Journal of Engineering and Applied Sciences 13 (Special Issue 5): 4592-4596, 2018
ISSN: 1816-949X

Collation Between Hierarchical and K-Means Clustering Algorithm

K.T. Athira, P.G. Gopika and G. Deepa
Department of Computer Science and IT, School of Arts and Sciences,
Amrita University, Kochi, India

Abstract: Clustering is a technique of keeping the closely related or in other words similar data into groups. Clustering is mainly a process in which a given data set is partitioned into homogenous groups on the basis of certain features like the similar objects are clustered into one group and dissimilar objects are clustered into another group. Therefore, clustering can be termed as unsupervised way of learning in which the aim is to find unlabeled data structure. Enormous clustering techniques are available such as partitioning, density-based, grid-based, hierarchical, model-based and soft-computing methods. In this study, we propose a comparative study between k-means and hierarchical clustering methods, claiming that the quality of hierarchical clustering increases as compared to k-means clustering for the same number of iterations and splitting percentage as the clustered instances will be more for hierarchical clustering. When performance is considered, k-means stands over hierarchical clustering. Also, we propose that when data is transformed by normalization which is a data preprocessing task results in improved accuracy and quality of hierarchical clustering as of now.

Key words: Hierarchical, k-means, clustered instances, data preprocessing, normalization, preprocessing task

INTRODUCTION

Data mining is a process of knowledge extraction from a set of historical data and results in the prediction of outcomes. An important data mining task is clustering. Clustering is mainly a form of unsupervised learning in which similar form of data are grouped into a single cluster and this similar group of data into another. The main aim of clustering is to determine whether two objects are similar or dissimilar by providing various measures and criteria. Clustering can be easily done using WEKA tool. Data clustering can be done mainly using two methods:

- Hierarchical and
- Partitioning methods

In this study, we bring out a just a position between hierarchical clustering and k-means clustering. By taking normalized dataset and performing the clustering techniques and simulating which clustering mechanism is better than the other on what terms? Hierarchical clustering can it self be split into two:

- Agglomerative algorithms and
- Divisive algorithms

k-means clustering algorithm comes under partitioning method. It is one of the efficient and simplest unsupervised learning algorithm that can easily make clusters. It is a method for finding the position of clusters that will reduce the distance from data points to that of a cluster. Hierarchical Clustering (HCA) is an algorithm used for cluster analysis which aims at building hierarchy of clusters.

Literature review: The study of Kaushik and Mathur (2014) “Comparative study of k-means and hierarchical clustering techniques”, provides a basic comparison between k-means and hierarchical clustering. It brings out the strengths and weaknesses of both the clustering methods. Comparative study done in this study claims that the k-means algorithm performance is better that hierarchical clustering algorithm whereas the quality of hierarchical algorithm is higher than that of k-means clustering algorithm.

Another study titled “Comparison Between k-means and Hierarchical Algorithm Using Query Redirection” (Kaur and Kaur, 2013) analyses hierarchical clustering algorithm and k-means algorithm by applying validation measures such as f-measure, entropy, coefficient of variance and time. The result depicts that k-means clustering method is more efficient when compared to hierarchical as it takes less execution time.

Comparison between data clustering algorithms (Abbas, 2008), this study compare different data clustering algorithms such as hierarchical clustering algorithm, k-means algorithm, expectation maximization clustering algorithms and self-organizing map algorithm based on the factors such as number of clusters, software type used and the dataset size. The conclusion made from

Corresponding Author: K.T. Athira, Department of Computer Science and IT, School of Arts and Sciences, Amrita University, Kochi, India

the analysis is that the performance of k-means and expectation maximization clustering is more than that of hierarchical clustering whereas the accuracy of hierarchical is greater. Hierarchical cluster quality is more than other algorithms. A study of hierarchical clustering algorithm (Rani and Rohil, 2013), this study depicts various improved hierarchical clustering algorithms such as (CURE (Clustering Using Representatives), BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies), CHEMELEON algorithm and linkage algorithms) belongs to agglomerative hierarchical methods, leaders-sub leaders, Bisecting k-means which is a divisive hierarchical clustering algorithm. The study suggests that the pure hierarchical clustering technique quality suffers from the disability to perform adjustment.

