Impact Factor (SJIF): 3.632

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-2444

Volume 3, Issue 5, May-2016

Clustering: A significant data mining technique

[1]Garima Goel
[2]Raman Chawla
[1,2]*Department of Computer Science & Engineering*
[1,2]N C College of Engineering, Israna
Haryana, India
[1]garimagoel06@gmail.com
2ramanchawla.cs @ncce.edu

***Abstract*— Data mining approach and all its technology is used to extort the unknown patterns from a huge set of data for real time and business applications. Using clustering and classification algorithm, the unlabeled data from huge data set can be classified in a supervised and unsupervised manner. Cluster analysis or clustering is the delegation of a set of observations into subsets (called *clusters*) so that clarification in the same cluster is related in some sense. This paper conveys a concise overview of clustering which is one of the major concepts used in data mining.**

Keywords*— *Data Mining,Clustering,K-Means,MCL.

I. Introduction

Data Mining, the extraction of hidden prognostic information from giant databases, is a dominant new technology with enormous potential to help companies focus on useful information in their data warehouses. It is the analysis step of the "knowledge discovery in databases" process.[1]

The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown, interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule mining).

A "clustering" is essentially a set of such clusters, usually containing all objects in the data set. It is the task of grouping a set of objects in such a way that objects in the same group are more similar in some sense or another to each other than to those in other groups . Data modelling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis. From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. From a practical perspective clustering plays an outstanding role in data mining applications such as scientific data exploration, information retrieval and text mining, spatial database applications, Web analysis, CRM, marketing, medical diagnostics, computational biology, and many others[2].

The notion of a "cluster" cannot be precisely defined, which is one of the reasons why there are so many clustering algorithms. There is a common denominator: a group of data objects. However, different researchers employ different cluster models, and for each of these cluster models again different algorithms can be given. The notion of a cluster, as found by different algorithms, varies significantly in its properties. [3] With the user centric nature of service and the user's lack of prior knowledge on the distribution of the raw data, one challenge is on how to associate user quality requirements on the clustering results with the algorithmic output properties (e.g. number of clusters to be targeted).[4]

Clustering algorithms can be categorized based on their cluster model, as listed below:

1.1 Connectivity-based clustering

Connectivity based clustering, also acknowledged as *hierarchical clustering*, has its foundation based on objects being more allied to close by objects than to objects beyond away. These algorithms hook up "objects" to form "clusters" based on their distance.

1.2 Centroid-based clustering

In centroid-based clustering, clusters are represented by a central vector, which may not necessarily be a member of the data set. When the number of clusters is fixed to k, *k*-means clustering gives a formal definition as an optimization problem: find

the k cluster centers and assign the objects to the nearest cluster center, such that the squared distances from the cluster are minimized.

1.3 Distribution-based clustering

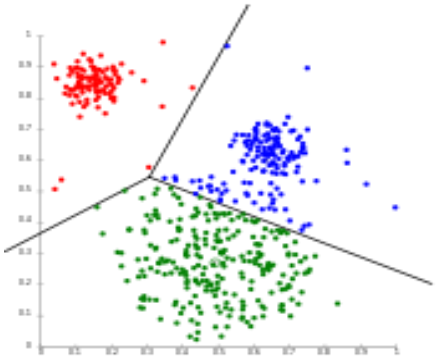
The clustering model most closely related to statistics is based on distribution models. Clusters can then easily be defined as objects belonging most likely to the same distribution. A convenient property of this approach is that this closely resembles the way artificial data sets are generated: by sampling random objects from a distribution.

1.4 Density-based clustering

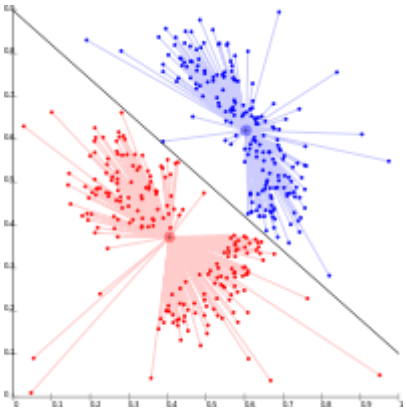
In density-based clustering, clusters are defined as areas of higher density than the remainder of the data set. Objects in these sparse areas - that are required to separate clusters - are usually considered to be noise and border points.

