attributes and 50 entries. Clustering of the data set is done with each of the clustering algorithm using Weka tool.

Now we are going to compare K-Means and Hierarchical Algorithm in case of small data sets such as 10 attributes and only 50 entries.

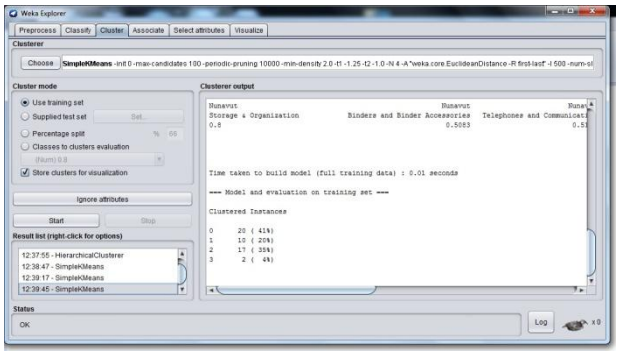

Figure 3 : Result of K-Means clustering with small dataset.

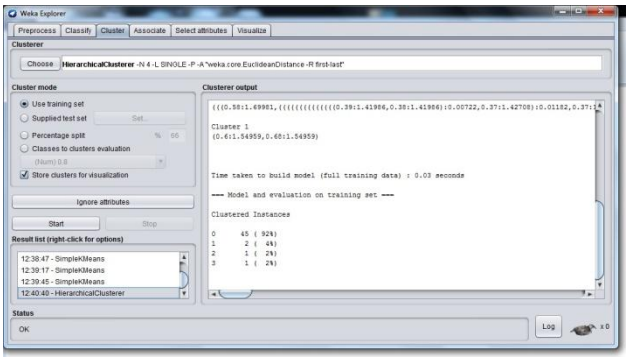

Figure 4 : Result of hierarchical clustering with small dataset.

In the above two diagram, it represents the comparisons of algorithms with small datasets. That is 10 attributes with 50 entries. Here Number of Clusters are only 4.

Table 1 : Comparison result of algorithms with small dataset using weka tool.

Hierarchical Algorithm	K – Means Algorithm	Name
4	4	No.of clusters
0:45 (92%) 1:2 (4%) 2:1 (2%) 3:1 (2%)	0:20 (41%) 1:10 (20%) 2:17 (35%) 3:2 (4%)	Cluster Instances
	6	No. of Iterations
	123.37875	Within clusters sum of squared errors
0.03 seconds	0.01 seconds	Time taken to build model
0	0	Unclustered Instances

Now we are going to compare K-Means and Hierarchical Algorithm in case of large data sets such as 10 attributes and 1000 entries.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 7, July 2018

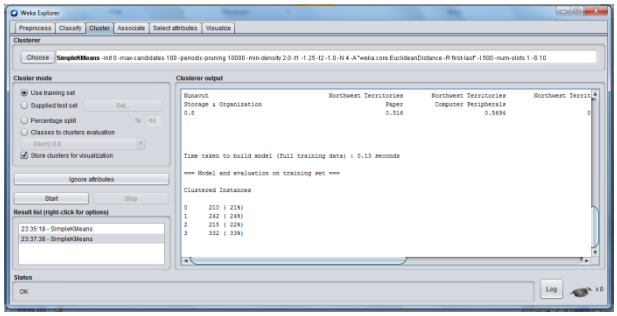

Figure 5 : Result of K-Means clustering with large dataset.

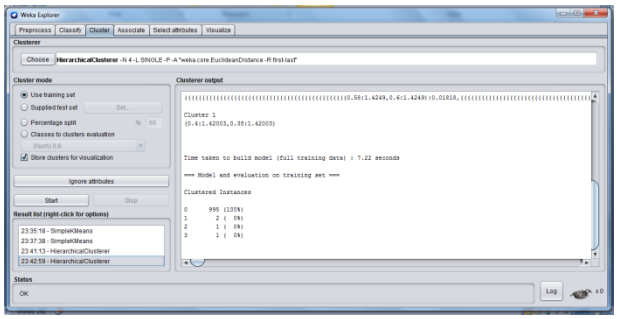

Figure 6 : Result of hierarchical clustering with large dataset.