Data mining using hierarchical agglomerative clustering algorithm in distributed cloud computing environment (Srivastava *et al*., 2013). This study shows a method to implement hierarchical agglomerative clustering algorithm in a way to cluster large dataset and increasing the efficiency by parallel task execution which results an increase in linear data set growth of execution time.

Hierarchical clustering algorithms for document datasets (Zhao *et al*., 2005) shows document clustering algorithms for building hierarchical solutions that will present a study based on partitional and agglomerative algorithms using different merging techniques and criterion functions. This study also depicts a new class of algorithms for clustering called constrained agglomerative algorithms that will be combining both the features of partitional and agglomerative approaches thus enhancing the clustering solutions quality.

Another study is “Comparing the various clustering algorithms of WEKA tool (Sharma *et al*., 2012). In this study, comparison is made between various clustering algorithms designed inside WEKA tool and to point out which algorithm is superior to than other algorithms as similar data objects are assembled together.

Comparing between k-means and hierarchical algorithm on the basis of normalization (Nair *et al*., 2016) represents dataset normalization to collate k-means and hierarchical clustering algorithm by implementing validation techniques such as covariance and entropy and comes to a conclusion that k-means algorithm shows improved results than after normalization.

MATERIALS AND METHODS

Clustering algorithm: Clustering is a way of assembling a group of abstract objects into cluster of similar objects. It is type of unsupervised learning in which there will be no predefined set of classes, so that, the clusters are not known until the execution if clustering algorithm. Clustering in data mining is necessary to bring scalability, the ability to deal with various types of attributes for discovering clusters with attribute shape, high dimensionality to deal with noisy data and for interpretability as the cluster results must be comprehensible. The main categories of clustering are as in Fig. 1. In this study, we bring a collation between the high lighted fields in the figure.

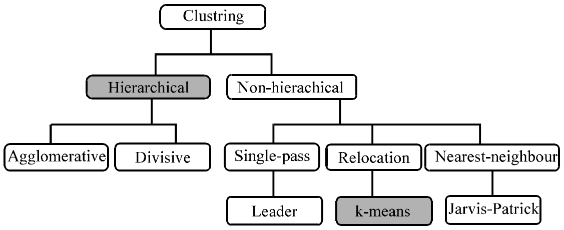


Fig. 1: Clustering hierarchy

k-means clustering algorithm: Clustering is the method of dividing a large group of objects into smaller groups of similar ones.

k-means algorithm is one of the most popular clustering algorithm used in data mining. It is also known as Lloyd’s algorithm. k-means clustering algorithm comes under unsupervised learning technique. The main advantage of using k-means over hierarchical clustering is that it is computationally faster than the latter one, if we keep k smalls. Also, k-means clustering algorithm produces more tightly coupled clusters than hierarchical clustering. But the problem lies in the prediction of K value. The final clusters are dependent on the initial partition of those clusters.

Algorithm 1; k-means clustering algorithm:

```
Step 1; From the dataset choose k objects as initial center randomly
Step 2; Each object in the cluster is reassigned with the closet center
Step 3; Mean value for each cluster is calculated thus, updating the centroid
Step 4; Repeat these steps till there is no change in the observation
```

Hierarchical clustering algorithm: Hierarchical clustering is not done in one steps. It follows repeated partitions to form clusters.

Hierarchical clustering is of two types, divisive method and agglomerative method. In agglomerative partition, each observation is considered as a single cluster. Using any one of the distance calculation, similarity of the clusters is calculated and forms most similar ones. Agglomerative hierarchical clustering can be done using the following methods. Agglomerative method can be visualized as (Fig. 2).

