**DATA MINING ALGORITHM TECHNIQUES**

**A Comparative Study of Various Clustering Algorithms in Data Mining**

**INTRODUCTION:**

This paper is based on a indepth study of various algorithms used during the clustering process of a particular dataset. Data clustering is nothing but a process of arranging or placing similar data into individual groups. The partition can be of several group based on an individual entity. This research papers reviews six different techniques used while performing clustering process. WEKA tool is used inorder to analyse these six techniques. Clustering separates the similar as well as the dissimilar data into different clusters. This is a major data mining algorithm technique as it is considered as the most important unsupervised learning technique. On a whole clustering is the unsupervised classification of patterns in groups such as observations , items or feature vectors. Clustering algorithms are not only used to organize and categorize data but also useful for data compression and model construction.Another reason for clustering is to discover relevance knowledge in data. Data cluster are created to meet specific requirements that cannot created using any of the categorical levels .

**PUBLICATIONS AND RESEARCHERS :**

About the publishers, this paper is published at international journal of engineering research and applications (IJERA) on june 2012. Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar, Nidhi Gupta together from GLNA institute of technology. Mathura has completed this paper based on clustering algorithm techniques.

**DATASET:**

These six techniques are tested using a banking dataset related to customer information in weka tool. The banking dataset consists of 11 attributes and 600 entries.

The test results are shown in a table format for each technique used.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | **Number of clusters** | **Cluster Instances** | **Number of Iterations** | **Within clusters sum of squared errors** | **Time taken to build model** | **Log likelihood** | **d Unclustered Instances** |
| K-Means Algorithm | 2 | 0:254(42%) 1:346(58%) | 4 | 2016.6752520938 053 | 0.08 Seconds |  | 0 |
| EM Algorithm | 6 | 0:31 (5%) 1:97 (16%) 2:65 (11%) 3:184(31%) 4:92(15%) 5:131(22%) |  |  | 76.94 seconds | -21.09024 | 0 |
| DBSCAN | 3 | 0:10 (40%) 1:6 (24%) 2:9 (36%) |  |  | 1.03 Seconds |  | 575 |
| Hierarchical Clustering | 2 | 0:599(100%) 1:1 (0%) |  |  | 1.16 Seconds |  | 0 |
| Density based Clusters | 2 | 0:239(40%) 1:361(60%) | 4 | 2016.6752520938 053 | 0.06 Seconds | -22.04211 | 0 |
| OPTICS | 0 |  |  |  | 1.37 seconds |  | 600 |

**TECHNIQUES:**

In this section a detailed discussion of each technique is presented. Implementation and results are presented in the following sections.

**K-Means clustering:**

This technique is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belong to the cluster with nearest mean.The algorithm is called k-means, where k is the number of clusters we want, since a case is assigned to the cluster for which its distance to the cluster mean is the smallest. The action in the algorithm centers around finding the k-means.Next, we compute the cluster means again, using the cases that are assigned to the cluster; then, we reclassify all cases based on the new set of means. We keep repeating this step until cluster means don’t change much between successive steps. Finally, we calculate the means of the clusters once again and assign the cases to their permanent clusters.

**Hierarchical clustering:**

This technique builds a cluster hierarchy 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.

**Agglomerative (bottom up)**

1. Start with 1 point (singleton).

2. Recursively add two or more appropriate clusters.

3. Stop when k number of clusters is achieved

**Divisive (top down)**

1. Start with a big cluster.

2. Recursively divides into smaller clusters.

3. Stop when k number of clusters is achieved.

Steps of hierarchical clustering:

* Start by assigning each item to a cluster, so that if we have N items, we now have N clusters, each containing just one item.
* Let the distances (similarities) between the clusters the same as the distances (similarities) between the items they contain.
* Find the closest (most similar) pair of clusters and merge them into a single cluster
* Compute distances (similarities) between the new cluster and each of the old clusters.

**DB Scan Clustering:**

This technique finds all clusters properly, independent of the size, shape, and location of clusters to each other, and is superior to a widely used Clarans method.DBscan is based on two main concepts: density reachability and density connectability. These both concepts depend on two input parameters of the dbscan clustering: the size of epsilon neighborhood e and the minimum points in a cluster m.An open set in the Euclidean space can be divided into a set of its connected components. The implementation of this idea for partitioning of a finite set of points requires concepts of density, connectivity and boundary.

Density Functions:

* Hinneburg & Keim [1998] shifted the emphasis from computing densities pinned to data points to computing density functions defined over the underlying attribute space. They proposed the algorithm DENCLUE (Densitybased Clustering).

**OPTICS :**

OPTICS is the accronym of “"Ordering Points to Identify the Clustering Structure” is an algorithm for finding density-based clusters in spatial data.Its basic idea is similar to DBSCAN, but it addresses one of DBSCAN's major weaknesses: the problem of detecting meaningful clusters in data of varying density .In order to do so, the points of the database are (linearly) ordered such that points which are spatially closest become neighbors in the ordering. It is required to cut off the density of clusters that is no longer considered to be interesting and to speed up the algorithm this way.

**EM Algorithm:**

This technique is an iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables . Tis iteration alternates between performing an exception step.

* **Expectation**: Fix model and estimate missing labels.
* **Maximization**: Fix missing labels (or a distribution over the missing labels) and find the model that maximizes the expected log-likelihood of the data.

Steps for EM algorithm clustering:

Alternate steps until model parameters don’t change much:

* **M step**: Choose new parameters for model to maximize expected log-likelihood of observed data and hidden variables.
* **E step**: Estimate distribution over labels given a certain fixed model.

**Major Findings:**

After analyzing the results of testing the algorithms and running them under different factors and situations, we can obtain the following conclusions:

1. Performance of K-Means algorithm increases as the RMSE decreases and the RMSE decreases as the number of cluster increases.
2. The performance of K-Means algorithm is better than Hierarchical Clustering algorithm
3. All the algorithms have some ambiguity in some (noisy) data when clustered.
4. The quality of EM and K-Means algorithm become very good when using huge dataset.
5. DBSCAN and OPTICS does not perform well on small datasets
6. K-Means and EM algorithm are very sensitive for noise in dataset. This noise makes it difficult for the algorithm to cluster data into suitable clusters, while affecting the result of the algorithm.
7. K-Means algorithm is faster than other clustering algorithm and also produces quality clusters when using huge dataset.
8. Hierarchical clustering algorithm is more sensitive for noisy data.

One important factor is normalization.Comparing between the results of algorithms using normalized data and non-normalized data will give different results. Of course normalization weill affect the performance of the algorithm and the quality of the results

**Potential relevance to the work:**

On overviewing the above six algorithm techniques , Two techniques were used in our work such as “K-Means” and “DBSCAN” were used based on the training and the test data .This study of techniques was useful in training our dataset in such a way that both the algorithm were perfectly induced on gaining the expected results. Hence the similar and the dissimilar data were grouped in different clusters based on its expectation and maximization.