Volume 3, Issue 11, November 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Survey Paper on Recent Clustering Approaches in Data Mining

Amanpreet Kaur Toor
Student of Master of Technology
Department of Computer Science & Engineering,
Amritsar College of Engineering &Technology,
Manawala, Amritsar, Punjab, India

Amarpreet Singh
Associate Professor
Department of Computer Science & Engineering,
Amritsar College of Engineering &Technology,
Manawala, Amritsar, Punjab, India.

Abstract:-A significant issue of data mining task is fast retrieval of the relevant information from databases. The important goal of data mining process is to collect the information from large dataset and then transform that data into some meaningful manner. Clustering is one of the most essential task in data mining. A clustering algorithm partition a data set into several groups so that data objects that share some kind of similarity can be placed and one group and the data objects that are dissimilar can be placed into different group. In this paper presents a survey about recent clustering algorithm that are K-Means, Kohenen SOM, HCA. The main objective of this paper is to discover detailed concepts and techniques related with clustering.

Keywords:- Data Mining, Clustering, Types of Clusters, Clustering algorithm

I. INTRODUCTION

Data Mining process can be defined as the mining or discovery of new information in terms of patterns or rules from vast amount of data set. Data mining is a process that takes data as input and outputs knowledge. It is also called KDD (Knowledge Discovery in Databases), it is the process of analyzing data from the different prospective and summarizing it into useful information. According to SAS Institute Inc."Data Mining is advanced methods for exploring and modeling relationships in large amounts of data"*[1]*. Data Mining process includes following basic steps:-

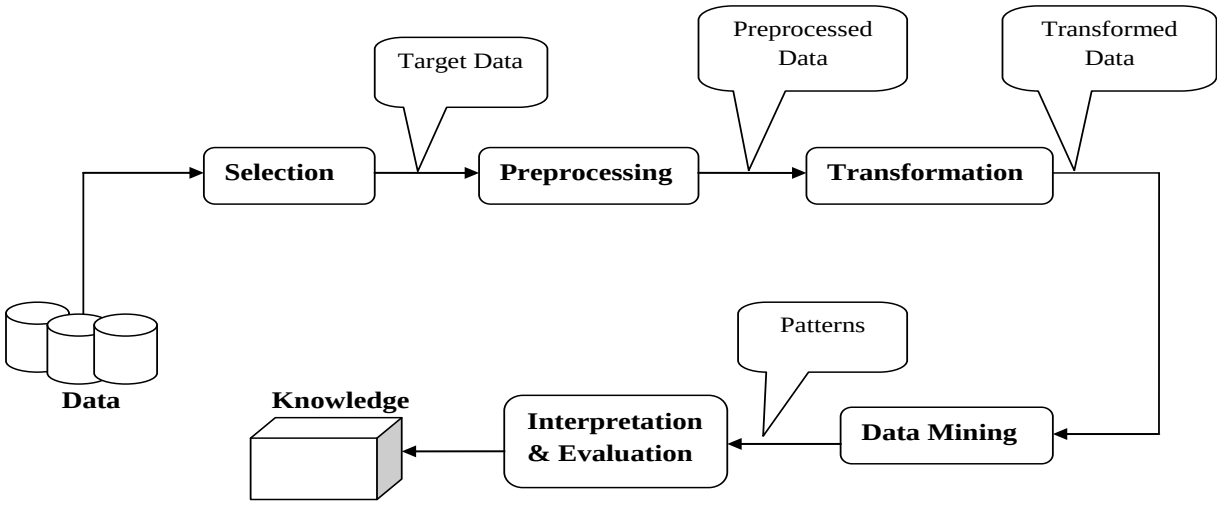

Figure 1: Steps of Data Mining Process

The process of Data Mining consists of iterative sequence methods as follows:-
1. *Selection*: This step deals with the Selection of data relevant to the analysis task from the database.
2. *Preprocessing*: Removing noise and inconsistent data and combining multiple data sources.
3. *Transformation*: Transforming data into appropriate forms to perform data mining.
4. *Data mining*: Choosing a data mining algorithm which is appropriate to pattern in the data; extracting data patterns.
5. *Interpretation/Evaluation*: Interpreting the patterns into knowledge by removing repeated or unrelated patterns from database and then translating the useful patterns into terms into human understandable format.

