Comparative Study on Hierarchical and Partitioning Data Mining Methods

Agrawal Rammohan Ashok[1], Chaudhari Rahul Prabhakar[2], Pathak Aniket Dyaneshwar[3]

[1,2,3] Assistant Professor, Department of Computer Science & Engineering
Shri Sant Gadge Baba College of Engineering & Technology, ZTC, Bhusawal – 425203, Maharashtra, India
agrawalrammohan@gmail.com
rahul.chaudhari7@gmail.com
aniketpathak89@gmail.com

Abstract: ***Data mining is the exploration and analysis of large quantities of data in order to discover valid, novel, potentially useful and ultimately understandable patterns in data. The goal of this paper is to study hierarchical along with partitioning method and its recent issues and present a comparative study on the above mentioned clustering techniques that are related to data mining. This paper presents a tutorial overview of the main clustering methods used in data mining. The goal is to provide a self contained review of the concepts and the mathematics underlying clustering techniques along with some experimental results. Paper begins by providing some measures and criteria that are used for determining whether two objects are similar or dissimilar. Further on the paper explores the study of clustering methods with some experimental results which is a study based experimental conclusion. Finally we conclude our paper with the problems we faced during dealing with the cluster classification and in getting the output.***

Keywords: Data Mining (DM) and clustering.

1. Introduction

The concept of clustering in the field of data mining has been around for a long time. Clustering comes under the term of *unsupervised learning* problem. It deals with finding a *structure* in a collection of unlabeled data. Clustering may also be defined as the process of organizing objects into groups whose members are similar in some characteristics". A *cluster* is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. It has several applications, particularly in the context of information retrieval and in organizing web resources. *The ultimate aim of the clustering is to provide a grouping of similar records*. Clustering is often confused with classification, but there is some difference between the two. In classification the objects are assigned to pre defined classes, whereas in clustering the classes are formed. The term "class" is in fact frequently used as synonym to the term "cluster". In database management, data clustering is a technique in which, the information that is logically similar is physically stored together. In order to increase the efficiency of search and the retrieval in database management, the number of disk accesses is to be minimized. In clustering, since the objects of similar properties are placed in one class of objects, a single access to the disk can retrieve the entire class. If the clustering takes place in some abstract algorithmic space, we may group a population into subsets with similar characteristic, and then reduce the problem space by acting on only a representative from each subset.

Clustering techniques are used for combining observed examples into clusters (groups) which satisfy two main criteria:

A. Each group/cluster is homogeneous; examples that belong to same group are similar to each other.

B. Each group/cluster should be different from other clusters. Depending on the clustering technique, clusters can be expressed in different ways:

- ✓ Identified clusters may be exclusive.
- ✓ They may be overlapping.
- ✓ They may be probabilistic.

2. Literature Survey

In [2] the authors have provided a comprehensive review of different clustering techniques in data mining. We now consider their implementation of hierarchical clustering with partitioning and try to extend and implement the algorithm on the k-means clustering that will yield results on some of the music profiles which is a study based analysis.

3. Clusters Analysis

Cluster analysis means finding the groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups. Here we randomly choose the debt – vs-income analysis of various employees in an organization and group them into various clusters. The following Figure 1, summaries this:

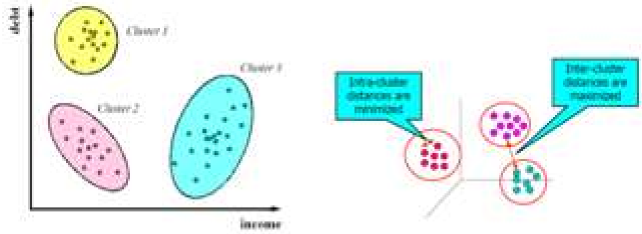

debt
Cluster 1
Cluster 3
Cluster 2
income
Intra-cluster distances are minimized
Inter-cluster distances are maximized

Figure 1: Analyzing Clusters

2.1. What is not Cluster Analysis?

People often misunderstand the following with cluster analysis:

- ➢ Supervised classification or learning
- ➢ Simple segmentation rules
- ➢ Results of a query generated
- ➢ Graph partitioning

2.2. Notion of a Cluster can be Ambiguous

Figure 2 solves our query of ambiguity of a cluster. It helps in grouping cluster into various groups of six, two and four clusters respectively.

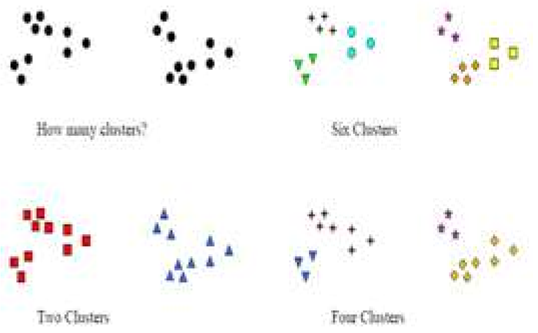


Figure 2: Cluster Notation

2.3. Well-separated clusters

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.

2.4. Center-based clusters

Cluster is a set of objects such that an 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.

2.5. Contiguous clusters

A cluster is a 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.

2.6. Density-based clusters

A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density. Used when the clusters are irregular or intertwined, and when noise and outliers are present.

4. Classification of Clustering Algorithm

Categorization of clustering algorithms is neither straightforward, nor canonical. Clusters can be classified into many; some of the methods are listed below. In reality, groups below overlap. Some of the methods in clustering are:

- ✓ *Hierarchical Methods*
- ✓ *Partitioning Methods*

Some of the important issues should be taken into considerations. The properties of the clustering algorithm which we take in consideration in data mining includes

- ✓ Type of attributes algorithm can handle.
- ✓ Scalability to large datasets.
- ✓ Ability to find clusters of irregular shape.
- ✓ Handling outliers.
- ✓ Data order dependency.
- ✓ Interpretability of results.

5. Hierarchical Methods

Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters, or in other words a tree of clusters, also known as a dendrogram. Every cluster node contains child clusters, sibling clusters partition the points covered by their common parent. The basics of hierarchical clustering includes Lance-Williams formula, idea of conceptual clustering, now classic algorithms SLINK, COBWEB, as well as newer algorithms CURE and CHAMELEON. The hierarchical algorithms build clusters gradually. Hierarchical Clustering is subdivided into *agglomerative* methods, which proceed by series of fusions of the n objects into groups, and *divisive* methods, which separate n objects successively into finer groupings. Agglomerative techniques are more commonly used, and this is the method implemented in XLMiner. Hierarchical clustering may be represented by a dendrogram as shown in figure 3, and it illustrates the fusions or divisions made at each successive stage of analysis.

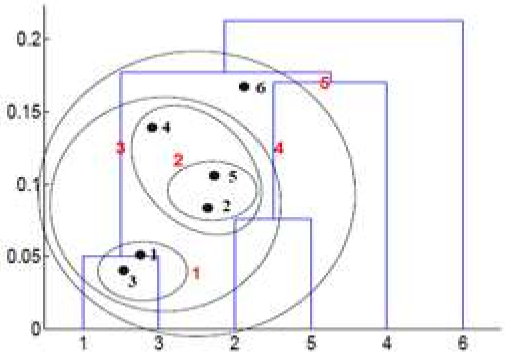


Figure 3: Representation of a nested cluster

4.1.1. Agglomerative

This is a *"bottom up"* approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. The algorithm forms clusters in a bottom-up manner, as follows:

1. Initially, put each article in its own cluster.
2. Among all current clusters, pick the two clusters with the smallest distance.
3. Replace these two clusters with a new cluster, formed by merging the two original ones.

4. Repeat the above two steps until there is only one remaining cluster in the pool.
5. Thus, the agglomerative clustering algorithm will result in a binary cluster tree with single article clusters as its leaf nodes and a root node containing all the articles.

4.1.2. Divisive Algorithm

This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

1. Put all objects in one cluster
2. Repeat until all clusters are singletons
   a. Choose a cluster to split
   b. Replace the chosen cluster with sub-cluster.

Advantages of hierarchical clustering

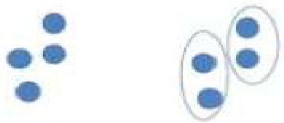
- ✓ Embedded flexibility regarding the level of granularity
- ✓ Ease of handling of any forms of similarity or distance
- ✓ Consequently, applicability to any attributes types.

Disadvantages of hierarchical clustering

- ✓ Vagueness of termination criteria
- ✓ The fact that most hierarchical algorithms do not revisit once constructed (intermediate) clusters with the purpose of their improvement.

4.2. Partitioning Methods

The partitioning methods generally result in a set of M clusters, each object belonging to one cluster. Each cluster may be represented by a centroid or a cluster representative; this is some sort of summary description of all the objects contained in a cluster. The precise form of this description will depend on the type of the object which is being clustered. In case where real-valued data is available, the arithmetic mean of the attribute vectors for all objects within a cluster provides an appropriate representative; If the number of the clusters is large, the centroids can be further clustered to produces hierarchy within a dataset.



Original Points Partitioned Points

Figure 4: Partition Clustered points in a domain

4.2.1 k-means Methods

In k-means case a cluster is represented by its centroids, which is a mean (usually weighted average) of points within a cluster. This works conveniently only with numerical attributes and can be negatively affected by a single outlier. The k-means algorithm [Hartigan 1975; Hartigan & Wong 1979] is by far the most popular clustering tool used in scientific and industrial applications. The name comes from representing each of k clusters C by the mean (or weighted average) c of its points, the so-called centroid. The sum of discrepancies between a point and its centroids expressed through appropriate distance is used as the objective function. Each point is assigned to the cluster with the closest centroid. Number of clusters k must be specified. The basic algorithm is as follows:

Input: S (instance set), k (total number of cluster)
Output: Clusters

1. Start
2. Select k points as initial centroids.
3. Repeat step 2 till desired criteria not get.
4. Form k clusters by assigning each point to its closest centroids.
5. Re-compute the centroids of each cluster until centroids do not change.

The k-means algorithm may be viewed as a gradient decent procedure, which begins with an initial set of K cluster-centers and iteratively updates it so as to decrease the error function.

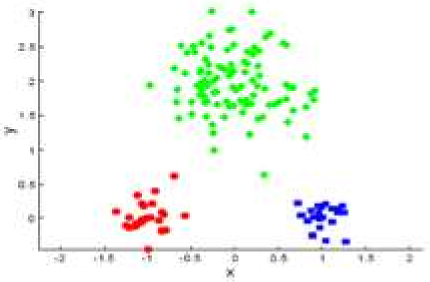


Figure 5: Original points in a domain

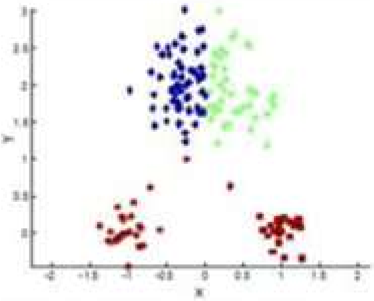


Figure 6: Sub –optimal Clustering on Original points

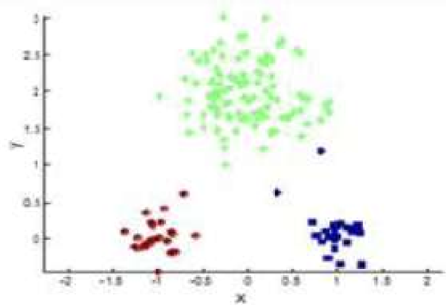


Figure 7: Optimal Clustering on Original points.

4.2.2. Implementing k-means clustering on some musical profiles using DM – A study based analysis

The k-means clustering algorithm is a straightforward technique that attempts to find a classification of the vectors, putting them in clusters of users that are similar in their musical preferences. Their definition to ending up in the same cluster is that they are all closest to their cluster's centre point (with respect to Euclidian distance).

Algorithm: k-means algorithm

// This algorithm will require some set of input items, x, in Euclidean space; desired number of clusters, k.

```
1.  Start
2.  Repeat till $1 \leq i \leq k$ do
        a.  kmeans[i] $\longleftarrow$ random item from data
        b.  centroid[i] $\longleftarrow$ 0
        c.  count[i] $\longleftarrow$ 0
3.  repeat
        a.  for all x $\in$ 2 items do
        b.  mindist $\longleftarrow$ 1
4.  for $1 \leq i \leq k$ do
5.  if$\|x - \text{kmeans}[i]\|2 < \|x - \text{kmeans}[\text{mindist}]\|2$ then
6.   mindist $\longleftarrow$ i
7.  cluster[x] $\longleftarrow$ mindist
8.  centroid[mindist] $\longleftarrow$ centroid[mindist] + x
9.  count[mindist] $\longleftarrow$ count[mindist] + 1
10. for $1 \leq i \leq k$ do
11. kmeans[i] $\longleftarrow$ centroid[i]/count[i]
12. centroid[i] $\longleftarrow$ 0, count[i] $\longleftarrow$ 0;
13. until no items reclassified or repetition count
    exceeded
14. each x $\in$ 2 items is now classified by cluster[x]
15. Exit.
```

The k-means algorithm idea is based around clustering items using centroids. These are points in the metric space that define the clusters. Each centroid defines a single cluster, and each point from the data is associated with the cluster defined by its closest centroid. The algorithm proceeds in rounds: in each round, every input point is inspected and compared to the k centroid points to find which is closest. At the end of every round, we compute a new set of centroids based on the points in each cluster. For each cluster, we compute the centroid of that cluster, as the "center of mass" of the points. The center of mass can be found efficiently by finding the mean value of each co-ordinate. This leads to an efficient algorithm to compute the new centroids with a single scan of the data: for each of the k clusters, compute the sum of each co-ordinate value of all points that are associated with that cluster, and the count of the number of points in the cluster. The new centroids can then be easily computed after all points have been allocated to clusters.

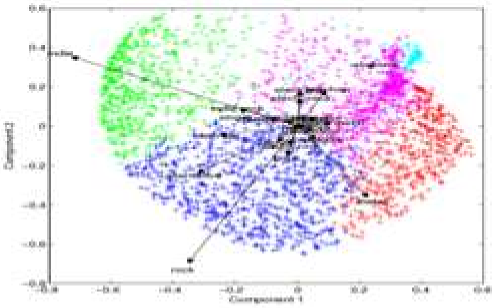


Figure 8: Cluster Classification using Colors

The two other clusters are not separated clearly when depicted with respect to two principal components, but the third and fourth components show that the remaining clusters separate users that listen to "hip-hop", and those that don't. "Electronic" is the adjective used for music that is produced electronically, where "Electronica" is the actual genre. For each of these clusters, we can build the common tag cloud, describing the average of musical genres that are enjoyed by the users in the clusters with little or no overlap with respect to the musical genres.

