ISSN: 2321-7782 (Online)
e-ISJN: A4372-3114
Impact Factor: 7.327

Volume 6, Issue 4, April 2018

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

Survey Paper on Clustering Techniques of Data Mining

Bhoomika Batra[1]
Research Scholar, Department of Computer Engineering
YMCAUST
Faridabad, India

Shilpa Sethi[2]
Assistant Professor, Department of Computer Engineering
YMCAUST
Faridabad, India

Dr. Ashutosh Dixit[3]
Associate Professor, Department of Computer Engineering
YMCAUST
Faridabad, India

Abstract: ***Data mining is an active and interesting research area that has emerged to discover knowledge from large amount of continuously generated data. Clustering is an unsupervised technique of data mining where the set of similar objects are grouped in one cluster. In this context, several clustering algorithms have been proposed by the researchers in the past. The knowledge discovery process intrinsically requires development of clustering algorithms capable of performing incremental processing of data objects in a quick fashion. The major challenges faced by existing clustering algorithms are to deal with unbound, non-stationary and unstructured data. The paper surveys various clustering algorithms and provides a thorough discussion on their major components. In addition, the work also identifies strengths and weakness of each algorithm. Finally, a comparative analysis among prevalent clustering algorithms is provided.***

Keywords: Data Mining, Clustering, Classification, Hierarchical Clustering, Partitioning Clustering.

I. INTRODUCTION

In order to find out some useful and hidden information among the data sets, data mining is used [15]. Traditional techniques of data mining performed well till the data sets were small in size. But with the advent of internet data sets accessible to users increased tremendously. Handling such a large volume of data and obtaining useful information from them by using traditional techniques of data mining is a tedious task.

The data mining techniques suffers from several challenges while extracting information from large datasets due to i) data is raw ii) data is incomplete iii) data is uncertain. Clustering is one of the most important techniques used in data mining to group similar objects together [16]. It is an unsupervised learning technique used for finding hidden patterns in the raw data and helps in making data mining techniques more efficient. It plays a very significant role in data mining applications like text mining, web analysis, information retrieval, customer relationship management (CRM), medical diagnostics etc.

In this paper a survey of clustering techniques and their comparison has been conducted. Section 2 consists of clustering overview, classification of clustering techniques on different basis and discussion of these methods in detail. Comparison between prevalent clustering techniques is done in section 3. Section 4 concludes this survey with a brief discussion of future scope in the field of clustering.

II. CLUSTERING OVERVIEW AND ITS CLASSIFICATION

Clustering [1, 2] also known as cluster analysis is defined as a technique in which data objects available in data sets which are partitioned into groups called clusters. Each cluster comprises of data objects which are similar among themselves but they are not similar to data objects contained in other groups.

Classification of clustering techniques [3] is not straightforward. Subtypes of clustering algorithms overlap with each other. Different authors provide different classification of clustering techniques. This paper discusses the classification of clustering techniques on the basis of their cluster model (as shown in Fig. 1)-

- Hierarchical clustering
- Partitioning clustering

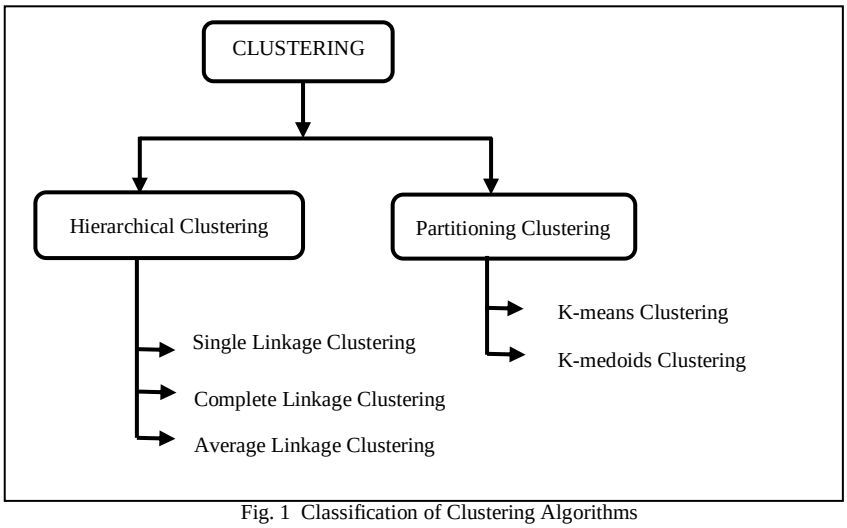

Fig. 1 Classification of Clustering Algorithms

Details of these subtypes are described in subsequent sections.

A. Hierarchical Clustering

This technique uses algorithms [4] which build clusters gradually over the time. Basic process of hierarchical clustering when set of N data objects which needs to be clustered is given includes following steps-

Step 1. Each item is first assigned to clusters, hence if N data objects are given, N clusters are formed. Assume distances (or similarities) among the clusters as equivalent to distances among the data objects contained in clusters.

Step 2. Closest pair (pair which is most similar) of clusters is found and merged to form a single cluster and in result number of clusters get decreased by one.

Step 3. Distance (similarity) between new cluster and every old cluster is computed.

Steps 2 & 3 are repeated until a single cluster is formed which has size N.

Step 3 can be implemented in three different ways which led to further classification of hierarchical algorithms into three subtypes-

- Single linkage clustering
- Complete linkage clustering
- Average linkage clustering

While performing all the three categories of hierarchical clustering, a proximity matrix of size N*N is taken and represented as is D = [d (i, j)]. The clusterings are arranged in sequence and are given sequence numbers 0, 1…. (n-1).Level of

kth clustering is represented as L (K). A cluster having sequence number m is represented (m).The proximity between clusters (s) and (q) is represented d [(s), (q)].

Single Linkage Clustering: This type of clustering [5, 6] is also known as connectedness or minimum clustering method. In this technique distance between one cluster and another cluster is considered same as shortest distance which can be calculated from any item of any cluster to any item of other cluster. The algorithm for single linkage clustering is given in fig 2-

Single linkage cluster ( )

Input: n clusterings arranged in sequence number as 0,1…….,(n-1) having one item each;

Output: A cluster C having all n items.

Method-

Step 1. Start working with the clustering which has sequence number m = 0 and level L (0) = 0.

Step 2.Find the pair which is most dissimilar pair of clusters in current clustering, for an example take pair (s), (q), using equation-

$$d[(s),(q)] = \min d[(i),(j)] \quad \text{.....eq(1)}$$

Step 3. Sequence number is incremented by 1.

m= m +1

Clusters (s) and (q) are merged into a single cluster and the next clustering m is formed. Level of this clustering is set to

L (m) = d [(s), (q)]

Step 4. Proximity matrix is updated by erasing the rows and columns which belong to clusters (s) and (q) and a row and column which belongs to new cluster is added. Proximity among the new cluster and old cluster is calculated using equation-

$$d[(k), (s, q)] = \min d[(k),(s)], d[(k),(q)] \quad \text{.....eq(2)}$$

Step 5. If all the items are places in same cluster, stop. Otherwise step 2 is repeated.

Fig. 2 Algorithm for Single Linkage Clustering

This technique works very efficiently for non-global shape clusters. But this technique is sensitive to noise and outliers and hence leads to chaining effects. Moreover this technique, most of the time generates long and elongated clusters.

Complete Linkage Clustering: This type of clustering [7, 8] is also known as diameter or maximum clustering method. In this technique distance between one cluster and another cluster is considered same as greatest distance which can be calculated from any item of any cluster to any item of other cluster. The algorithm for complete linkage clustering is given in Fig. 3.

Complete linkage cluster ( )

Input: n clusterings arranged in sequence number as 0,1…….,(n-1) having one item each;

Output: A cluster C having all n items.

Method-

Step 1. Start working with the clustering which has sequence number m = 0 and level L (0) = 0.

Step 2.Find the pair which is most dissimilar pair of clusters in current clustering, for an example take pair (s), (q), using equation-

$$d[(s),(q)] = \max d[(i),(j)] \quad \text{.....eq(3)}$$

Step 3. Sequence number is incremented by 1.

m= m +1

Clusters (s) and (q) are merged into a single cluster and the next clustering m is formed. Level of this clustering is set to

L (m) = d [(s), (q)]

Step 4. Proximity matrix is updated by erasing the rows and columns which belong to clusters (s) and (q) and a row and column which belongs to new cluster is added. Proximity among the new cluster and old cluster is calculated using equation-

d[(k), (s, q)] = max d[(k),(s)], d[(k),(q)]……..eq(4)

Step 5. If all the items are places in same cluster, stop. Otherwise step 2 is repeated.

Fig. 3 Algorithm for Complete Linkage Clustering

This technique is not susceptible to noise and outliers and hence solves the problem of chaining which occurs in single linkage clustering technique. Moreover more balanced clusters having almost equal diameter are generated using this technique. But in cases when diameter becomes very large this technique does not perform efficiently.