II. CLUSTERING

Clustering response is a primitive exploratory approach in data analysis with little or no prior knowledge about the number of clusters. Clustering can be can be considered as the most important unsupervised learning technique. In

clustering a group of data objects is divided into a number of homogenous subgroups on the basis of some chosen measure of similarity. Any cluster should exhibit two main properties:-

1. High intra-class similarity
2. Low inter-class similarity

Clustering can be used as the pre-processing step to separate data into manageable parts *[2,3]*, as knowledge discovery tool *[4,5]*, indexing and compression *[6]*, etc. The most popular use of clustering is to allot labels to unlabelled data-data for which no pre-existing grouping is known. Clustering can be used in any field where data is processed and utilized. The problem province and applications of clustering are numerous.

Advantages: 1. Automatic recovery from failure that is recovery without user intervention.

2. It provides incremental growth to group new data, because it is impossible to collect all data before starting clustering with the increased use of personal computers and mobile technology.

Disadvantages: 1. Inability to recover from the database corruption.

III. TERMS

Cluster: A cluster is an group of data objects, which have some common features. The objects belong to an interval [a,b], in our case [0,1].

Cluster centroid:

The centroid of a cluster is a point whose parameter values are the mean of the parameter values of all the points in the clusters.

$$\mathbf{m} = \frac{1}{|C|}\sum_{\mathbf{x}\in C}\mathbf{x}$$

Distance:

The distance between two clusters involves some or all elements of two clusters. The clustering method determines how the distance should be computed. The commonly used distance measure is the Euclidean metric which defines the distance between two points p= (p1, p2, ...) and q = (q1, q2, ...) is given by :

$$d = \sqrt{\sum_{i=1}^{k} (\mathbf{p}_i - \mathbf{q}_i)^2}$$

Radius: The Square root of average distance from any point from any point of the cluster to cluster centroid.

$$R = \sqrt{\frac{\sum_{\mathbf{x}\in C}(\mathbf{x}-\mathbf{m})^2}{|C|}}$$

Diameter: Square root of average mean squared distance between all pairs of points in the cluster.

$$D = \sqrt{\frac{\sum_{\mathbf{x}\in C}\sum_{\mathbf{y}\in C,\mathbf{y}\neq\mathbf{x}}(\mathbf{x}-\mathbf{y})^2}{|C|\cdot(|C|-1)}}$$

Similarity Measure: A similarity measure SIMILAR (Di, Dj) can be used to represent the similarity between documents. Typically similarity generates values of 0 for documents exhibiting no agreement among the assigned terms, and 1 when perfect agreement is detected.

Threshold: The lowest possible input value of similarity required to join two objects in one cluster.

Cluster Seed: First object of a cluster is defined as the initiator of the cluster i.e. every incoming object‘s similarity is compared with the initiator. The initiator is called the cluster seed.

IV. TYPES OF CLUSTERS

A. Well-Separated Clusters

A cluster is a set of points such that any point in a cluster is closer to every other point in the cluster than to any point not in the cluster.

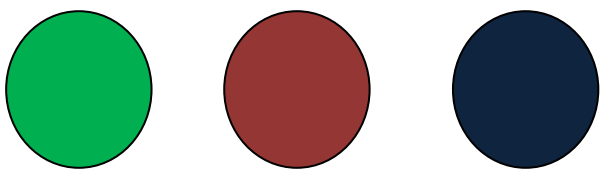

Fig 2: 3 Well-Separated Clusters

B. Center-based Clusters

A cluster is a set of objects such that the object in a cluster is closer to the "center" of a cluster, than to the center of any other cluster. The center of a cluster is often a "centroid", the average of all the points in the cluster or a "medoid", the most representative point of a cluster.

Fig 3: 2 Center-based Clusters

C. Contiguity-based Clusters

A Contiguity-based Cluster is also called nearest neighbour or Transitive clusters. A cluster is set of points such that a point in a cluster is closer to one or more other points in the cluster than to any point not in the cluster.

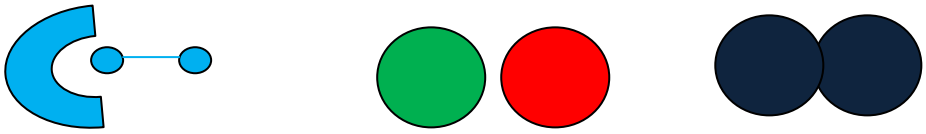

Fig 4: 5 Contiguity-based Clusters

D. Density-based Cluster

A cluster is a dense region of points, which is separated by low–density regions of high density regions, from other regions of high density. These clusters are used when noise and outliers are present.

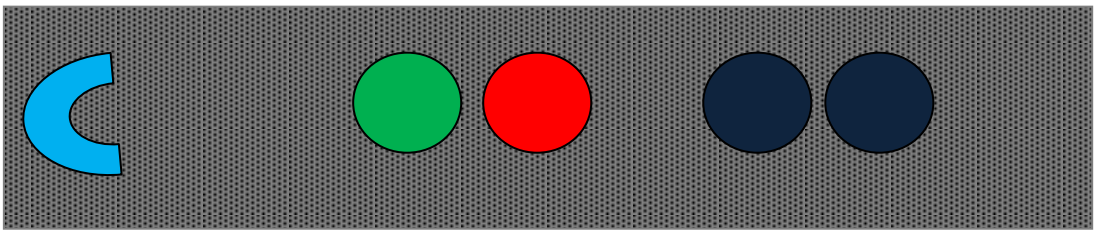

Fig 5: 5 Density-based Cluster

E. Conceptual Clusters

Conceptual Clusters are also called Shared Property Clusters. It finds clusters that share some king of common property or represent a particular concept.

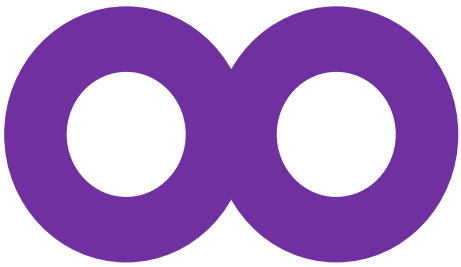

Fig 6: 2 Overlapping Circles

V. APPROACHES

Clustering is a division of data into groups of similar objects. Clustering algorithms can be divided into following main categories that are:-

A. K-Means Clustering Algorithm

K-Means Clustering method probably the most well known. K-Means Clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. The algorithm is called k-means, where k is the number of clusters we want. The algorithm starts k initial seeds of clustering, one for each cluster. All the data objects are then compared with each seed by means of Euclidian Distance and assigned to the closest cluster seed. This procedure is repeated again and again until the algorithm stops when the changes in the cluster seed from one stage to next are close to zero or smaller than the predefined value.

Algorithmic steps for k-means clustering Let X = {x1,x2,x3,……..,xn} be the set of data points and V = {v1,v2,…….,vc} be the set of centers.

1) Randomly select *'c'* cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.

4) Recalculate the new cluster center using

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

Where, *'ci'* represents the number of data points in *ith* cluster.
5) Recalculate the distance between each data point and new obtained cluster centers.
6) If no data point was reassigned then stop, otherwise repeat from step 3.

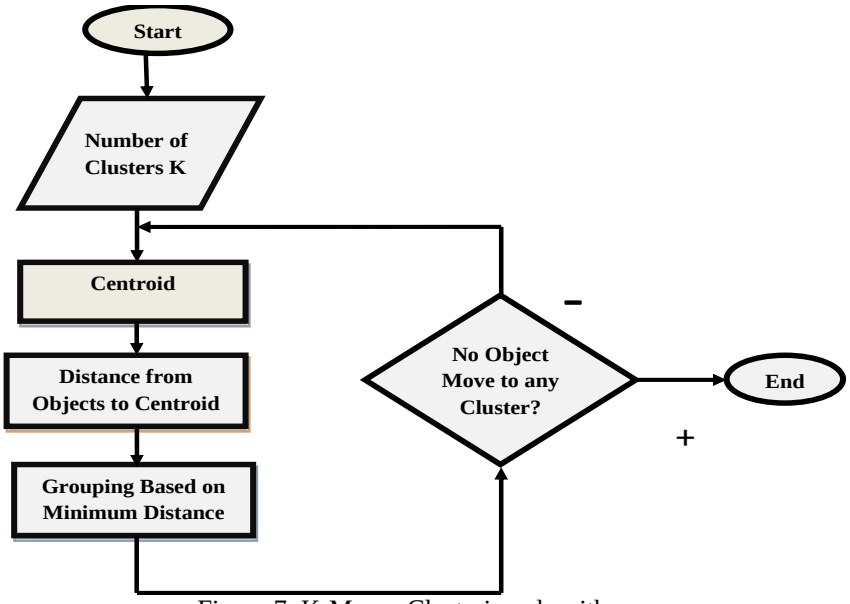

Figure 7: K-Means Clustering algorithm

Advantages**:** 1. K-Means algorithm is simple and understandable.
2. Data objects are automatically assigned to clusters.
3. Relationship to Gussian mixture model.
4. K-Means may produce tighter clusters.
5. It gives best results when data set are distinct or well separated from each other.

Disadvantages: 1. Need to specify K, the number of clusters in advance.
2. Often terminates at local optimum.
3. Unable to handle noisy data and outliers.
4. Not suitable to discover clusters with non- convex shapes.
5. Applicable only when mean is defined.

B. Kohonen Self Organizing Map (SOM)

The SOM is a new, valuable software tool for the visualization of high-dimensional data. It converts complex, nonlinear statistical relationships between high-dimensional data items into simple geometric relationships on a low-dimensional display.

Kohonen SOM is an artificial neural network that learns representations of data via an unsupervised learning algorithm. That is while many other ANN learns to model target data; the SOM learns the patterns within the data itself. It consists of two layers of artificial neurons: the input layer, which accepts the external input signals, and the output layer (also called the output map), which is usually arranged in a two-dimensional structure. Every input neuron is connected to every output neuron, and each connection has a weighting value attached to it. When an input vector is presented to the SOM, the Euclidean distance between the input vector and the incoming weight vector of each output map neuron is calculated. The output neuron with the smallest distance is declared the winner (This is also known as the Best Matching Unit (BMU)).

$$distFromBMU^2 = (bmuI - nodeI)^2 + (bmuJ - nodeJ)^2$$

SOM learning is an iterative process, during which training examples are propagated through the network, and connection weights modified according to equation. The SOM algorithm is essentially a clustering algorithm that will assign training examples to neurons, where each neuron is equivalent to the centre of one cluster. The winning neuron for each vector thus determines which cluster the vector belongs to.

The stages of the SOM algorithm can be summarised as follows:
1. *Initialization* – Choose random values for the initial weight vectors w*j*.
2. *Sampling* – Draw a sample training input vector x from the input space.
3. *Matching* – Find the winning neuron *I*(x) with weight vector closest to input vector.
4. *Updating* – Apply the weight update equation $\Delta w_{ji} = \eta(t)\ T_{j,I(\mathbf{x})}(t)\ (x_i - w_{ji})$.

5. *Continuation* – keep returning to step 2 until the feature map stops changing.

Advantages:- 1. Better Visualization and interpretation of nodes closer in 2D space will have corresponding cluster also closer.
2. Easier to regroup similar clusters.
3. SOM is Algorithm that projects high-dimensional data onto a two-dimensional map.
4. The projection preserves the topology of the data so that related data items will be mapped to nearby locations on the map.
5. SOM has many applications in pattern recognition, speech analysis, industrial and medical diagnostics, data mining.
Disadvantages:- 1. SOM can be viewed as a restricted form of k-Means (K-Means restricted on 2D geometry).
2. SOM results in suboptimal in some sense.
3. SOM is heuristic algorithm.
4. Large quantity of good quality representative training data required.
5. No generally accepted measure of 'quality' of a SOM.

C. Hierarchical Clustering Algorithm (HCA)

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:

i. 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. The single cluster becomes the hierarchy's root. For the merging step, it finds the two clusters that are closest to each other, and combines the two to form one cluster.

Steps: 1. Place each object in its own cluster (a singleton).
2. Merge in each step the two most similar clusters until there is only one cluster left or the termination condition is satisfied.

The distances between each pair of clusters are computed to choose two clusters that have more chance to merge. There are several methods to calculate the distances between the clusters Ci and Cj.

Table 1: Linkage Method

Single Linkage	$d_{12} = \min_{ij} d(X_i, Y_i)$	The distance between the closest members of the two clusters.
Complete Linkage	$d_{12} = \max_{ij} d(X_i, Y_i)$	The distance between the farthest apart members.
Average Linkage	$d_{12} = \frac{1}{kl} \sum_{i=1}^{k} \sum_{j=1}^{l} d(X_i, Y_j)$	This method involves looking at the distances between all pairs and averages all of these distance.

ii. 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.

Steps: 1. Start with one big cluster containing all the objects.
2. Divide the most distinctive cluster into smaller clusters and proceed until there are *n* clusters or the termination condition is satisfied.

Advantages: 1. Does not require to specify the number of clusters in advance.
2. It can produce ordering of objects which may be more informative for data display.
3. Easy to implement and gives best results in some cases.

Disadvantages: 1. Algorithm can never undo what is done previously.

2. Use of different distance metrics for measuring distance between clusters may generate different results.
3. Sensitive to noise and outliers.
4. Sometimes it is difficult to identify the number of clusters by Dendrogram.

VI. COMPARATIVE ANALYSIS

A detailed comparative study of different clustering algorithms proposed under the different methods of considering the different aspects of clustering is given in Table 2.

Clustering Algorithm	Proposed By	Year	Input Parameters	Computation Time	Memory Requirement	Capability of tackling High Dimensional Data
K-Means	Steinhaus	1955	No of Clusters	O(k t m n)	O(mn + kn)	No
	Lloyd	1957				
	Ball & Hill	1965				
	Mcqueen	1967				
SOM	C Vonder Malsburg	1970's	Maximum number of Neighbours	Not specified	Not Specified	Yes
	Proff. Teuvo Kohonen	1981-82				
HCA	Cormack	1971	No need to specify number of Clusters	O(m n^2 log(n))	O(mn + n^2)	No
	Anderberg	1973				
	Sneath & Sokal	1973				
	Hartigan	1975				
	Everitt	1980				

VII. CONCLUSION

Clustering is a process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters. The goal of data clustering is to discover the *natural* grouping(s) of a set of patterns, points, or objects. In this paper, an attempt has been made to give the basic concepts of clustering by firstly providing the definition of different clustering algorithms along with their advantages and disadvantages and some basic terms used in clustering. At last the complexity in terms of time and space has been discussed along with the dimensionality.

ACKNOWLEDGEMENT

I would like to thank the Department of Computer Science & Engineering of Amritsar College of Engineering & Technology, Manawala, GT Road Amritsar, Punjab, India.

REFERENCES

[1] Anonymous, A. Introduction to Data Mining and Knowledge Discovery, 3rd Edition. Two Crows Corporation, 2005.
[2] H. Frigui, Advances in Fuzzy Clustering and Feature Discrimination with Applications. John Wiley and Sons, ch. Simultaneous Clustering and Feature Discrimination with Applications, pp. 285–312 , 2007.
[3] W. Bo and R. Nevatia, Cluster boosted tree classifier for multi-view, multi-pose object detection, in Proc. ICCV, October 2007.
[4] S. Khan, G. Situ, K. Decker, and C. Schmidt, GoFigure: Automated Gene Ontology annotation, Bioinf., vol. 19, no. 18, pp. 2484–2485, 2003.
[5] The UniProt Consotium, The universal protein resource (UniProt), Nucleic Acids Res., vol. 35, pp. D193–D197, 2007.
[6] S. Gunnemann, H. Kremer, D. Lenhard, and T. Seidl, Subspace clustering for indexing high dimensional data: a main memory index based on local reductions and individual multi representations," in Proc. Int. Conf. Extending Database Technology, Uppsala, Sweden, pp. 237–248.2011.
[7] Prof. Neha Soni, Dr. Amit Ganatra, Comparative Study of Various Clustering Algorithms, International Journal of Advance Research, Volume 2, Number 4, Issue 6, December 2012.
[8] Bharat Chaudhari, Manan Parikh, A Comparative Study of clustering algorithms Using weka tools , International Journal of Application or Innovation in Engineering & Management, ISSN 2319 - 4847 , Volume 1, Issue 2, October 2012.
[9] Mu-Chun Su and Chien-Hsing Chou, ―A Modified Version of the K-Means Algorithm with a Distance Based on Cluster Symmetry, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 6, pp.674-680, Jun. 2001.

[10] Tapas Kanungo, David M. Mount,Nathan S. Netanyahu, Christine D.Piatko, Ruth Silverman, and Angela Y. Wu, ―An Efficient k-Means Clustering Algorithm: Analysis and Implementation, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 881-891, Jul 2002.

[11] Malay K. Pakhira, ‖A Modified k-means Algorithm to Avoid Empty Clusters,‖ International Journal of Recent Trends in Engineering, vol. 1, no. 1, pp.220-226, May 2009.

[12] Narendra Sharma, Aman Bajpai, Mr.Ratnesh Litoriya, Comparison the various clustering algorithms of weka tools, International Journal of Emerging Technology and Advanced Engineering, vol. 2, pp.73-80, May 2012.

[13] Fabrizio Angiulli, Stefano Basta, Stefano Lodi, and Claudio Sartori, Distributed Strategies for Mining Outliers in Large Data Sets, IEEE 2013.

[14] Ariel E. Baya and Pablo M. Granitto, How Many Clusters: A Validation Index for Arbitrary-Shaped Clusters, IEEE/ACM 2013.

[15] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding, Data mining with big data, IEEE 2013.