In the above two diagram, it represents the comparisons of algorithms with large datasets. That is 10 attributes with 1000 entries. Here Number of Clusters are only 4.

Table 2 : Comparison result of algorithms with large dataset using weka tool.

Hierarchical Algorithm	K – Means Algorithm	Name
4	4	No.of clusters
0:995 (100%) 1:2 (0%) 2:1 (0%) 3:1 (0%)	0:210 (21%) 1:242 (24%) 2:215 (22%) 3:332 (33%)	Cluster Instances
	21	No.of Iterations
	3143.72007981549	Within clusters sum of squared errors
7.22 seconds	0.13 seconds	Time taken to build model
0	0	Unclustered Instances

VI. CONCLUSION AND FUTURE SCOPE

CONCLUSION

The K - mean algorithm has the advantage of clustering large data sets and its performance increases as the number of clusters increases. The performance of K- mean algorithm is better than Hierarchical Clustering Algorithm. Performance of K-Means algorithm increases as the RMSE decreases and the RMSE decreases as the number of cluster increases. All the algorithms have some noise or ambiguity in some data when clustered. The quality of all algorithms becomes very good when using huge dataset. K-Means is very sensitive to noise in the dataset. This noise makes it difficult for the algorithm to cluster data into suitable clusters, while affecting the result of the algorithm. K-Means algorithm is faster than other clustering algorithm and also produces quality clusters when using huge dataset. Hierarchical clustering algorithm is more sensitive for noisy data.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 7, July 2018

FUTURE SCOPE

As a future work, comparison between these algorithms (or may other algorithms) may be done using different parameters other than considered in this paper. One important factor is normalization. Comparing between the results of algorithms using normalized data and non-normalized data will give different results. Of course normalization will affect the performance of the algorithm and the quality of the results.

REFERENCES

[1] Abbas ,O.A., Jordan, "Comparisons Between Data Clustering Algorithms, "The International Arab Journal of Information Technology, vol. 5, no. 3, pp. 320-326, Jul. 2008.
[2] Aastha Joshi and Rajneet Kaur "A Review : Comparitive Study of Various Clustering Techniques in Data Mining." ISSN: 2277 – 128X International Journal of Advanced Research in Computer Engineering & Technology, Volume 3, Issue 3, March 2013.
[3] Verma ,M., Srivastava., Chack,N., Diswar, A .K ., Gupta,N.," A Comparative Study of Various Clustering Algorithms in Data Mining," International Journal of Engineering Research and Applications (IJERA), Vol. 2, Issue 3, pp. 1379-1384, 2012.
[4] Tajunisha , Saravanan, "Performance analysis of k-means with different initialization methods for high dimensional datasets", International Journal of Artificial Intelligence & Applications (IJAIA), vol. 1, no.4, pp. 44-52, Oct. 2010.
[5] Chandelier, N. S., Nandavadekar, V. D.,"Comparative Analysis of Two Algorithms for Intrusion Attack Classification Using KDD CUP Dataset**,** "International Journal of Computer Science and Engineering (IJCSE), Vol. 1, pp. 81-88, Aug 2012.
[6] Koyukuk, M., Grama,A., Krishnan, N. R., "Compression, Clustering, and Pattern Discovery in Very High-Dimensional Discrete-Attribute Data Sets", IEEE Transactions On Knowledge And Data Engineering, Volume 45 Issue3, page No -377-401, July 2006.
[7] Bishop C. M. , Michael, E. Tipping, "Hierarchical Latent Variable Model for Data Visualization", IEEE Trans. Pattern Anal. Mach Intell., 20 (3), 281-293,1998.
[8] "Tools Pros and Cons of Clustering Algorithm using Weka Tools" by Ayyoob M.P, Assistant Professor, Dept. of Computer Science, Sullamussalam Science College, Areacode- Kerala in International Journal of Computer Application (ISSN: 0975-8887)
[9] Sharma,N., Bajpai,A., Litoriya,R., "Comparing the various clustering algorithms of Weka tool",International Journal of Emerging Technology and Advanced Engineering ,2(5),2012.