A Survey: Clustering Algorithms in Data Mining

Sonamdeep Kaur
M.Tech (C.S.E) Amity University, Haryana

Sarika Chaudhary
Assistant Professor Amity University, Haryana

Neha Bishnoi
Assistant Professor Amity University, Haryana

ABSTRACT
In data mining Clustering is a technique that's aims to single out the data elements into different clusters based on useful features. In this technique data elements that are *similar* to one another are placed within the same cluster and those which are *dissimilar* are placed in different clusters. Many algorithms have been proposed in the literature but the most active research algorithms are unsupervised clustering methods of data mining:Partitioning and Hierarchical Methods for clustering. The choice of a particular clustering method depends on many factors or themes. The key idea of this paper is categorizing the methods on the bases of different themes so that it helps in choosing algorithms for any further improvement and optimization.

General Terms
Data Mining, Clustering, Pattern recognition.

Keywords
Partitional clustering, Hierarchical clustering,Statistical methods,*k*-Means,BIRCH, CURE.

1. INTRODUCTION
Data mining is the technique of extracting useful information or knowledge from a given data which can be small or large, nominal or categorical, temporal or spatial. Clustering is a Statistical data analysis technique which groups together similar data to recognise useful patterns in the data. It is an unsupervised learning mode which finds itsapplication in plant and animal taxonomies derivations, gene expression analysis, outlier detection, verification of precipitation forecasts [1], detecting anomalies in credit card usage. It finds its importance as it provide scalability and high dimensionality solutions. Clustering can be of two types (1) inclusive clustering dividing a dataset into clusters such that one datapoint belongs to more than one cluster and (2) exclusive clustering i.e dividing the dataset into different clusters such that one datapoint belongs to one cluster.This belongingness of datapoint is calculated with the help of proximity matrix. A number of distance functions are used to calculate the proximity between datapoints.

In this paper we examined various partioning and hierarchical methods for clustering. The algorithms compared are *k*-Means, *k*-Medoids, CLARANS, BIRCH, CURE, CHAMELEON, and ROCK.So far number of unsupervised clustering methods has been introduced for both types of clustering. Algorithms can be categorized on different aspects and the aspects according to which we have compared are scalability, type of variable/attribute, physical shape and capacity to handle outliers.

2. PARTITIONING CLUSTERING TECHNIQUE
The partitional clustering methods generate *k* partitions of the given dataset where each partition represents a cluster. It uses iterative relocation technique that attempts to improve the grouping by moving datapoints from one cluster to another. Techniques using partitioning method for clustering generally prompt clusters of globular shape [2].

2.1*k*-Means
k-Means[7] clustering algorithm is the most widely used algorithm of all the clustering algorithms. The *k* in the K-Means algorithms refers to the fact that the algorithm is going to look for *k* different clusters which means when applied on a dataset the algorithm is going to break the dataset into *k* clusters. The value of *k* has to be specified before clustering the dataset. Firstly we start with *k* datapoints as initial cluster centres and then these datapoints are reassigned to clusters with which the datapoint is more similar. It then calculates the new mean of the cluster. This procedure iterates until no further change is possible. The merging of these reassigned clusters is done by proximity matrix and finally the output of this method is *k* clusters.

2.2 *k*-Medoids
k-Medoids[3] is an improvement to *k*-Means as it is sensitive to outliers due to its centroid approach to represent a cluster which means that large datapoint value distorts the distribution of data. So, instead of taking mean of the datapoints as cluster representatives we consider actual datapoints to represent the clusters and then use medoidusing the absolute-error criterion. Therefore this approach is much robust in the presence of noise. The most common realisation of *k*-medoids is PAM (Partitioning AroundMedoids)

2.3 CLARANZ
It (Clustering Large Applications Based on Randomized Search) is a clustering process which draws sample of neighbours dynamically by finding out the spatial clusters present in the data. It takes each node in the graph as a possible solution. It dynamically draws a random sample at every new search. It makes the most *"natural"*clusters with the help of *silhouette coefficient* which tell the amount of belongingness of a datapoint to an underlying cluster. It outperforms PAM (also a Partitioning clustering method) in terms of run time and cluster quality.

3. HIERARCHICAL CLUSTERING TECHNIQUE

This clustering technique is, as its name suggests, partitions the clusters into levels such that they can be represented in top-down or bottom-up manner. The Hierarchical clustering algorithms build a cluster hierarchy that is commonly displayed as a tree diagram called a dendrogram. A main division of hierarchical algorithms can be done as: Agglomerative Nesting (AGNES), starting with each datapoint being considered a different cluster, then we iteratively merge most appropriate elements one by one until we get a single cluster and Divisive Analysis (DIANA), assign all the datapoints to a single cluster and further partition the cluster into two so that dissimilarity is more. A major disadvantage of this method is its rigidness of being unchangeable. Once a merge or a split is done it cannot be undone in future.We have discussed only the Agglomerative clustering algorithms in this paper.

3.1BIRCH

BIRCH (Balanced Iterative Reducing and Clustering Using Hierarchies) is a method which works best for large data. It is highly efficient in handling the outliers. BIRCH methodology is basically divided in four phases. It runs on the consideration that not all datapoints are equally important. It uses a (Clustering Feature) CF-Tree in memory in its first phase. Then in the second it condenses the CF-Tree. Third phase is Global Clustering phase where we perform simple traditional clustering of CF and not datapoints. Last phase is optional to get good to better clusters by redistributing datapoints to its closest seed and forming new clusters.BIRCH is reasonably fast, but has twoserious drawbacks: data order sensitivity and inability to deal with non-spherical clusters of varying size [8].

3.2 CURE

First step of CURE(Clustering using Representatives) is to draw a random sample and further partition the sample, so as to have partially clustered partitions. The datapoints taken for random sample can be well-scattered. These selected datapoints decide the shape of the sample. Therefore, this algorithm forms clusters of non-spherical shape. After, these partial partitions are shrunk to remove outliers. In this Shrink factor between 0.2-0.7 is considered to find right clusters. It follows a central outlook between the centroid-based and all point extremes. This algorithm basically addresses to the major problems of hierarchical clustering i.e. outliers and clusters of only spherical shape.It ignores the interconnectivity amongst the clusters and gives preference to distance between the representative points of two clusters[6]. Also, it fails to consider prominent features of a particular cluster thus affecting the clusters merging decisions. CURE only works for metric data.

3.3 ROCK

ROCK (Robust Clustering using Links) introduced the concept of link and neighbour. Link incorporates comprehensive information of other similar enough neighbours so that not only two points are considered every time merging or splitting clusters. Bigger is the link, higher the probability of points being in same cluster. Traditional algorithms used functions for Boolean and categorical attributes but here concept of links (common neighbours) is introduced. ROCK has demonstrated its power by being successfully used for real datasets [7].

3.4 CHAMELEON

It measures the similarity between pair of clusters using a dynamic modelling approach. It has two phases defined as finding initial sub-clusters and merging these sub-clusters using a dynamic framework. Formation of initial sub-clustersis done by using graph partitioning method to easily get high-quality partitioning for wide range of unstructured graphs [4].

4. CONCLUSION

Table 1 shows the algorithms being compared on the basis of different aspects. Therefore, we conclude that from the algorithms discussed that:

- *k*-Means is good for small databases but is very much sensitive to noise. But, it is more efficient than k-Medoids[3] but *k*-Medoids is not much sensitive to noise.
- CLARANS have better performance with very small databases.
- BIRCH gives good performance for both large and small databases but gives spherical clusters [6].
- CURE gives best performance for non-spherical clusters formed in large databases [5].
- ROCK is the most efficient method for large databases having Boolean and categorical data variables. It works worst in metric space datapoints.
- CHAMELEON performs best for any kind of data and forms non-spherical clusters.

As we see in the following table CHAMELEON and CURE are considerably much better than the other algorithms but the major limitation of CURE is its inability to consider special characteristics of a single cluster which makes wrong merging decisions. Further optimization of CURE is possible taking into account its restriction to achieve more commending results.

Table 1.Comparision Methodology

Algorithm	**Capacity to handle outliers**	**Types of variable**	**Physical shape of clusters**	**Type of data set**
k-Means	Bad	Numeric	Spherical clusters	Small And medium datasets
k-Medoids	Good	Numeric	Spherical clusters	Small And medium datasets

CLARANS	Good	Numeric	Spherical clusters	Large datasets
BIRCH	Appreciable	Numeric	Spherical clusters	Large And Small datasets
CURE	Very Good	Numeric	Non-Spherical clusters	Large datasets
ROCK	Good	Categorical	Non-Spherical clusters	Large datasets
CHAMEL - EON	Very Good	All types of data	Non-Spherical clusters	Large datasets

5. ACKNOWLEDGEMENT

I would like to thank Ms SarikaChaudhary, Ms NehaBishnoi and Computer Science and Engineering Department of Amity University, Haryana for their continuous support and encouragement.

6. REFERENCES

[1] Stefano Serafin, AlessionBerto, Dino Zardi. (2005) Application of Cluster Analysis Technique to the Verification of Quantitative Precipitation Forecasts. University of Trento.

[2] Shalini.S.Singh, N.C.Chauhan (2011).K-Mean v/s K-Medoids: A comparative study. National Conference on Recent Trends in Engineering & Technology.

[3] Dr. T. Velemurugan. (2012) Efficiency of k-Means and K-Medoids Algorithms for Clustering Arbitrary Data Points. ISSN: 2229-6093.

[4] George Karypis, Eui-Hong (Sam) Han, Vipin Kumar.(1999).*CHAMELEON*:AHierarchical *Clustering*Algorithm.Using Dynamic Modelling.

[5] Guha, Sudipto; Rastogi, Rajeev; Shim, Kyuseok (2001). "CURE: An Efficient Clustering Algorithm for Large Databases".

[6] Pooja Gupta,, Monika Jena. (2013).Comparative study of different clustering algorithms for association rule mining.

[7] Jiawei Han and MichelineKamber. (2006). Data Mining: Concepts and Techniques.

[8] Malwindersingh,Meenakshibansal. (2014).Survey On Clustering And Optimization Techniques ToDevelop Hybrid Clustering Technique International Journal Of Computer Engineering AndApplications.