ijournals International Journal of Software & Hardware Research in Engineering
ISSN No: 2347-4890 Volume 2 Issue 6, June 2014

Comparative Study of K-Means and Hierarchical Clustering Techniques

Dr. Manju Kaushik[1], Mrs. Bhawana Mathur[2]
Associate Professor, JECRC University, Jaipur[1];
Research Scholar, JECRC University, Jaipur[2] ,

ABSTRACT

Clustering is a process of keeping similar data into groups. Clustering is an unsupervised learning technique as every other problem of this kind; it deals with finding a structure in a collection of unlabeled data. Many types of clustering methods are—hierarchical, partitioning, density –based, model-based, grid –based, and soft-computing methods. In this paper compare with k-Means Clustering and Hierarchical Clustering Techniques. Strength and weakness of both Clustering Techniques and their methodology and process.

Keywords

Data clustering, K-Means Clustering, Hierarchical Clustering, unsupervised learning technique.

1. INTRODUCTION

Clustering is a data mining technique. Clustering is grouping sets of data objects into multiple groups or clusters so that objects within the cluster have high similarity, but are very dissimilar to the objects in the other clusters. Dissimilarities and similarities are determined based on the attribute values describing the objects.

To organize data, categorize data, for data compression and model construction, for detection of outliers, etc. by Clustering algorithms .The goal of clustering is to provide measures and criteria that are used for determining whether two objects are similar or dissimilar. The aim of clustering is descriptive, and classification is predictive.

All clustering techniques are to find the cluster's center that will represent each cluster this is a Common approach.

Cluster center will represent with input vector can say which cluster this vector belongs to Researcher measuring a similarity metric between input vector and all cluster centers and determining which cluster is nearest or most similar one [1]. Cluster analysis can be used as a standalone data mining tool.

To achieve into the data distribution, or as a preprocessing step for other data mining algorithms operating on the detected clusters. Inherent geometric properties of numeric data may be exploited to naturally define a distance function between data points. Categorical data may be derived from either quantitative or qualitative data. Where observations are directly observed from the counts.

2. Literature Review

Comparisons between Data Clustering Algorithms [2]. The algorithms related with k-means algorithm, hierarchical clustering algorithm, self-organizing map algorithm and expectation maximization clustering algorithm. These algorithms are compared according to the following factors: size of the dataset, the number of clusters, type of dataset and the type of software used. Few conclusions that are extracted belong to the performance, quality, and accuracy of the clustering algorithms.

Comparative Study of Various Clustering Algorithms in Data Mining [3]. The six types of clustering techniques- k-Means Clustering, Hierarchical

Clustering, DB Scan clustering, Density Based Clustering, Optics, and EM Algorithm.

Performance analysis of k-means with different initialization methods for high dimensional data [4]. In this paper, Researchers have analyzed the performance of our proposed method with the existing works.

K-Means Clustering Algorithm for High Dimensional Data Set using Dimensionality Reduction Method [5]. K-means clustering algorithm often does not work well for high dimension, hence, to improve the efficiency, apply PCA to the original data set and obtain a reduced dataset containing possibly correlated with variables. In this paper principal component analysis and linear transformation is used for dimensionality reduction and initial centroid is computed, and then applied to K-Means clustering algorithm.

Evolving limitations in K-means algorithm in data mining and their removal" [6]. K-means is popular because it is conceptually simple and is computationally fast and memory efficient, but various types of limitations in k means algorithm that makes extraction difficult.

Comparative Analysis of Two Algorithms for Intrusion Attack Classification Using KDD CUP Data set [7].

Compression, Clustering, and Pattern Discovery in Very High-Dimensional Discrete-Attribute Data Set [8].

A Hierarchical Latent Variable Model for Data Visualization [9].Researchers introduce a hierarchical visualization algorithm which allows the complete data set to be visualized at the top level, with clusters and sub clusters of data points visualized at deeper levels.

A Modified K-Means Algorithm for Circular Invariant Clustering [10]. Several important pattern recognition applications are based on feature vector extraction and vector clustering. Enhanced Moving K-Means (EMKM) Algorithm for Image Segmentation [11].

A Modified Version of the K-Means Algorithm with a Distance Based on Cluster Symmetry [12]. In this paper, Researchers propose a modified version of the K-means algorithm to cluster data.

An Efficient k-Means Clustering Algorithm: Analysis and Implementation [13]. In this paper, Researchers present a simple and efficient implementation of Lloyd's k means clustering algorithm, which Researchers call the filtering algorithm. This algorithm is easy to implement, requiring a kd-tree as the only major data structure.

