Research Paper Engineering

An Empirical Comparison of K-Means and DBSCAN Clustering Algorithm

* Kaushik H. Raviya ** Kunjan Dhinoja

* Assistant Professor, B.H. Gardi Vidyapith, Anandpar, Gujarat, India

** B.E. (CSE) Student, B.H. Gardi Vidyapith, Anandpar, Gujarat, India

ABSTRACT

Data mining is the process of automatically finding useful information in large data repositories. It is commonly used in marketing, fraud detection, AI, information retrieval, machine learning ,pattern recognition and now gaining a broad way in other fields also. Clustering is a data mining technique used for discovering groups and identifying interesting distributions in the underlying data. Clustering algorithms used in data mining such as K-means, Density based, K- medoids, Hierarchical, Optics etc. This paper presents comparison on two clustering techniques which are K-means and DB scan respectively. The goal of this research is to enumerate the best technique from above analyzed under a given dataset and provide a fruitful comparison result with respect to time, cost and effort for the comprehensive review of different clustering algorithms in data mining and can be used for further analysis or future development.

Keywords : Clustering, K-means, DB- scan

1. Introduction

These days the role of data generation and collection, are producing data sets from variety of scientific disciplines.

Today's world large quantities of data is being accumulated and seeking knowledge from massive data is one of the most fundamental attribute of Data Mining. It consist of more than just collecting and managing data but to analyze and predict also.

The scope of this paper is modest: to provide an introduction to **cluster analysis** in the field of data mining, where, it is to define data mining to be the discovery of useful, but non-obvious, information or patterns in large collections of data.

Clustering is a process of dividing a dataset into mutually exclusive groups such that the members of each group are as "close" as possible to one another, and different groups are as "far" as possible from one another, where distance is measured with respect to all available variables.

Data could be large in two senses : in terms of size & in terms of dimensionality.

Also there is a huge gap from the stored data to the knowledge that could be clustered from the data. In the present paper comparison on two clustering technique is done with its fundamental knowledge.

2. K means:

K-Means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem where each cluster is represented by center of gravity of the cluster. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters . k|means used to cluster observation into related observation without any prior knowledge of those relationships.

The k-means algorithm works as follows:

a) Select initial centroid of the k clusters. Repeat steps b through c until the cluster membership stabilized.
b) Generate a new partition by assigning each data to its closest cluster centroid.
c) Compute new cluster centroid for each cluster.

Popular K-means algorithm is Lloyd's algorithm which we call filtering algorithm. Lloyd's algorithm is based on the simple observation that the optimal placement of a center is at the centroid of the associated cluster. For e.g. Given any set of k centers Z, for each center z ÎZ, let V(z) denote its neighborhood, that is, the set of data points for which z is the nearest neighbor. In geometric terminology, V(z) is the set of data points lying in the Coronoid cell of z . Each stage of Lloyd's algorithm moves every center point z to the centroid of V(z) and then updates V(z) by recomputing the distance from each point to its nearest center. These steps are repeated until some convergence condition is met. This algorithm is simple to implement and run, relatively fast, easy to adapt, and common in practice.

The most widely used convergence criteria for the k-means algorithm is minimizing the SSE.

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

Denotes the mean of cluster cj and mi denotes the no. of instances in ci . As good clustering with smaller K can have a lower SSE than a poor clustering with higher K. The k-means algorithm always converges to a Local minimum. The k-means algorithm updates cluster centroids till local minimum is found. So the computational complexity of this algorithm is O (nkl), where n is total no of objects in data set, k is a required no of clusters and l is a no of iteration. And for high dimensionality data set time complexity is O (nklm),where m is no of dimensions.

3. DBSCAN (Density-based spatial clustering of applications with noise):-

DB scan is one of Density based algorithm. In Density based algorithm two algorithms are 1)DB scan 2)SNN. Clusters are identified by looking at the density of points. Regions with a high density of points depict the existence of clusters whereas regions with a low density of points indicate clusters of noise or clusters of outliers. The algorithm grows regions with suf-

ficiently high density into clusters and discovers clusters of arbitrary shape in spatial databases with noise. DBSCAN having two parameters: epsilon e (eps) and (minPts) the minimum number of points that must be exists in eps. It starts with an arbitrary starting point that has not been visited. This point's e-neighborhood is retrieved, and if it contains sufficiently many points, a cluster is started. Otherwise, the point is labeled as noise. Note that this point might later be found in a sufficiently sized e -environment of a different point and hence be made part of a cluster.

If a point is found to be a dense part of a cluster, its e neighborhood is also part of that cluster .hence all points within a cluster are added and this process is repeated until the density connected clusters completely found and then further processes are proceed and leading to discovery of further cluster or noise. DB scan visits each points of database possibly multiple times. The algorithm which executes one of the query for each points and the overall run time complexity is obtained as O (n*logn).

Parameters	K-means	DB scan
Fundamentals	K means algorithm can be build by classify data to groups of objects based on their attributes/ features in to k number of groups.	DB scan is based on the density reachability and density connectivity Where dense area objects separated by less dense areas.
Work of classifier	This algorithm partitions objects in a data set into a fixed number of K disjoint subsets. For each cluster, the partitioning algorithm maximizes the homogeneity	In this algorithm objects are initially assumes to be unassigned. Db scan then chooses arbitrary object p from data set .if it finds p as a core object then finds all density connected objects bases on eps and mints and if p is not a core object then considered as a noise and move to next object.
Pros	- simple to implement -which is very fast to execute -scalable -works well for Euclidian data -convergence can be done to local minima not global minimum	- It can find arbitrarily shaped clusters and also find clusters completely surrounded by different clusters. - it is robust to noise -do not need any priori k deterministic -it requires just two points which are very insensitive to the ordering of the points in the database.
Cons	- possibility of many loop turns without significance changes in clusters - works only for well-shaped clusters - sensitive to outlier, noise -must know k priori	- datasets with varying densities are problematic -requires connected region of sufficiently high density
Applications	-Market segmentation, computer vision, astronomy to agriculture, geostatic etc.	-Satellite images, anomaly detection in temperature data , x –ray crystallography etc.
Summary	- in k-means the k is a number partitions to construct. - it uses the iterative relocation technique that attempts to improve partitioning for moving objects.	-the clusters based on the notion of density -It either grows clusters according to the density of neighborhood objects (e.g.DB scan) or according to some density function (such as in DENCLUE). OPTICS is a densitybased method that generates an augmented ordering of the clustering structure of the data.

Table 1 : Comparison of K Means and DB scan

5. Experimental Work:

For the practical scenario comparison on these techniques is done through WEKA. WEKA is a data mining system developed by the University of Waikato in New Zealand that implements data mining algorithms using the JAVA language. Here we have used a common database (which is a vote database) for all the two mechanisms, so that we can differentiate their parameters on a single instance. This vote database consist of 17 attributes (attributes like immigration, class, crime etc) and 435 instances.

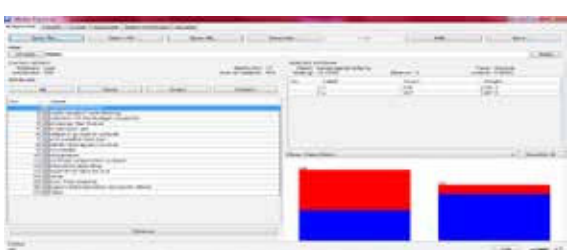

Figure 1: WEKA 3.7.7 - Explorer window

The above is the explorer window in WEKA tool with the iris dataset loaded. Here we can also analyze the data in the form of graph as shown above in visualization section with red and blue code. In WEKA, all data is considered as instances features in the data are known as attributes. The simulation results are partitioned into several sub items for easier analysis and evaluation. On the first part, correctly and incorrectly clustered instances will be partitioned in numeric and percentage value and subsequently some of squared errors , their time taken to build and loglikelyhood.

Figure 2 : Clustering Result

This dataset is measured and analyzed with 7 parameters as shown in table . Based on Table 2 given below we can clearly see that the accurate clustering is provided in less time by K means. From experimental comparison we can say that k means is best from two.

Clustered (Total Instances, 150)	Number of clusters	Number of iterations	Clustered Instances % (value)	unflustered Instances % (value)	Time Taken To build model (Seconds)	Within cluster sum of squared errors	Log likelihood
K means	2	3	0:214(49%) 1:221(51%)	0	0.07sec	1510.0	
DB scan	2	3	0:205(47%) 1:230(53%)	0	0.1sec	1510.0	-8.03831

Table 2 : Simulation Result of each Algorithm

6. Conclusion:
Clustering is one of the most popular techniques in data mining when complexity comes in to picture. Here we have compared algorithms theoretically and experimentally on parameters such as time taken to build model, clustered instances, etc. Apart from these data sets we have analyzed algorithms on 7 different datasets and from that we conclude that k means is efficient then DB scan.

7. Future work:
In future we will work increasing the accuracy of the k means algorithm where it is lacking behind other recently developed mechanisms.

REFERENCES

[1] Performance analysis of k-means with different initialization methods for high dimensional data by Tajunisha1 and Saravanan Department of Computer Science, Sri Ramakrishna College of Arts and Science (W) Coimbatore, Tamilnadu, India(IJAIA) | [2] A Novel Density based improved k-means Clustering Algorithm – Dbkmeans by K. Mumtaz1 and Dr. K. Duraiswamy 2,1 Vivekanandha Institute of Information and Management Studies, Tiruchengode, India | [3] A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise by Martin Ester, Hans-Peter Kriegel, J&g Sander, Xiaowei Xu | Institute for Computer Science, University of Munich Oettingenstr.6 7, D-80538M tinchen, Germany | [4] Thair Nu Phyu "Survey of Classification Techniques in Data Mining",Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol IIMECS 2009, March 18 - 20, 2009, Hong Kong | [5]] A Comparative Study of Various Clustering Algorithms in Data Mining Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar, Nidhi Gupta GLNA Institute of Technology, Mathura | [6] Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data by Levent Ertöz , Michael Steinbach,Vipin Kumar University of Minnesota Minneapolis, MN USA | [7] Extending K-Means Clustering for Analysis of Quantitative Structure Activity Relationships (QSAR) by Robert William Stanforth School of Computer Science and Information Systems Birkbeck College, University of London | [8] An Efficient Clustering Algorithm for Outlier Detection by S.Vijayarani , S.Nithya Bharathiar University , Coimbatore