International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-3 Number-3 Issue-12 September-2013

Comparative Study of K-means and Robust Clustering

Shashi Sharma[1], Ram Lal Yadav[2]

Abstract

Data mining is the mechanism of implementing patterns in large amount of data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. Clustering is the very big area in which grouping of same type of objects in data mining. Clustering has divided into different categories – partitioned clustering and hierarchical clustering. In this paper we study two types of clustering first is Kmeans which is part of partitioned clustering. Kmeans clustering generates a specific number of disjoint, flat (non-hierarchical) clusters. Second clustering is robust clustering which is part of hierarchical clustering. This clustering uses Jaccard coefficient instead of using the distance measures to find the similarity between the data or documents to classify the clusters. We show comparison between Kmeans clustering and robust clustering which is better for categorical data.

Keywords

Data mining, clustering, Kmeans, Robust, Partitioned, Hierarchical, Jaccard coefficient, analysis

1. Introduction

Data warehouses an implementation mechanism which is being used for database analysis and reporting. Data warehouse is a fundamental warehouse of data which is created by integrating data from one or more distinct sources. **Data mining** is an interdisciplinary field of computer science for computational process of finding patterns in huge amount of data sets involving methods at the intersection of artificial intelligence, statistics, machine learning, and database systems.

Shashi Sharma, Kautilya Institute of Technology & Engineering, Jaipur.
Ram Lal Yadav, Kautilya Institute of Technology & Engineering, Jaipur.

The ultimate aim of the data mining methods is to get information from available data and convert it into a useful structure which can be us in future. **Cluster analysis** or **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) is more similar (in some sense or another) to each other than to those in other groups.
A good clustering process can generate high quality clusters in which:

- The similarity between intra-class and intra-clusters is high.
- The similarity between inter-classes should be low.
- The performance of a clustering result also depends on both the similarity measure used by the method and its implementation.

Applications of Clustering
Clustering has wide applications in

- Economic Science (especially market research).
- WWW:
  - Document classification
  - Cluster Weblog data to discover groups of similar access patterns
- Pattern Recognition.
- Spatial Data Analysis:
  - Create thematic maps in GIS by clustering feature spaces
- Image Processing

Categories of clustering
Clustering is divided into two major categories:-

- partition clustering
- hierarchical clustering

Kmeans clustering is a part of partition clustering and robust clustering is a part of hierarchical clustering. Generally the Partitioned clustering begins with a random partitioning. Hierarchical clustering is also known as connectivity based algorithm because hierarchical clustering matches objects based on the distance from clusters.

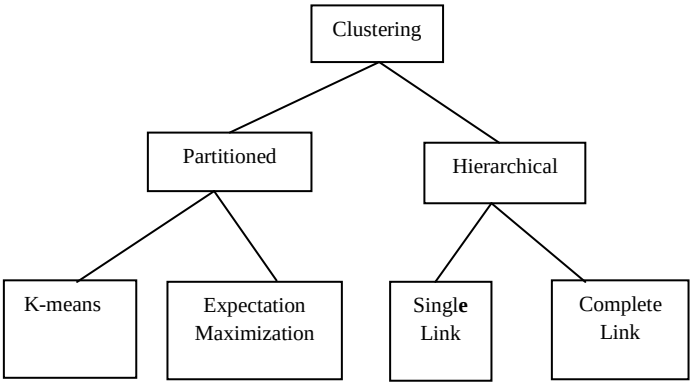

Figure 1: classification of clustering

2. Introduction of Kmeans clustering

K-Means method generates a particular number of non-hierarchical, disjoint clusters. It is the best way to generating globular clusters. The Kmeans clustering is statistical, non-supervised, possibilistic and iterative. Kmeans clustering has k clusters always. Each cluster always has minimum one item, and these clusters are unstructured and they do not overlap. Each item of cluster has closer property for its cluster with any other cluster because center of cluster is not always involved for closeness. Vector quantization method of kmeans clustering is used for processing of signal. For data mining, we use kmeans which is very popular cluster analysis method. Kmeans method uses dividing of n samplings for k clusters where each sampling is part of nearest mean of the cluster.

2.1 Kmeans algorithm process-

- The dataset is partitioned into K clusters and the data points are randomly assigned to the clusters resulting in clusters that have roughly the same number of data points.
- **For each data point:**
  - Calculate the distance from the data point to each cluster.
  - If the data point is closest to its own cluster, leave it where it is. If the data point is not closest to its own cluster, move it into the closest cluster.
- Repeat the above step until a complete pass through all the data points' results in no data point moving from one cluster to another. At this point the clusters are stable and the clustering process ends.
- The choice of initial partition can greatly affect the final clusters that result, in terms of inter-cluster and intracluster distances and cohesion.

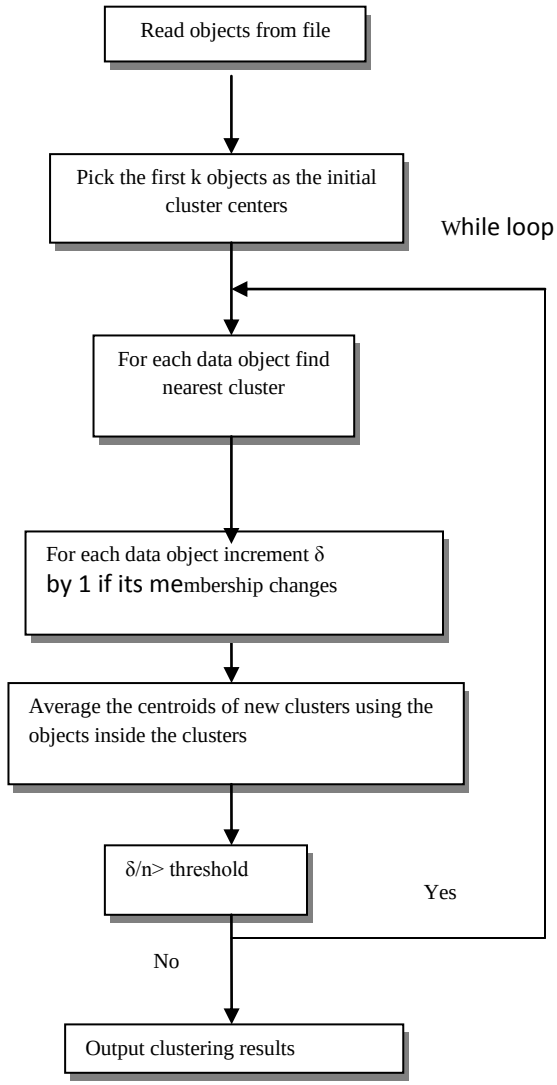

Figure 2: Kmeans Clustering Process

By using this clustering Result can vary significantly depending on initial choice of seeds. This clustering is simple and understandable and its items automatically assigned to clusters, but it can't produce high quality of clusters.

3. Introduction of Robust clustering

The process for hierarchical clustering can be depend as either being agglomerative and bottom-up or divisive and top-down, based on how the hierarchical

method is used on data sets. Robust hierarchical clustering algorithm which follows a more comprehensive approach to clustering that is, two similar points have similar neighborhoods, and then only the two points can be merged together in the same cluster. Robust Clustering using performs agglomerative hierarchical clustering and explores the concept of links for data with categorical attributes. Robust clustering method is based on the notion of neighbors & links. The goal of clustering is to group the similar data together. Robust clustering uses Jaccard coefficient because Jaccard's coefficient is a good similarity measure because it can find the similarity between the categorical data. For sets A and B of keywords used in the documents, the Jaccard coefficient may be defined as follows:

Similarity (A, B) = (|A ∩ B|) / (A U B|)

The ROCK algorithm:

- Initially each point is a separate cluster.
- The number of links between each pair of clusters is computed
- A goodness measure is calculated for all pair of clusters.
- The pair of clusters whose goodness measure is maximums merged
- The goodness measure between clusters is calculated again and the process of merging of cluster continues until a user specified number of clusters remain
- A second termination criteria can be that links between all the clusters become zero
- The Goodness Measure:

$$g(C_i, C_j) = \frac{link[C_i, C_j]}{(n_i + n_j)^{1+2f(\theta)} - n_i^{1+2f(\theta)} - n_j^{1+2f(\theta)}}$$

HereCi and Cj are two clusters

Link [Ci, Cj] stores the total number of cross links between them.

4. Comparison between Kmeans and robust clustering

Kmeans clustering is based on selection number of clusters in advance. It does not produce high quality clusters. But Robust clustering merges two points to the same cluster only when having similar neighborhoods. Simply, if clusters are to be meaningful, the similarity measure should be selected affectively. The target is to maximize the criterion function so that the intracluster similarity can be maximized and intercluster similarity can be minimized. And during clustering, goodness measure function helps in merging clusters which have highest goodness measure at each step of robust algorithm with the intent of maximizing criterion function. There is no need for selection of clusters in advance because of it works on links and neighbors.

So robust clustering does produce high quality clusters. Kmeans clustering is not capable to handle noisy data and outliers because an object with a very huge amount of value may significantly disfigure the allocation of the data. But robust clustering use Jaccard coefficient for calculating good similarity. By using Jaccard coefficient similarity can be fined in categorical data. Robust clustering also identified some noise which was totally dissimilar with every other document. So Kmeans clustering measures for categorical data may not suited well than robust clustering. In Kmeans clustering different initial partitions can result in different final clusters. It is helpful to rerun the program using the same as well as different K values, to compare the results achieved. It does not work well with non-globular clusters. Hence robust clustering is well suited for categorical data than Kmeans clustering.

Table 1: comparison b/w Kmeans and Robust Clustering

S.No.	K-means	Robust
1.	Faster(if k small)	Slower
2.	Cluster selection based on K cluster	Cluster selection based on link and neighbors
3.	It produce tighter cluster if data is globular	It produce loose cluster
4.	Not produce high quality clusters	Produce high quality clusters
5.	Not suited for categorical data	Suited for categorical data

5. Conclusion and future work

In this paper we study two types of algorithm Kmeans and robust clustering. Kmeans is partitioned type of clustering and robust is hierarchical clustering. We conclude that Kmeans clustering is simple and understandable but this clustering measure is not suitable for categorical data. Robust clustering works on links and neighbors. It uses Jaccard function as similarity function to find the

similarity between categorical data so it is well suited for categorical data. In future we will try other clustering in my project and extend my work in this area.We increase more databases to perform accuracy and efficiency using different types clustering.

Acknowledgement

The satisfaction that accompanies the successful completion of any task would be incomplete without the mention of people whose ceaseless cooperation made it possible, whose constant guidance and encouragement crown all efforts with success. I would like to place on record my deep sense of gratitude to Associate professor Ram Lal Yadav, Deptt. of computer Science and Engineering, KITE-SOM, Jaipur, Rajasthan, India for his generous guidance, help and useful suggestions.

References

[1] V. Ganti, J. Gehrke, R. Ramakrishan. CACTUS Clustering Categorical Data Using Summaries. KDD'99.
[2] A. K. Jain and R. C. Dubes. Algorithms for Clustering Data. Printice Hall, 1988.
[3] G. Caryopsis, E.-H. Han and V. Kumar. CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling. COMPUTER, 32(8): 68-75, 1999.
[4] D. Gibson, J. Kleinberg, and P. Raghavan. Clustering categorical data: An approach based on dynamic systems. In Proc. VLDB'98.
[5] S. Guha, R. Rastogi, and K. Shim. Cure: An efficient clustering algorithm for large databases. SIGMOD'98.
[6] S. Guha, R. Rastogi, and K. Shim. ROCK: A robust clustering algorithm for categorical attributes. In ICDE'99, pp. 512-521, Sydney, Australia, March 1999.
[7] A. Likas, Vlassis and J. J. Verbeek, "The global k-means clustering algorithm," in Pattern ecognition, vol. 36, no. 2, pp.451-461, 2003.
[8] S. Kantabutra and A. Couch. "Parallel K-means clustering algorithm on NOWs," in NECTEC Technical Journal, Vol. 1,No. 6, pp. 243–247, 2000.
[9] Dutta, Mala, A. Kakoti Mahanta, and Arun K. Pujari. "QROCK: A quick version of the ROCK algorithm for clustering of categorical data." Pattern Recognition Letters 26, no. 15 (2005): 2364-2373.
[10] P. Michaud. Clustering techniques. Future Generation Computer systems, 13, 1997.
[11] Huang, Zhexue. "A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining." In DMKD, p. 0. 1997.
[12] R. Ng and J. Han. Efficient and effective clustering method for spatial data mining. VLDB'94.
[13] Park, Sangmin, and Chandrajit Bajaj. "Feature selection of 3d volume data through multi-dimensional transfer functions." Pattern recognition letters 28, no. 3 (2007): 367-374.
[14] Anna Huang, "Similarity Measures for Text Document Clustering", Volume: 2008, Issue: April, Pages: 49–56, Mendeley.
[15] Boriah, Shyam, