Clustering of Customers from Massive Customer Transaction Data

Neethu CM[1], Anitha Abraham[2], Linda Sebastian [3]

[1]Post Graduation Student, College of Engineering Kidangoor

[2, 3]Assistant Professor, College of Engineering Kidangoor

Abstract*: In this internet era, more and more people use online shopping. Analysing massive customer transaction data about these online activities can be used to improve the business and to satisfy customer demands in a better way. In this research paper we try to study different methods employed to analyse the customer transaction data. In our study we have studied methods like K-Means clustering,PAM clustering,Agglomerative, Divisive and Density Based clustering methods. Based on our study we have identified that K-Means is the widely used clustering method*.

Keywords: Clustering, Partitional clustering, Hierarchical Clustering

1. Introduction

Clustering can be considered the most important unsupervised learning problem. It deals with finding a structure in a collection of unlabelled data [1] et al says the definition of clustering could be the process of organizing objects into groups whose members are similar in some way. A cluster is therefore a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters.[2] Clustering is most important technique used in datamining. The most commonly used algorithms in Clustering are Hierarchical and Partitioning. [3] Pramod Gupta says that applications of clustering in various field like In Biology, clustering has been used to find groups of genes that havesimilar functions. In Information Retrieval, clustering canbe used to group search results of a query into a small numberof clusters, each of which capturing a particular aspectof the query. In Geology,cluster analysis has been appliedto find patterns in the atmospheric pressure of polar regionsand areas of the ocean that have a significant impact on landclimate. In Medicine, cluster analysis can also be used to detectpatterns in the spatial or temporal distribution of diseaseslike cancer, autism, etc.

2. Scope of this Survey

The scope of this survey is to find out the best clustering technique which help us to analyse the customer's transaction behavior based on their purchases. Today most of the people depend the online shopping so to do this marketers analyse the customer's behaviors through their transaction record.

3. Literature Survey

After Reviewing the literature in the existing domain we have find out that there exists the following classification for solving the problem [6] Different starting points and criteria usually lead to different taxonomies of clustering algorithms [7],[8],[9],[10],[11] widely agreed frame is to classify clustering techniques as hierarchical clustering and partitional clustering, and density based clustering based on the properties of clusters generated.

3.1 Partitional Method

In partition method data objects are divided into non overlapping clusters so that every and each object is in one subset. [12] There are totally different [13] algorithms for partitioning like means, K-Medoids, PAM (Partitioning AroundMedoids), CLARA (Clustering Large Applications) and CLARANS.

3.1.1 K-Means

K-means is the most easy and wide used partitioning ways. This methods divide the information into k clusters ($C_{1,}C_{2.....}C_K$) with their centres and mean. The centre points of all clusters is calculated by taking the mean of all knowledge points belonging to the cluster. For that we've got used the subsequent equation to search out the mean of information points in [14],[15],[16],[17],[18],[19].

$$M = \frac{1}{N}\sum_{1}^{N} Xq$$

Where M is the mean of all points of cluster k and N is the number of points belonging to that clusters k.

K-Means Algorithm
Input: S (instance set), K (number of cluster)

Output: Clusters

```
1) Initialize K cluster centers.
2) whereas termination condition is not satisfied do
3) Assign instances to the nearest cluster center.
4) Update cluster centers based on the assignment.
5) end while
```

L.Rokach et al [19] says that the K-means algorithmmay 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.The short coming of the K Means algorithm is that it works well only with data points having isotropic clusters. In addition, the method is applicableonly when mean is defined. K-Means algorithm expects the number of clusters to be formed, is defined in advance. Noise and outliers affect the performance of K-Means algorithm.

3.1.2 PAM(Partition Around Medoids)

PAM is similar to K-means and find median instead ofdoing mean. PAM is to find the representative object for each cluster, this representative object called medoids which means that the most centrally located point in the cluster. The data points which is closest to the medoids is grouped into the cluster since the medoids is found. Consider O_j is data object and O_m is selected medoids [24] we can say that the O_j belongs to the cluster O_m. The quality of clustering is determined by average dissimilarity between an object and the medoids of its cluster. That is an absolute criterion function is used and is defined as,E $=\sum_{j=1}^{n} \sum p \in c_j \, |P - Q|$, where E is the sum of the absolute error for all objects in the data set; p is the point in the space representing given object in cluster C_j; and O_j is the representative object of C_j. The short coming of the PAM algorithm is that processing is more costly than K-Means. In addition, algorithm expects the number of clusters to be formed, is defined in advance. [21] RuiXu et al proposed algorithm as follows,

PAM Algorithm
Input, k: the number of clusters, D: a data set containing n objects
Output: Clusters

1) arbitrarily choose k objects in D because the initial representative objects or seeds;
2) repeat
3) assign each remaining object to the cluster with the the closest representative object;
4) randomly select a nonrepresentative object, O_{random};
5) compute the total cost,S,of swapping representative object,O_j, with O_{random}.
6) if S < 0 then swap $O_j$with $O_{random}$to form the new set of k representative objects;
7) until no change;

3.1.3 CLARA(Clustering Large Applications)

J.Han et al [15] proposed CLARA algorithm is used to handle large data set and relies on sampling. It draws a sampleof data set and applies PAM algorithm on it and find the medoids of the sample. The important factor is the point in the sample is drawn randomly and medoids of sample should be the medoids of entire data samples. CLARA drawn multiple samples and gives best clustering results. Quality of clustering is determined with average dissimilarity of all objects in the entire data set in [19] describes the algorithm as follows,

CLARA Algorithm

1. For i=1 to 5, repeat the following steps.
2. Draw a sample of 40+2k from the whole data set randomly,
and apply PAM to the data set for locating the
medoids.
3. For each object $O_j$within the entire set, realize the k-medoids is that the most just like O_j.
4. Compute the average dissimilarity of the clustering obtained within the previous step.If the values is a smaller amount the current minimum, and retain the k-medoidsfound in step2 because the best set of medoids obtainedto this point.
5. Return to step one to begin the next iteration

3.1.4 CLARANS (Clustering LARge Applications based on RANdomised Search)

Han et al [15] proposed CLARANS method for clustering polygonal objects. CLARANS is main memory clustering techniques and it uses K-medoids method for clustering. Because K-medoids is very robust and to avoid outliers. Resulting clusters are not depend on the order of the data set. It also handle very large data set efficiently.Unlike CLARA, CLARANS does not check neighbour of a node and it drawn a sample of neighbours in each step. Search by clarans gives high quality clustering than clara, and it requires the small number of searches. The algorithm contain two parameters maxneighbour and numlocal, the higher value of maxneighbourcloser is CLARANS to PAM and longer is each search of local minima. Application is that CLARANS is used to cluster spatial coordinates. since the parameter maxneighbour is set to high,it is very effectively same as the quality ofthe clustering produced by PAM. Same as for lower value of maxneighbour produces a lower clustering quality.

3.2 Hierarchical Method

Hierarchical method is one of the clustering method. It can be categorised into two, they are Agglomerative method and Divisive method.

3.2.1 Agglomerative Hierarchical Method

Cen Li et al [22] proposed an agglomerative method. Itstarts with many small clusters so it is called bottom up approach. Each and every iteration each smaller clusters combine to form large clusters. And finally we got a large cluster. [15]The hierarchical clustering methods could be further divided according to the manner that the similarity measure is calculated.

Single-link cluster(also known as the connectedness,the minimum technique or the nearest neighbour method)methods that consider the distance between two clusters to be equal to the shortest distance from anymember of onecluster to any member of the other cluster. If the data consist of similarities, the similarity between a pair of clusters is considered to be equal to the greatest similarity from any member of one cluster to any member of the other cluster.

Complete-link cluster (also called the diameter, the maximum method or the furthest neighbour method) – methods that consider the distance between two clusters to be equal to the longest distance from any member of one cluster to any member of the other cluster.

Average-link cluster(also called minimum variance method), methods that consider the distance between two clusters to be equal to the average distance from any member of one cluster to any member of the other cluster. Single link cluster have a defect known as chain effect", (ie) few points that form a bridge between 2 clusters that ends up in these clusters into one [23]et al proposed Balanced IterativeReducing and Clustering using Hierarchies (ie)BIRCH it is an agglomerate technique, it is used for big databases. This method clusters the incoming multi-dimensional metric data points incrementally and dynamically to provide quality clusters. Its I/O cost is linear in size with reference to data points. It provides good cluster since one scan of the database may be finished with BIRCH. It is the first algorithm in the field of data mining to handle noise. During data clustering, discovers the distribution pattern of the data set. Generally in this method there are two types of attributes are used to be clustered, metric and non-metric attribute. BIRCH offers parallelism since it used metric attributes. Mainly there are four phases in BIRCH as follows,

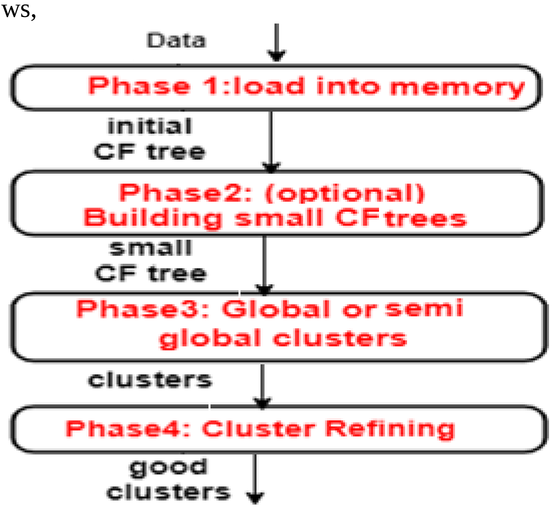

Figure 1: Overview of BIRCH

In Phase 1, all data points are scanned only once. And insert the points into trees. If it runs out of memory before it finishes scanning the data, it increases the threshold value, rebuilds a new, smaller CF tree, by re-inserting the leaf entriesof the old tree. Load data into memory using CF treeswhich uses the maximum of memory spaces. Data points are build in the form of CF tree. Condense points are treated as a single cluster and sparse are as a outliers. CF tree concept used here, CF tree is a tree with 2 parameters , branching factor B and threshold value T. Each non-leaf node contains at most B entries of the form[CF, child] where CF is a pointer to its child node and child is a sub clusters represented by its child. In Phase 2, it is an optional and it acts as a bridge between phase 1 and phase 3. Phase 3 uses the global or semi-global algorithm to cluster all leaf entries. Here different input size and perform well in terms of speed and quality. After phase 3 a set of clusters are formed. Phase 4 is also optional and refine the clusters later. The data was scanned only once and although the tree and outlier information may be scanned multiple times. we can extend the phase 4 if we desired by the user. It also provide discarding of outliers, a point which is so far from the seed considered outliers.

[23]Optionally, we can use R bytes of disk space for handling outliers which are leaf entries of low density. It is judged to be unimportant, with respect to the overall pattern.When we reconstruct the CF tree by re-inserting the old leaf entries, the size of the new tree is reduce in two ways. First, we increase the threshold value, thereby allowing each leaf entry to absorb more points. Second, we treat some leaf entries as potential outliers and write them out to disk. An old leaf entry is considered to be a potential outlier if it has far fewer data points than the average. Far fewer, is of another heuristics. Periodically, the space could run out, and the potential outliers are scanned to see if they'll be reabsorbed into this tree without causing the tree to grow in size. An increase in the threshold value or a change in the distribution due to the potential outlier not qualities as an outlier. The potential outliers left, when all data has been scanned, and the disk space must be scanned to verify if they are original outliers. If a potential outlier cannot be verified at this last chance and it consider only to be a real outlier and can be removed. There is a drawback is that entire cycle insufficient memory leads to rebuilding of the tree, insufficient disk spare leads to re-absorbing of outliers, etc could be repeated many times before the data itself is scanned. This effort will cause the cost of scanning the data in order to assess the of Phase 1 accurately.

Guha et al [4] developed another agglomerative hierarchical clustering algorithm, ROCK. Grouping data with attributes or distinct non-metric attributes. This method is a measurement of link used to reveal the relation between combine of objects and their common behaviour. Link captures the neighbourhood related data and it provides more robust solutions. This method also use random sampling and used to deal with large data sets. Rock basically uses the Jaccard coefficient as the distance measure between clusters

$$\text{Sim(A,B)} = \frac{|A \cap B|}{|A \cup B|}$$

similarity levels at most min {A+B}+1 values. Rock freefrom the chain impact phenomenon.

Algorithm: ROCK's hierarchical clustering algorithm is as follows, (Data Draw random sample Cluster with links Label data in disk). Receives the set S of n sampled points to be clustered (randomly drawn from original data set). The procedure begins with computing the number oflinks. A local heap is created in each cluster and maintain these during the execution of algorithm. Clusters in the heap are decreasing order of their goodness measure. Complexity of ROCK is O(maxfn2;nmmma;n2logng where ma and mm are the average and maximum number of neighboursof a points. It is given by Computation of links,(O(n2ma)ma) is the average number, Update of the local heaps with the new cluster w, O(n2logn). Space complexity is O(minn2;nmmma). Rock reduces the complexity by reducing the big set of data sets. In which the sample points used to build clusters, that not affects the quality of clusters. Goodness value is calculated by using the criterion function When the criterion function is high we tend to got the best clustering points. Then, we define the goodness measure g(Ci,Cj) for merging clusters Ci,Cjas follows.

$$g(C_i, C_j)=\text{link}(C_i, C_j)(n_i+n_j)^{1}+2\ f\ (\emptyset)\ -\ n_i^{1}+2\ f\ (\emptyset)\ -\ n_j^{1}+2\ f\ (\emptyset)$$

where, link[Ck,Cj]=cross links between clusters. The best pair of clusters to be merged at any given step when above goodness measure is maximum.