A Modified k-means Algorithm to Avoid Empty Clusters [14]. This paper presents a modified version of the *k*-means algorithm that efficiently eliminates this empty cluster problem. Researchers have shown that the proposed algorithm is semantically equivalent to the original *k*-means and there is no performance degradation due to incorporated modification. Results of simulation experiments using several data sets to prove our claim.

Comparing the various clustering algorithms of Weka tool[15]. In Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. Their main aim to show the comparison of the different-different clustering algorithms on WEKA and find out which algorithm will be most suitable for the users.

3. DATA CLUSTERING TECHNIQUES

3.1 K-Means Clustering

It is a partition method, a technique which finds mutual exclusive clusters of spherical shape. A specific number of disjoint, flat (non-hierarchical) clusters are generated.

The statistical method can be used to cluster to assign rank values to the cluster categorical data. Categorical data have been converted into numeric by assigning rank value. K-Means algorithm organizes objects into k – partitions.

Where each partition represents a cluster. Researchers start out with the initial set of means and classify cases based on their distances to their centers. Next, Researchers compute the cluster means again, using the cases that are assigned to the clusters; then, Researchers reclassify all cases based on the new set of menus. Researchers keep repeating this step until cluster means don't change between successive steps. Finally, Researchers calculate the means of cluster once again and assign the cases to their permanent clusters.

3.1.1 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.

3.1.2 K-Means Algorithm Process

The dataset is partitioned into K clusters and the data points are randomly assigned to the clusters resulting in clusters that have roughly the same number of data points.

For each data point:

- Calculate the distance from the data point to each cluster.
- If the data point is closest to its own cluster, leave it where it is. If the data point is not closest to its own cluster, move it into the closest cluster.
- Repeat the above step until a complete pass through all the data points' results in no data point moving from one cluster to another. At this point the clusters are stable and the clustering process ends.
- The choice of initial partition can greatly affect the final clusters that result, in terms of inter-cluster and intra cluster distances and cohesion [6].

3.1.2.1 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

3.1.2.2 Weaknesses of k-means

- The algorithm is only applicable if the mean is defined.
  - ✓ For categorical data, *k*-mode - the centroid is represented by most frequent values.
- 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.

  Outliers could be errors in the data recording or some special data points with very different values.
- Weaknesses of k-means: To deal with outliers
- One method is to remove some data points in the clustering process that are much further away from the centroids than other data points.
  - ✓ To be safe, we may want to monitor these possible outliers over a few iterations and then decide to remove them.

- Another method is to perform random sampling. Since in sampling we only choose a small subset of the data points, the chance of selecting an outlier is very small.
  - ✓ Assign the rest of the data points to the clusters by distance or similarity comparison, or classification
- The k-means algorithm is not suitable for discovering clusters that are not hyper-ellipsoids (or hyper-spheres).

In Short, K Means is:-

- Despite weaknesses, k-means is still the most popular algorithm due to its simplicity, efficiency and
  - ✓ Other clustering algorithms have their own lists of weaknesses.
- No clear evidence that any other clustering algorithm performs better in general
  - ✓ Although they may be more suitable for some specific types of data or applications.
- Comparing different clustering algorithms is a difficult task. No one knows the correct clusters!

3.2 Hierarchical Clustering

A hierarchical method creates a hierarchical decomposition of the given set of data objects. A dendrograms is built due to Tree of clusters. Every cluster node contains child clusters, sibling cluster partition the points covered by their common parent.

In hierarchical clustering, Researchers assign each item to a cluster such that if Researchers have N items then Researchers have N clusters. Find closest pair of clusters and merge them into a single cluster. Compute distance between new cluster and each of old clusters. Researchers have to repeat these steps until all items are clustered into K no. of clusters.

3.2.1 Strengths of Hierarchical Clustering

- 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.

3.2.2 Weakness of hierarchical Clustering

- Merging /splitting of clusters is permanent => Erroneous decisions are impossible to correct later.
- Divisive methods can be computational hard.
- Methods are not (necessarily) scalable for large datasets.

Table 1. Types of Hierarchical Clustering

Hierarchical Clustering	
Agglomerative (bottom up)	**Divisive (top down)-**
Start with each document being a single cluster.	Starts with all documents belong to the same cluster.
Eventually all documents belong to the same cluster	Eventually each node forms a cluster on its own.

- Does not require the number of clusters ***k*** in advance.
- Needs a termination/readout condition.
- The final mode in both Agglomerative and Divisive is of no use.

3. Comparisons on K –Means and Hierarchical Clustering

Properties	K –Means	Hierarchical Clustering
Definition	K Means Clustering generates a specific number of disjoint, flat (non-hierarchical) Clusters.	Hierarchical Clustering method construct a hierarchy of Clustering, not just a single partition of objects.
Clustering Criteria	It is well suited to generating globular Cluster.	Use a distance matrix as Clustering Criteria. A termination Condition can be used .Example –A number of Clusters.
Performance	The performance of K- mean algorithm is better than Hierarchical Clustering Algorithm.	Hierarchical Clustering Algorithm performance is less as compare to K- mean algorithm.
Category Data	K- Means can be used in categorical data is first converted into numeric by assigning rank.	Hierarchical algorithm was adopted for categorical data, and due to its complexity a new approach for assigning rank value to each categorical attribute.
Sensitive To Noise	K-Means is very sensitive to noise in the dataset.	It is less sensitive to noise in the dataset.
Cluster	There are always K.	The number of Clusters k is not required as an input.
Execution Time	K -mean algorithm also increases its time of execution.	Hierarchical algorithm its performance is better.
Quality	K-Means algorithms Shows less quality.	Hierarchical algorithm shows more quality.
Data Set	k -mean algorithm is good for large dataset.	Hierarchical is good for small datasets.

5. CONCLUSION

The K - mean algorithm has the big advantage of clustering large data sets and its performance increases as the number of clusters increases. But these conditions apply when Researchers use is limited to numeric values.

Hierarchical algorithm was adopted for categorical data, and due to its complexity a new approach for assigning rank value to each categorical attribute using K- means can be used in which categorical data is first converted into numeric by assigning rank.

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 ambiguity in some (noisy) 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.

When using huge dataset, K-Means algorithm is faster than other clustering algorithm and also produces quality clusters.

6. Future Scope

As a future work, comparison between K-Means and Hierarchical Clustering Technique can be implemented on the basis of normalization, by taking normalized and un-normalized data will give different results.

7. References

[1] Joshi, A., Kaur, R., A Review: Comparative study of various clustering techniques in data mining, International Journal of Advanced Research in Computer Science and Software Engineering, 3 (1), March, 2013.

[2] 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.

[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] Napoleon, D., Pavalakodi, S., "A New Method for Dimensionality Reduction using K-Means Clustering Algorithm for High Dimensional Data Set", International Journal of Computer Applications (0975– 8887), Vol. 13, no. 7, pp. 41-46, Jan 2011.

[6] Singh, K., Malik D., Sharma, N.,"Evolving limitations in K-means algorithm in data mining and their removal, "IJCEM International Journal of Computational Engineering &Management, vol. 12, pp. 105-109, Apr. 2011.

[7] 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.

[8] 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.

[9] Bishop C. M. , Michael, E. Tipping, "Hierarchical Latent Variable Model for Data Visualization", IEEE Trans. Pattern Anal. Mach Intell., 20 (3), 281-293, 1998.

[10] Charalampidis C. M.**,,"** A Modified K-Means Algorithm for Circular Invariant Clustering ", IEEE Transactions on Pattern Analysis and Machine Intelligence, 27 (12), 2005.

[11] Siddiqui ,F., Isa A.M.,,"Enhanced Moving K-Means (EMKM) Algorithm for Image Segmentation ", IEEE Trans. Consumer Electronics, 12 (4), 2014.

[12] Su ,M.C,Chou,C.H.,"A Modified Version of the K-Means Algorithm with a Distance Based on Cluster Symmetry", IEEE Transactions on Pattern Analysis and Machine, 23 (6), Aug 7, 2002.

[13] Kanungo, T., David M., Piatko C.D., Silverman,R.,Angela Y.Wu, An Efficient k-Means Clustering Algorithm: Analysis and Implementation, IEEE Transactions on Pattern Analysis and Machine Intelligence, 24,881-892,2002.

[14] Pakhira, M. K.," A Modified k-means Algorithm to Avoid Empty Clusters", International Journal of International Journal of Recent Trends in Engineering 1 (01), 2009

[15] 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.

[16] Han, J., Kamber, M. Data Mining: Concepts and Techniques, 3rd Ed, 443-491, 2012.

[17]Grabmeier, J., Manila,F., "Techniques of Cluster Algorithms in Data Mining", Data Mining and Knowledge Discovery, 6, 303–360, 2002.

[18] Patnaik, S. K., Sahoo,S., Swain D.K.,, "Clustering of Categorical Data by Assigning Rank through Statistical Approach," International Journal of Computer Applications 43 (2), 1-3, 2012.

[19] Arockiam, L., Baskar, S. S. Jeyasimman. L.,"Clustering Techniques in Data Mining: A Review", Asian Journal of Information Technology, 11 (1), 40-44,2012.