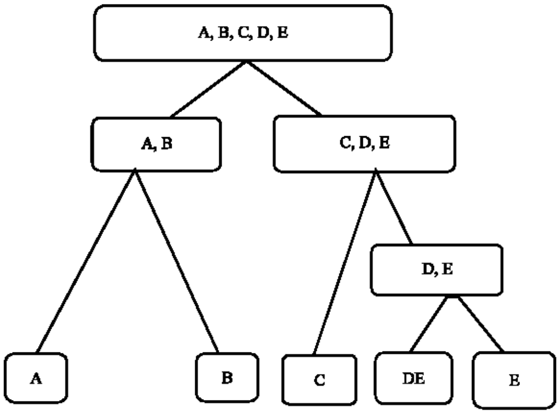


Fig. 2: Agglomerative hierarchical clustering method

- Single linkage clustering
- Complete linkage clustering
- Average linkage clustering
- While using Hierarchical clustering, there is no need to predict the number of clusters. i.e., the k value
- It produces best results and easy to implement
- But there is no scope to undo the previous step once it is completed
- Time complexity is also larger than k means clustering algorithm

RESUTS AND DISCUSSION

Contrast between hierarchical and K-means cklustering algorithms: In this study, we bring out the major contrast between hierarchical and k-means clustering algorithms in most of the study previously published claim that of k-means clustering reserach better than hierarchical. While here we claim that hierarchical clustering is better. Here, we take a dataset for instance election dataset and performed k-means and hierarchical clustering on the same dataset. By making the attribute percentage split and maxIterations constant for both the clustering algorithms, the clustered instances generated was considerably larger for hierarchical clustering (Fig. 3) rather than k-means (Fig. 4). So, from the formation of clustered instances we come to an conclusion that when accuracy and quality is considered hierarchical clustering algorithm is efficient than k-means clustering algorithm.

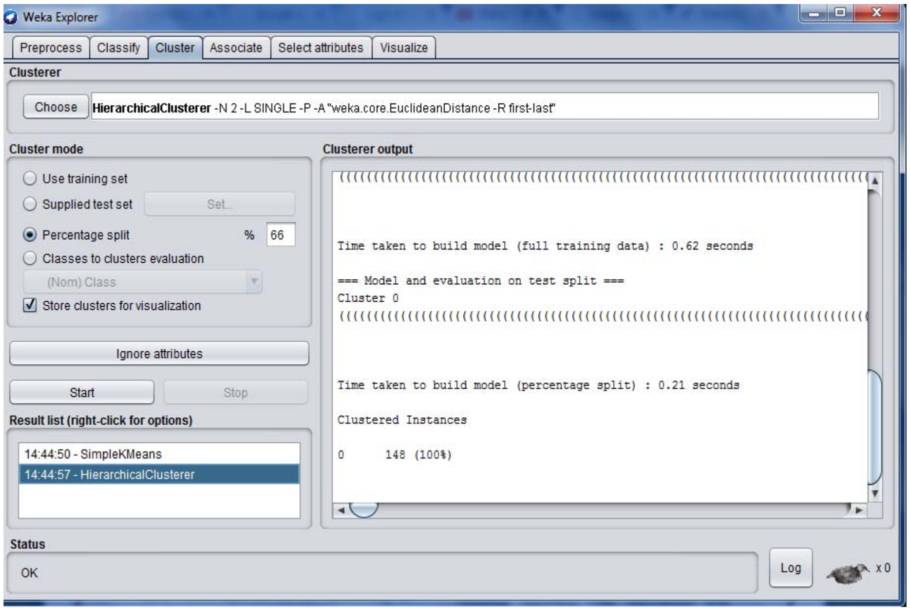
When performance of both clustering methods is taken to consideration k-means performs better than hierarchical clustering because the clusters formed will be comparatively higher for k-means while in hierarchical clusters will be only one in most cases. Figure 3 shows the output in WEKA explorer the clustered instances of k-means and hierarchical. Other comparisons are the following.