Issues in ROCK is labeling data on disk: Assigning the remaining data points in database to the clustersgenerated a set of points Li from each cluster for each remaining point. Compute Ni, the neighbours in Li , p is assigned to cluster I that Ni is maximum. scalability is the main advantage of this method (i.e) minimum execution time because of the combination of random sampling and labelling. Computation of links more efficient as Q(global heap) increases and qualityof clustering improves as random sample size increases. Agglomerative graph based clustering: The properties of graph theory is used to describing the problem of clusters by using graphs. Based on graph theory, node V represents the data points in the space and Edge E represents the closeness between the each pair of data points. Single linkage agglomeration is equivalent to maximally connected subgraph whereas complete linkage agglomeration is equivalentto maximally complete subgraphs.

George Karypis et al [24] developed a technique known as CHAMELEON is type of clustering algorithm based on agglomerative HC with k-nearest neighbour graph. In whichan edge is eliminated if both vertices are not within the k-nearest closest points related to each other. In this method, authors ignored the issue of scaling to large data sets that cannot fit in the main memory. It is a two-phase approach describedas follows,Phase -I Uses a graph partitioning algorithmic rule to divide the info set into a group of individual clusters.Phase-II uses agglomerative hierarchical mining algorithmrule to merge the clusters. Basically it works asfollows,

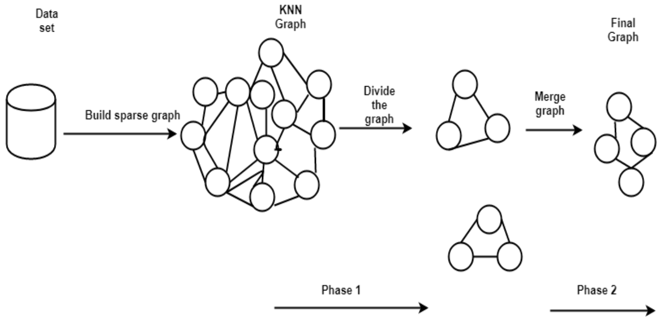

Figure 2: Steps involved in Chameleon

In Phase I, this method divides the connectivity graphinto sub clusters with the minimum edge cut i.e, the sum of the weight of the edges that straddle partitions, is minimized. CHAMELEON uses multilevel graph partitioning algorithms to find the initial sub-clusters. This method also uses hMETIS which is a part of graph partitioning rule. hMETIS is used to quickly produce high-quality partitioning for a wide range of unstructured graphs and hypergraphs [25] Chamelon is flexible i.e, it provides the characteristics of potentialclusters by combining the relative inter-connectivityand relative closeness. Relative inter connectivity is obtainedfrom the total weights of edges connecting the two clusters over the closeness of the clusters. A minimum cut (mincut) procedure, which is used to separate a graph with a minimum number of edges. CHAMELEON primarily use hMETIS libraries to split a cluster into two sub-clusters. Edge cut between 2 clusters is minimized and each one of these sub-clusters contains at least 25% of nodes in one cluster.Balance constraint is a graph partitioning approach tofind the sub-clusters. Within the allowed balance constraints hMETIS is effective in operating to find a bisection that minimizes the edge-cut. However, this balance constraint can break a natural cluster by using hMETIS.

CHAMELEON obtains the initial set of sub-clusters as follows, it first starts with all the points belonging to the same cluster. It then repeatedly selects the largest sub-cluster among the currentset of sub-clusters and uses hMETIS to bisect. This process ends when the larger sub-cluster contains fewer than a specified number of vertices i.e,MINSIZE. In general, MINSIZE should be set to a value that is smaller than the size of most of the clusters that we expect to find in the data set. The MINSIZE parameter controls the granularity of the initialclustering solution. At the same time, MINSIZE shouldbe sufficiently large such that most of the sub-clusters contain a sufficiently large number of nodes to allow us to evaluate the inter-connectivity and closeness of the items in each sub-cluster in a meaningful fashion.