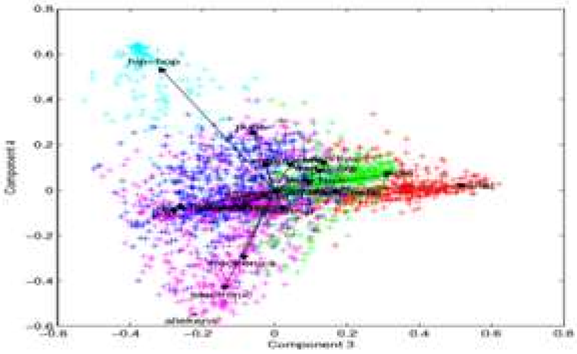


Figure 9: Cluster dense on a region

This can also be observed by the relative mutual exclusivity of "rock" and "indie" in their tag clouds. The "electronic" cluster packs a lot of different genres, and is the most versatile cluster. The above figure attempts to separate this "electronic" cluster into smaller ones by choosing a higher number of clusters failed, this cluster remained intact. The sub clusters of this big cluster are "pop", "Japanese", "ambient", "electronic", "industrial" and "punk." Sub populations of the other main clusters also define more specific target audiences each of them having their own needs of music sound profiles.

Limitations of k-means

k-means has problems when clusters are of differing sizes, densities, non-globular shapes and when the data contains outliers. Our goal is to overcome these limitations. We here present another method to solve this.

Bisecting k Means Method

This is an extension of k-means method. The basic concept is as follows that to obtain k clusters split the set of all points into two clusters select one of them and split and repeat this process until the k clusters have been produced.

Problems with Clustering:

Some of the clustering problems are:

- ✓ Results may be random (in many ways).
- ✓ Clustering techniques do not address all requirements adequately.
- ✓ Time complexity is directly proportional to experimental data.

6. Conclusion & Future Enhancement

In Clustering, the ability to discover highly correlated regions of objects when their number becomes very large is highly desirable. We conclude that as data sets grow their properties and data interrelationships also changes. Also when we try to study the experiment of clustering on colors we found that this cluster classification depends purely on various sound profiles and also the mood of the users listening to this music. For example: if many people like “rock” sound profile then the cluster will be found dense in the “rock” profile region. Further on we are now trying to eliminate the problems that we faced in clustering.

References

[1] M. Marin, A. van, Deursen, and L. Moonen. Identifying Aspects Using Fan-in Analysis. In Proceedings of the 11th Working Conference on Reverse Engineering (WCRE2004), pages 132-141. IEEE Computer Society, 2004.

[2] Pradeep Rai, Shubha Singh, A Survey of Clustering Techniques, International Journal of Computer Applications (0975 – 8887), Volume 7– No.12, October 2010.

[3] Orlando Alejo Mendez Morales. Aspect Mining Using Clone Detection. Master's thesis, Delft University of Technology, The Netherlands, August 2004.

[4] D. Shepherd and L. Pollock. Interfaces, Aspects, and Views. In Proceedings of Linking AspectTechnology and

[5] Evolution Workshop(LATE 2005), March 2005.

[6] P. Tonella and M. Ceccato. Aspect Mining through the Formal Concept Analysis of Execution Traces. In Proceedings of the IEEE Eleventh Working Conference on Reverse Engineering (WCRE 2004), pages 112_121, November 2004.

[7] L. D. Baker and A. McCallum. Distributional clustering of words for text classification. In SIGIR ’98: Proceedings of the21st Annual International ACM SIGIR, pages 96–103. ACM, August1998.

[8] R. Bekkerman, R. El-Yaniv, Y. Winter, and N. Tishby. On feature distributional clustering for text categorization. In ACM SIGIR, pages 146–153, 2001.

[9] P. Berkhin and J. D. Becher. Learning simple relations: Theory and applications. In Proceedings of the The Second SIAM International Conference on Data Mining, pages 420–436, 2002.

[10] B. E. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In COLT, pages 144–152, 1992.

[11] P. S. Bradley and O. L. Mangasarian. k-plane clustering. Journal of Global Optimization, 16(1): 23–32, 2000.

[12] S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan. Using taxonomy, discriminants, and signatures for navigating in text databases. In Proceedings of the 23rd VLDB Conference, Athens, Greece, 1997.

[13] T. M. Cover and J. A. Thomas. Elements of Information Theory. John Wiley & Sons, New York, USA, 1991.

[14] Zahn, C. T., Graph-theoretical methods for detecting and describing gestalt clusters. IEEE trans. Computer. C-20 (Apr.), 68-86, 1971