Average Linkage Clustering: In this technique [9] distance between one cluster and another cluster is considered same as greatest distance which can be calculated from any item of any cluster to any item of other cluster. The algorithm for average linkage clustering is given in Fig. 4.

Average linkage cluster ( )

Input: n clusterings arranged in sequence number as 0,1…….,(n-1) having one item each;

Output: A cluster C having all n items.

Method-

Step 1. Start working with the clustering which has sequence number m = 0 and level L (0) = 0.

Step 2.Find the pair which is most dissimilar pair of clusters in current clustering, for an example take pair (s), (q), using equation-

d [(s),(q)] = Tsq/(N*N)……..eq(5)

Tsq is here is used as representation of sum of distance pairs of cluster s and cluster q.

Step 3. Sequence number is incremented by 1.

m= m +1

Clusters (s) and (q) are merged into a single cluster and the next clustering m is formed. Level of this clustering is set to

L (m) = d [(s), (q)]

Step 4. Proximity matrix is updated by erasing the rows and columns which belong to clusters (s) and (q) and a row and column which belongs to new cluster is added. Proximity among the new cluster and old cluster is calculated using equation-

$$d[(k), (s,q)] = \frac{(N*d[(k),(s)])+(N* d[(k),(q)])}{N*N} \text{ ...eq(6)}$$

Step 5. If all the items are places in same cluster, stop. Otherwise step 2 is repeated.

Fig. 4 .Algorithm for Average Linkage Clustering

This technique is similar to complete average clustering in the sense that it is also not susceptible to noise and outliers and hence solves the problem of chaining which occurs in single linkage clustering technique and also in cases when cluster size becomes very large this technique does not perform efficiently. Difference lies in the way of finding proximity between clusters which in turn reduces the time taken to perform the average linkage clustering.

It may be noted that complexity of single linkage clustering algorithm and complete linkage clustering algorithm is O (n^2) and complexity of average linkage clustering algorithm is O (n^2 log n). Average linkage clustering algorithm has better time complexity than single linkage clustering algorithm and complete linkage clustering algorithm and hence works more efficient than them.

B. Partitioning Clustering Algorithms:

These kinds of algorithms build clusters directly. In these kinds of algorithms division of set on N data objects is done into non-overlapping clusters or subsets in a form that each subset or cluster contains exactly one data object. The disadvantage of these kinds of algorithms is that in some cases due to overlapping the result becomes misleading.

Clustering, in general, have a major problem associated with them and that problem is choosing a value of k i.e. number of output clusters. Partitioning techniques [10] help in taking this major design decision. This is done by producing clusters having optimized criterion function which is defined either on the subset of patterns i.e. globally or on all patterns i.e. locally. For this algorithm is run several times having different starting state each time. After all the runs best possible configuration is obtained from them which is then used as output clustering.

When a data set of n data objects is given and K is the number of clusters to be formed, basic principle of these kinds of algorithms includes organizing n data objects into k partitions(k<=n).Here each partition is considered as a cluster. Data objects of same clusters are similar and data objects of different clusters are dissimilar. Partitioning clustering techniques are of two types-

- k-means clustering
- k-medoids clustering

K-means Clustering: It is a popular method used for clustering in data mining [11].It groups given set of data objects into clusters on the basis of their proximity to each other using square error function. For which first mean value or centre is found by randomly selecting a data object. After that most similar objects are assigned in same clusters on the basis of mean value. This is done by finding data objects pairs which have minimum squared error value than all the other data objects pair. This algorithm finds k partitions i.e. clusters which minimizes square error function. Algorithm for k-means clustering is given in Fig. 5.

k-means cluster( )
Input: Data set having n data objects D={x1,x2,…..,xn};
Output: A set of k clusters C= {C1,C2,…..,Cn};

Method-
Step1. k objects are selected as initial clusters centres. This selection is done randomly.
Step2. While (no change in any cluster is done)
{
2.1.Most similar cluster on the basis of mean value or centre of data objects using equation-

$$E = \sum_{i=1}^{k} \sum_{xp \epsilon Ci} (xp - Mi)^2 \dots\dots\dots\dots (7)$$

where:
- E represents sum of squared error for each data object in set of n data objects
- xp represents data point i.e. data object
- Mi represents mean or centre of cluster Ci

2.2. Cluster mean or centre is updated using equation-

$$Mi^{k+1} = \frac{1}{|E^k|} \sum_{xi \in E^k} xi \dots\dots\dots\dots eq(8)$$

}

Fig. 5.Algorithm for k-means Clustering

K-means clustering technique works very efficiently when large data sets needs to be clustered and it is more scalable than k-medoids clustering algorithm. This technique can be applied only in cases where centre or mean can be defined. Moreover this technique is very sensitive to noise and outlier data objects which can affect centre or mean value because due to presence of noise or outlier data objects mean gets shifted from its original position.