K-Means Algorithm

Clustering has been a widely studied problem in a variety of application domains including neural networks, AI, and statistics. The k-means method has been shown to be effective in producing good clustering results for many practical applications. However, a direct algorithm of k-means method requires time proportional to the product of number of patterns and number of clusters per iteration. This is computationally very expensive especially for large datasets



K-means seperates data into voronoi cells which assumes equal sized clusters.



K-means cannot represent density-based clusters

K-means organizes all the patterns in a k-d tree structure such that one can find all the patterns which are closest to a given prototype efficiently. Through this all the prototype sare potential candidates for the closest prototype at the root level.[5] Practical approaches to clustering use an iterative procedure (e.g. K-Means, EM) which converges to one of numerous local minima. It is known that these iterative techniques are especially sensitive to initial starting conditions. They present a procedure for computing a refined starting condition from a given initial one that is based on an efficient technique for estimating the modes of a distribution. The refined initial starting condition allows the iterative algorithm to converge to a “better” local minimum. They demonstrate the

application of this method to the popular K-Means clustering algorithm and show that refined initial starting points indeed lead to improved solutions. [6]

Markov Chain Lloyd

Lloyd's algorithm is based on the simple observation that the optimal placement of a center is at the centroid of the associated cluster (see. Given any set of k centers Z, for each canter z 2 Z, let V .z. denote its neighbourhood, that is, the set of data points for which z is the nearest neighbour. In geometric terminology, V .z. is the set of data points lying in the Verona 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 re-computing the distance from each point to its nearest center. These steps are repeated until some convergence condition is met. Lloyd's algorithm assumes that the data are memory resident. Lloyd's algorithm can be applied as a post processing stage to improve the final distortion. A straightforward implementation of Lloyd's algorithm can be quite slow. The treatment quality is estimated by the quadratic error in the delivered profiles compared to the optimized, and the delivery time is estimated primarily by the number of segments.[7]

II. Conclusion

The K-means algorithm is a popular data-clustering algorithm. However, one of its drawbacks is the requirement for the number of clusters, K, to be specified before the algorithm is applied.[8] The basic procedure involves producing all the segmented images for 2 clusters up to *Kmax* clusters, where *Kmax* represents an upper limit on the number of clusters. [9] So to clarify this problem we will refer to generalized Lloyds 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.

Refrences

[1] In Year 2001, Glenn Fung performed a work,"A Comprehensive Overview of Basic Clustering Algorithms".
[2] In Year 1997, Pavel BerkhinHo performed a work,"Survey of Clustering Data Mining Techniques"
[3] In Year 2007, Andrea De Lucia, Michele Risi, and Genoveffa Tortora," Clustering Algorithms and LatentSemantic Indexing to Identify Similar Pages in Web Applications'.
[4] In Year 2011, YunWei Zhao, Chi-Hung Chi and Chen Ding performed a work, "Quality-Driven Hierarchical Clustering Algorithm For Service Intelligence Computation".
[5] Khaled Alsabti, Sanjay Ranka, Vineet Singh "An Efficient K-Means Clustering Algorithm" Information Technology Lab (ITL) of Hitachi America, Ltd. while K. Alsabti and S. Ranka were visiting ITL.
[6] Paul S. Bradley Usama M. Fayyad, "Refining Initial Points for K-Means Clustering" Appears in Proceedings of the 15th International Conference on Machine Learning (ICML98), J. Shavlik (ed.), and pp. 91- 99. Morgan Kaufmann, San Francisco, 1998.
[7] Bjorn Hardemark Henrik Rehbinder, Johan Lof, "Rotating the MCL between segments improves performance in step and shoot IMRT delivery" Ray Search Laboratories AB, Stockholm, Sweden.
[8] D T Pham_, S S Dimov, and C D Nguyen, "Selection of K in K-means clustering", Manufacturing Engineering Centre, Cardiff University, Cardiff, UK. The manuscript was received on 26 May 2004 and was accepted after revision for publication on 27 September 2004.
[9] Siddheswar Ray and Rose H. Turi, "Determination of Number of Clusters in *K*-Means Clustering and Application in Colour Image Segmentation" School of Computer Science and Software Engineering Monash University, Wellington Road, Clayton, Victoria, 3168, Australia.