Size of dataset: k-means can research with larger dataset compared to hierarchical which can afford mostly smaller dataset. Figure 2 and 3 show the time of execution of both the clustering methods.

Time taken: k-means takes comparatively lesser time for execution than hierarchical clustering.

Proposed way of comparing k-means and hierarchical clustering algorithms using normalized data: Until now we compared k-means and Hierarchical algorithms withthe dataset which is in denormalized form. Another way of comparing both the algorithms is by using the normalized dataset. Normalization is a basic method of data transformation which is major task coming under data preprocessing. Data preprocessing is a relevant process in data mining as the data in the real world will be incomplete, noisy, redundant, inconsistent data, so in order to transform this inconvenient initial data to complete and error free data preprocessing is done. A major task in preprocessing is data transformation. Data transformation is a major process in which the data is the revolutionized and solidified to a pattern which is satisfactory for mining. Normalization and aggregation are the methods converging under data transformation.

Normalization is a means in which attribute value can be mounted to a specific scope that range between 0-1. Attribute in the dataset can be normalized by any three means, min-max normalization, score normalization or normalization by decimal scaling. After performing normalization the dataset will be of less redundancies thus, leading to betterment in the creation of high quality clusters during data clustering. So, clustering techniques once done after normalization, i.e., by using normalized data will organize clusters efficiently. As we performed k-means and hierarchical clustering with the denormalized data and found that hierarchical clustering quality and accuracy is far better than that of k-means clustering. We can come up saying that its quality will be lifted by using a normalized dataset. So in this study, we also suggest that by performing clustering in a normalized data will create more convenient clusters. Also, it will eventually increase the accuracy of hierarchical clustering when calculated using other criteria's like entropy, covariance, etc. Also, the performance of k-means will be raised than hierarchical when used with normalized data. So major variation in the performance of Hierarchical and k-means clustering can be witnessed while using normalized data.



Clusterer

HierarchicalClusterer -N 2 -L SINGLE -P -A "weka.core.EuclideanDistance -R first-last"

Cluster mode

○ Use training set
○ Supplied test set
● Percentage split % 66
○ Classes to clusters evaluation
(Nom) Class
☑ Store clusters for visualization

Clusterer output

```
(((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((



Time taken to build model (full training data) : 0.62 seconds

=== Model and evaluation on test split ===
Cluster 0
(((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((



Time taken to build model (percentage split) : 0.21 seconds

Clustered Instances

0      148 (100%)
```

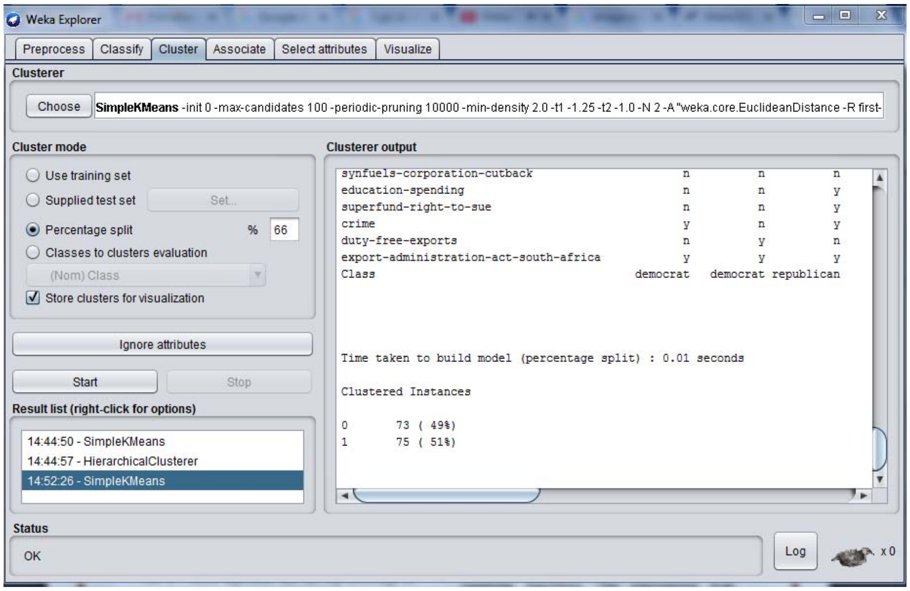
Result list (right-click for options)

14:44:50 - SimpleKMeans
14:44:57 - HierarchicalClusterer

Status

OK

Fig. 3: Clustered instances more for hierarchical clustering algorithm



Clusterer

SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-

Cluster mode

○ Use training set
○ Supplied test set
● Percentage split % 66
○ Classes to clusters evaluation
(Nom) Class
☑ Store clusters for visualization

Clusterer output

```
synfuels-corporation-cutback                         n          n          n
education-spending                                   n          n          y
superfund-right-to-sue                               n          n          y
crime                                                y          n          y
duty-free-exports                                    n          y          n
export-administration-act-south-africa               y          y          y
Class                                         democrat   democrat republican




Time taken to build model (percentage split) : 0.01 seconds

Clustered Instances

0       73 ( 49%)
1       75 ( 51%)
```

Result list (right-click for options)

14:44:50 - SimpleKMeans
14:44:57 - HierarchicalClusterer
14:52:26 - SimpleKMeans

Status

OK

Fig. 4: Clustered instances for same dataset using k-means showing lesser value

CONCLUSION

The study comes up with a comparison between Hierarchical and k-means clustering algorithms which comes to a conclusion that hierarchical clustering creates considerably more number of clustered instances thus, increasing its quality and accuracy than k-means clustering. So, it will be more recommended to use hierarchical clustering when accuracy is considered and when less execution time and performance is the major factor then hierarchical clustering runs out, k-means is better in that case. As a proposed work if clustering is performed on normalized data the clusters created will be more accurate in case of hierarchical. Also, it can handle larger dataset as it can handle only smaller dataset when it is not normalized. There will be major variation in the collation between k-means and hierarchical when using normalized data.

REFERENCES

Abbas, O.A., 2008. Comparisons between data clustering algorithms. Intl. Arab J. Inf. Technol., 5: 320-325.

Kaur, M. and U. Kaur, 2013. Comparison between K-mean and hierarchical algorithm using query redirection. Int. J. Adv. Res. Comput. Sci. Software Eng., 3: 1454-1459.

Kaushik, M. and M.B. Mathur, 2014. Comparative study of k-means and hierarchical clustering techniques. Intl. J. Software Hardware Res. Eng., 2: 93-98.

Nair, S.S., O. Rodrigues and A.J.S. Khandeparkar, 2016. Comparison between k-means and hierarchical algorithm on the basis of normalization. Intl. J. Innovative Res. Comput. Commun. Eng., 4: 11835-11839.

Rani, Y. and H. Rohil, 2013. A study of hierarchical clustering algorithm. Intl. J. Inf. Comput. Technol., 3: 1115-1122.

Sharma, N., A. Bajpai and M.R. Litoriya, 2012. Comparison the various clustering algorithms of WEKA tools. Int. J. Emerging Technol. Adv. Eng., 2: 73-80.

Srivastava, K., R. Shah, D. Valia and H. Swaminarayan, 2013. Data mining using hierarchical agglomerative clustering algorithm in distributed cloud computing environment. Int. J. Comput. Theory Eng., 5: 520-522.

Zhao, Y., G. Karypis and U. Fayyad, 2005. Hierarchical clustering algorithms for document datasets. Data Min. Knowl. Discov., 10: 141-168.