K-medoids Clustering: This technique [12] was used to reduce sensitivity of k-means clustering to noisy data and outliers.It uses absolute error function to group two sets of data objects. Actual objects are used instead of mean values which is known as representative of data object i.e. medoid which at initial stages is selected randomly. After that each representative data object is replaced by non-representative data object. This process continues till quality of clustering stops improving. Algorithm for k-medoids is given in Fig. 6.

k-medoids cluster( )
Input: Data set having n data objects D={x1,x2,…..,xn};
Output: A set of k clusters C={C1,C2,…..,Cn};
Method-
Step1. k objects are selected as representative data objects. This selection is done randomly.
Step2. Each remaining data object is assigned to its most cluster having most similar representative data object using equation-

$$E = \sum_{i=1}^{k} \sum_{xp \epsilon Ci} (xp - Mi)^2 \dots\dots\dots\dots eq(7)$$

where:
- E represents sum of absolute error for each data object in set of n data objects
- xp represents data point i.e. data object
- Oi is representative object of cluster Ci

Step3. While (no change in clusters or medoids occurs)
{

3.1. For each representative object Oi
{

3.1.1.A non-representative object Orandom is randomly selected.
3.1.2.Total cost S of swapping is calculated.
3.1.3If(S<0) Oi is replaced with Orandom.
}
}

Fig. 6.Algortithm for k-medoids Clustering

K-medoids clustering technique is robust because it very sensitive to noise and outlier data objects and do not get much affected by their presence. Specification of k is user dependent in this technique as that of k-means clustering technique. But this technique is more complex than k-means clustering. It also becomes very expensive in case of large values of k and n and it is not scalable like k-means clustering technique and hence difficult to apply in case of large datasets.

It may be noted that complexity of algorithm used in k-means clustering technique is O (nkt) where t is representation of number of iterations performed while applying the algorithm and complexity of algorithms used in k-medoids clustering technique is O (k (n-k) ^{2}). k-medoids clustering does not work well for large values of k and n.

III. COMPARISON BETWEEN CLUSTERING TECHNIQUES

Comparison between various clustering techniques [13, 14] is provided in Table 1.

TABLE I Comparison between Clustering Techniques

Summarization method	Main idea	Complexity	Advantages	Disadvantages
Single linkage clustering	In this technique distance between one cluster and another cluster is considered same as shortest distance which can be calculated from any item of any cluster to any item of other cluster.	Complexity of this technique is O (n^2).	1. This technique works very efficiently for non-global cluster shapes.	1. This technique generates long and elongated clusters. 2. This technique is very sensitive to noise and outliers.
Complete linkage clustering	In this technique distance between one cluster and another cluster is considered same as greatest distance which can be calculated from any item of any cluster to any item of other cluster.	Complexity of this technique is O (n^2).	1. This technique is less susceptible to noise and outliers. 2. More balanced clusters having equal diameter are generated using this technique.	1. This technique does not perform well in cases of large clusters.
Average linkage clustering	In this technique distance between one cluster and another cluster is considered same as greatest distance which can be calculated from any item of any cluster to any item of other cluster.	Complexity of this technique is O (n^2 logn).	1. This technique is less susceptible to noise and outliers. 2. This technique is more efficient and have less time complexity.	1. This technique does not perform well in cases of large clusters.
K-means clustering	It groups given set of data objects into clusters on the basis of their proximity to each other using square error function. For which first mean value or centre is found by randomly selecting a data object. After that most similar objects are assigned in same clusters on the basis of mean value. This is done by finding data objects pairs which have minimum squared error value than all the other data objects pair.	Complexity of this technique is O (nkt)	1. When large data sets need to be clustered this technique works more efficient than all other clustering techniques. 2. This algorithm is also more scalable than k-medoids clustering algorithm and other partitioning based clustering algorithms.	1. This technique can only be applied when mean or centre of a cluster can be defined. 2. Specification of k by users is compulsory before applying this algorithm. 3. It is very sensitive to noise and outlier data objects which can have an influence in finding mean value.
K-medoids clustering	It uses absolute error function to group two sets of data objects. Actual objects are used instead of mean values which is known as representative of data object i.e. medoid which at initial stages is selected randomly. After that each representative data object is replaced by non-representative data object. This process continues till quality of clustering stops improving.	Complexity of this technique is O (k (n-k) ^{2})	1. This technique is more robust in case noisy data or outliers are present as it is less sensitive to them.	1. This technique uses algorithms which are more complex than k-means clustering algorithm. 2. Specification of k by users is compulsory before applying it. 3. It does not work well in case of large datasets i.e. it is not as scalable as k-means clustering. 4. For large values of k and n this technique becomes costly than k-means clustering algorithm.

All the techniques have their advantages and disadvantages as well. User selects different clustering algorithms according to their requirements and situations.

IV. Conclusion and Future Scope

For analyzing the large data sets, various clustering methodologies are discussed in this paper. The techniques are evaluated on the basis of execution time and cluster quality. Paper also discusses merits and de-merits of each technique. It is observed after evaluating these techniques that they lack in clustering large volumes of data sets. Each technique suffers from one or other limitation. Single linkage and k-means techniques are sensitive to noisy data and outliers. Complete and average linkage techniques fail in case of large clusters. k-medoids technique is very complex and costly. A lot of research work is still needed to improve accuracy and efficiency of existing clustering algorithms

References

1. https://en.wikipedia.org/wiki/Cluster_analysis
2. XiaoyeWanga, XiaoruiChaia, Ching-HsienHsua, YingyuanXiaoa, Yukun Lia,"cluster analysis based on opinion mining," in 8th International Conference on Ubi-Media Computing,ISBN:978-1-4673-8270-0,2015.
3. P. S. Badase,G. P. Deshbhratar,A. P. Bhagat,"Classification and analysis of clustering algorithms for large datasets," in IEEE Sponsored 2nd International Conference on Innovations in Information Embedded and CommunicationSystems, ISBN:.978-1-4799-6818-3,2015.
4. https://en.wikipedia.org/wiki/Cluster_analysis#Connectivitybased_clustering_(hierarchical_clustering).
5. https://en.wikipedia.org/wiki/Single-linkage_clustering
6. https://onlinecourses.science.psu.edu/stat555/node/86
7. https://en.wikipedia.org/wiki/Complete-linkage_clustering
8. https://nlp.stanford.edu/IR-book/completelink.html
9. https://en.wikipedia.org/wiki/UPGMA
10. A.Dharmarajan ,T.Velmurugan,"Applications of partition based clustering algorithms: a survey," in IEEE International Conference on Computational Intelligence and Computing Research,ISBN:978-1-4799-1597-2,2013.
11. https://en.wikipedia.org/wiki/K-means_clustering
12. https://en.wikipedia.org/wiki/K-medoids
13. Preeti Arora, Dr. Deepali , ShipraVarshney, "Analysis of k-means and k-medoids algorithm for big data", in International Conference on Information Security & Privacy,Vol.3 No.2,Dec.2015.
14. JiWentian, GuoQingju,ZhongSheng,Zhou En,"improved k-medoids clustering algorithm under semantic web," in Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering,2015.
15. S. Sethi, A. Dixit, " An adaptive web search system based on web usages mining" International journal of computer engineering and application , Vol. X , No. 1, ISSN: 23213469 ,2016.
16. S. Sethi, A. Dixit ,"An Automatic User Interest Mining Technique for Retrieving Quality Data" International Journal of Business Analytics. Vol 4, No. 2, pp 62-79, ISSN: 2334-4547 ,2017.

AUTHOR(S) PROFILE

Bhoomika Batra, is currently pursuing M. Tech. in Computer Engineering YMCA University of Science & Technology, Faridabad Haryana. She is a currently a research scholar and doing research in the field of data mining. She has received her B. Tech degree from MD University, Rohtak, in the year 2014.

Shilpa Sethi, has received her Master in Computer Application from Kurukshetra University, Kurukshetra in the year 2005 and M. Tech. in Computer Engineering from MD University Rohtak in the year 2009. She is currently pursuing PhD in Computer Engineering and serving as Assistant Professor in the department of computer engineering at YMCA University of Science & Technology, Faridabad Haryana. She has published more than ten research papers in various International journals and conferences. Her area of research includes Internet Technologies, Web Mining and Information Retrieval System.

Dr. Ashutosh Dixit, received his PhD and M. Tech. in Computer Engineering from MD University Rohtak, in the years 2010 and 2004 respectively. He is presently serving as Associate Professor in the department of computer engineering at YMCA University of Science & Technology, Faridabad Haryana. He has published around 80 research papers in various International journals and conferences. His research interests include Internet Technologies, Data Structures and Mobile and Wireless networks.