In Phase II: Merging Sub-Clusters using a dynamic framework, CHAMELEON then uses an agglomerative hierarchical clustering method that combines together these small sub-clusters. The key step of agglomerative hierarchical algorithm is that of finding the pair of sub-clusters that are the most similar to each other. CHAMELEON select the most similar pairs of clusters by looking both at their relative inter-connectivity and their relative closeness by using the dynamic modelling framework. There are many ways to develop an agglomerative hierarchical clusteringalgorithm that takes into account both of these measures. CHAMELEON implemented in two different scheme i.e, the first scheme merges only those pairs of clusters whose relative inter-connectivity and relative closeness are user specified threshold. Once every cluster has been given the opportunity to merge with one of its adjacent clusters, and the combinations that have been selected are performed. The second scheme implemented in CHAMELEON uses a functionto combine the relative inter-connectivity and relativecloseness. The selected pair of clusters is to be merged that maximizes Chameleon function since our goal is to merge together pairs, for which both the relative inter-connectivity and the relative closeness are high.

The computational complexity of CHAMELEON depends on the amount of time required to compute the k-nearest neighbour graph.The amount of time required by CHAMELEONs two-phase clustering algorithm depends on the number of initial sub-clusters produced by the graph partitioning algorithm used in the first phase. The amount of time required by the Phase I of CHAMELEON depends on the amount of time required by hMETIS. The amount of time required by the second phase depends on the amount of time needed to compute the internal inter-connectivityand internal closeness for each initial as well as intermediate cluster, and the amount of time needed to select the most similar pair of clusters to merge.

CLICK is another agglomerative algorithm based onthe calculation of the minimum weight cut to form clusters [30].

Here, we use the weighted graph and the edge weights are assigned a new interpretation. By combining probability and graph theory, the edge weight between two node is calculated. CLICK assumes the similarity values within clustersand between the clusters and follow Gaussian distributions. CLICK recursively checks the current sub graph and generates a kernel list, which consists of the components satisfying some criterion function. Using kernels as basic set of clusters and CLICK carried out singleton clusters. These clusters contain only one node and merge to generate resultingclusters. Additional heuristics are provided to accelerate the algorithm performance.

CAST is another agglomerative algorithm used probabilistic model for graph based theory clustering algorithms [34]. Cast is the heuristics original theoretical version and creates clusters sequentially and each cluster begin with unassigned data point randomly. Relation between data pointand cluster is defined by

$a(i) = \sum I \varepsilon C0 S_{ij}$

where,

$$i \geq tC0$$

it means that the data point is highly related to the cluster and vice versa. Cast add more similar data points and deletes the less similar data points.

SudiptoGuha et al proposed CURE algorithm [26]. CURE is a novel hierarchical clustering Algorithm. In CURE, a constant number c of scattered points in a cluster are first chosen. The scattered points capture the shape and extent of the cluster and these points are next shrunk towards the centroid of the cluster by a fraction. These scattered points after shrinking are used as representatives of the cluster. The clusters with the closest pair of representative points are the clusters that are merged at each step of CUREs hierarchical clustering algorithm. There is an advantage is that it is fitted in main memory. Partitioning sample is that whenclusters in the data set became less dense, it implies poorquality of clustering. For speed up CURE algorithm sincewe have to increase the Random sampling. Moreover reduceexecution time in presence of large database using randomsampling and partitioning and increase the performance bypartition the data points into groups. It is 50% less expensivethan BIRCH and CURE algorithm detects cluster with non-sphericalshape. It works well when the database contains outliers, these are detected and eliminated additionally small clusters eliminated at the end of process.

3.2.2 Divisive Hierarchical Method

Divisive method is a hierarchical clustering method. This methods start with a large cluster, so it is also called top down approach. Each and every iteration large cluster is divided into smaller clusters and it continues until each individual cluster is formed. [5] TengkeXiong et al proposed an DHCC (Divisive Hierarchical Clustering of Categorical Data), it is a divisive hierarchical algorithm for categorical data. This method uses MCA(Multiple Correspondence Analysis) is carried out by performing a standard correspondence analysis(CA) on an indicator matrix. An indicator matrix is a binary objects-by-values matrix of relationships indicatingwhich object contains what categorical values. The burt matrix, also called a cross-table of categorical values, is a values-by-values matrix. This matrix representing the correlation of each pair of categorical values, in terms of the number of objects that share the pair of categorical values. MCA on an indicator matrix can analyse patterns among both objects and categorical values. But MCA on a burt matrix keeps the correspondence among categorical values butdiscards the correspondence among objects. Since our main goal is to cluster categorical objects rather than to discover the relationship among the categorical values. The indicator matrix rather than the burt matrix is used in DHCC. MCA calculation on the indicator matrix involves a measure of the Chi-square dissimilarity between a single object and a set of objects.

DHCC starts with an all clusters containing all the categorical objects, and repeatedly chooses one cluster to split into two sub clusters. A binary tree is employed to represent the hierarchical structure of the clustering results. In DHCC, divides the cluster $C_p$involves finding a sub optimal solution to the optimization problem on the data set $C_p$with K =2. DHCC consists of two phases, initial splitting phase andrefinement phase. In initial splitting based on MCA andis performed as follows, to bisect cluster modulus of cluster objects, and the apply MCA on the indicator matrix and we get principal coordinates of row. Each object whose first coordinate less than principal coordinates goes to left child and object whose first coordinate greater than principalcoordinates goes to right child. In this phase, MCA plays an important role in analysing the variance and data distribution of the objects.

The refinement phase makes an attempt to enhance the quality of the bisection by relocating the objects from the cluster itself. After the initial division, for every object from the parent cluster goes to sub cluster. The refinement phase tries to explore the division quality by finding the sub cluster. DHCC always improves the compactness of the cluster due to its division. Division of the cluster always increases the homogeneity of the resulting sub clusters. All the leaf clusters are subject to divide and the leaf clusters are often chosen either through a depth first or a breadth first search method. If divide a leaf node ends up in improvement of cluster quality and these leaf node may be a final cluster. The division is terminated until no leaf cluster are often divide to increase the global clustering quality. Initial division consumes more time than refinement of objects. Thus the refinement operation has linear time complexity. So each initial division is scalable to large data set. The time complexity of the entire DHCC algorithm is linear with respect to the number of the objects for clustering so DHCC is scalable for large data sets.

In DHCC, in the initial division phase, the coordinates of rows which are the output of the MCA calculation used to perform initial division are not affected by the order. The refinement phase runs like the traditional K-means algorithm that is, the cluster centers are updated at the end of each iteration, so as to called iterative refinement procedure. [23] Maurice Roux proposed, divisive hierarchical algorithms are built top-down i.e, starting with the whole sample in a unique cluster they split this cluster into two sub clusters which are in turn divided into sub clusters and so on. At each step the two new clusters make up so called bipartition of the former. Therefore it is too time consuming to base a division protocol on the trial of all possible bipartitions. [28] NajvaIzadpanah proposed a divisive hierarchical clustering based multidimensional indexing structure which is efficient in high dimensional feature spaces. A projection pursuit method has been used for finding a component of the data which data are projections onto it maximizes the approximation of entropy for preparing essential information in order to partitioning of the data space.

[29] R.Datta et al and [36] M s Lew et al proposed one of the current techniques for image retrieval is content-based image retrieval. In content based image retrieval approach visual features such as colour feature, texture feature, shape feature and local features are automatically extracted from the image objects and organized as feature vectors. Then at search phase, after selecting the query image by user, retrieval engine retrieves the most similar images to the query image by performing similarity comparison between query feature vector and all the feature vectors in database. The proposed multidimensional indexing structure NO-NGP tree method consist of four stages, from that one of the stage is clustering of data projections where a clustering algorithm such as k-means [31] will be applied on the data projections. As the result two projection sub clusters and their centroids will be obtained. k-means clustering with k=2 has been applied on projections. So two obtained centroids are considered as locations with more density and the mean of these two centroids is considered as the location with less density. M. Venkat Reddy [32] et al developed a divisive hierarchical clustering with K-means and Agglomerative hierarchical clustering. In hierarchical clustering by finding the initial k centroids in a fixed manner instead of randomly choosing them. In which k centroids are chosen by dividing the one dimensional data of a particular cluster into k parts and then sorting those individual parts separately. Then the middle elements id in each part is mapped to id of m-dimensional data. The m-dimensional elementswhose ids are matched, taken as initial k centroids of any cluster. Then applying the agglomerative hierarchical clustering on the result and each element has its own individual cluster, where the clusters are merger based on the centroid distance. Then finally obtaining k-clusters. A Divisive hierarchical clustering is one of the most important tasks in data mining and this method works by grouping objects into a tree of clusters. The top-down strategy is starting with all objects in one cluster. It subdivides the clusters into smaller and smaller pieces by k-means algorithm by choosing initial k centroids in a fixed manner to get an efficient result, until each object form a cluster on its own and by applying Agglomerative Hierarchical Clustering on the result to get the efficient k cluster with high accuracy.

3.2.2 Density Based Method

Jorg Sander et.al proposed [33] Density-Based Clustering for spatial databases. In which DBSCAN relies on a density based notion of clusters and is designed to discover clusters of arbitrary shape as well as to distinguish noise. The generalized algorithm called GDBSCAN can cluster point objects as well as spatially extended objects according to both their spatial and their non-spatial attributes. Moreover four applications using 2D points (astronomy), 3D points (biology),5D points (earth science) and 2D polygons (geography) are presented. The algorithm says that find a density connected set, GDBSCAN starts with an arbitrary object p and retrieves all objects density reachable from p with respect to NPred and MinWeight. If p is a core object, this procedure yields a density connected set with respect to NPred and MinWeight. Applications of GDBSCAN include cluster a spectral space (5D points) created from satellite images in different spectral channels which is a common task in remote sensing image analysis. Second application is molecular biology. Third application uses astronomical image data (2D points) showing the intensity on the sky at different radiowavelengths etc. A performance evaluation, analytic as well as experimental, showed the effectiveness and efficiency of GDBSCAN on large spatial databases. [35] Markus M. Breunig et.al proposed LOF: Identifying Density-Based Local Outliers. It introduce a new method for finding outliers in a multidimensional data set. We introduce a local outlier (LOF) for each object in the data set, indicating its degree of outlierness. This is, to the best of our knowledge, the first concept of an outlier which also quantifies how outlying an object is. The outlier factor is local in the sense that only a restricted neighbourhood of each object is taken into account.

4. Conclusion

Clustering will be thought-about the most necessary unsupervised learning Problem. It deals with finding a structure during a group of unlabel knowledge. A cluster is thus a set of objects that are similar between them and are dissimilar to the objects happiness to alternative clusters. The most usually used algorithms in clustering are hierarchical and Partitioning. Applications of clustering in numerous field like in Biology, Information Retrieval, Geology, Medicine, Pattern detection etc. So this paper shows the many clustering techniques and it deserves and shortcomings. So this techniques are largely used for business applications like customer segmentation, fraud detection, market segmentation etc that store great amount of information.

References

[1] Saroj, T. C., and Chaudhary, T., 2015. “Study on various clustering techniques”. International Journal of Computer Science and Information Technologies, 6(3),pp. 3031–3033.

[2] Tan, P.-N., Steinbach, M., and Kumar, V., 2013. “Data mining cluster analysis: basic concepts and algorithms”. Introduction to data mining.

[3] Gupta, P., 2011. “Robust clustering algorithms”. PhD thesis, Georgia Institute of Technology.

[4] Gupta, P., 2011. “Robust clustering algorithms”. PhD thesis, Georgia Institute of Technology.

[5] Xiong, T., Wang, S., Mayers, A., and Monga, E., 2012. “Dhcc: Divisive hierarchical clustering of categorical data”. Data Mining and Knowledge Discovery, 24(1), pp. 103–135.

[6] Berkhin, P., 2006. “A survey of clustering data mining techniques”. In Grouping multidimensional data. Springer, pp. 25–71.

[7] Everitt, B., Landau, S., Leese, M., and Stahl, D., 2001."Cluster analysis. 4th". Arnold, London.

[8] Hansen, P., and Jaumard, B., 1997. "Cluster analysis and mathematical programming". Mathematical programming, 79(1-3), pp. 191–215.

[9] Jain, A. K., and Dubes, R. C., 1988. "Algorithms for clustering data".

[10] Jain, A. K., Murty, M. N., and Flynn, P. J., 1999. "Data clustering: a review". ACM computing surveys (CSUR), 31(3), pp. 264–323.

[11] Kolatch, E., et al., 2001. "Clustering algorithms for spatial databases: A survey". PDF is available on the Web, pp. 1–22.

[12] Wilks, D. S., 2011. "Cluster analysis". In International geophysics, Vol. 100. Elsevier, pp. 603–616.

[13] Grabmeier, J., and Rudolph, A., 2002. "Techniques of cluster algorithms in data mining". Data Mining and knowledge discovery, 6(4), pp. 303–360.

[14] Migu´eis, V. L., Camanho, A. S., and e Cunha, 2012. "Customer data mining for lifestyle segmentation". Expert Systems with Applications, 39(10), pp. 9359–9366.

[15] Ng, R. T., and Han, J., 2002. "Clarans: A method for clustering objects for spatial data mining". IEEE transactions on knowledge and data engineering, 14(5), pp. 1003–1016.

[16] Huang, J. Z., Ng, M. K., Rong, H., and Li, Z., 2005."Automated variable weighting in k-means type clustering".IEEE Transactions on Pattern Analysis andMachine Intelligence, 27(5), pp. 657–668.

[17] Tsai, C.-Y., and Chiu, C.-C., 2008. "Developing a feature weight self-adjustment mechanism for a k-means clustering algorithm". Computational statistics & data analysis, 52(10), pp. 4658 4672.

[18] Chen, X., Xu, X., Huang, J. Z., and Ye, Y., 2013. "Tw-k-means: automated two-level variable weighting clustering algorithm formultiview data". IEEE Transactions on Knowledge and Data Engineering, 25(4), pp. 932–944.

[19] Rokach, L., and Maimon, O., 2005. "Clustering methods".In Data mining and knowledge discovery handbook.Springer, pp. 321–352.

[20] Green, P. E., Kim, J., and Carmone, F. J., 1990. "Apreliminary study of optimal variable weighting in k-meansclustering". Journal of Classification, 7(2),pp. 271–285.

[21] Xu, R., and Wunsch, D., 2005. "Survey of clusteringalgorithms". IEEE Transactions on neural networks,16(3), pp. 645–678.

[22] Li, C., and Biswas, G., 2002. "Unsupervised learningwith mixed numeric and nominal data". IEEE Transactionson Knowledge & Data Engineering(4), pp. 673–690.

[23] Zhang, T., Ramakrishnan, R., and Livny, M., 1996. "Birch: an efficient data clustering method for very large databases". In ACM Sigmod Record, Vol. 25, ACM, pp. 103–114.

[24] Karypis, G., Han, E.-H., and Kumar, V., 1999. "Chameleon: Hierarchical clustering using dynamicmodeling". Computer, 32(8), pp. 68–75.

[25] Karypis, G., and Kumar, V., 1998. "Multilevel k-way partitioning scheme for irregular graphs". Journal of Parallel and Distributed computing, 48(1), pp. 96–129.

[26] Guha, S., Rastogi, R., and Shim, K., 1998. "Cure: an efficient clustering algorithm for large databases". In ACM Sigmod Record, Vol. 27, ACM, pp. 73–84.

[27] Roux, M., 2015. "A comparative study of divisive hierarchical clustering algorithms". arXiv preprint arXiv:1506.08977.

[28] Izadpanah, N., 2015. "A divisive hierarchical clustering-based method for indexing image information".arXiv preprint arXiv:1503.03607.

[29] Datta, R., Joshi, D., Li, J., and Wang, J. Z., 2008. "Image retrieval: Ideas, influences, and trends of the new age". ACM Computing Surveys (Csur), 40(2), p. 5.

[30] Lew, M. S., Sebe, N., Djeraba, C., and Jain, R., 2006. "Content-based multimedia information retrieval: State of the art and challenges". ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2(1), pp. 1–19.

[31] MacQueen, J., et al., 1967. "Some methods for classification and analysis of multivariate observations". In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Vol. 1, Oakland, CA, USA, pp. 281–297.

[32] Reddy, M. V., Vivekananda, M., and Satish, R. "Divisive hierarchical clustering with k-means and agglomerative hierarchical clustering".

[33] Sander, J., Ester, M., Kriegel, H.-P., and Xu, X., 1998. "Density-based clustering in spatial databases: The algorithm gdbscan and its applications". Data mining and knowledge discovery, 2(2), pp. 169–194.

[34] Ben-Dor, A., and Yakhini, Z., 1999. "Clustering gene expression patterns". In Proceedings of the third annual international conference on Computational molecularbiology, ACM, pp. 33–42.

[35] Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J., 2000. "Lof: identifying density-based local outliers". In ACM sigmod record, Vol. 29, ACM, pp. 93–104.

Author Profile

Neethu C M ,post graduation student in college of engineering kidangoor. Specialization in computer and information science. Received the graduation in computer science and Engineering from Chennai University.

Anitha Abraham, She is working as an Asst.Professor at college of engineering kidangoor. She received graduation in computer science and engineering from college of engineering kidangoor and post graduation in communication and network technology from MG University. Her area of interest includes Cryptography,Image Processing, Artificial Intelligence.

Linda Sebastian, She is working as an Asst.Professor at college of engineering kidangoor. She received graduation in computer science and engineering from School of engineering CUSAT and post graduation in Computer and information Science from Dep.ofComputer Science CUSAT. Her area of interest includes DataMining,Image Processing, Natural Language Processing,and Information retrieval.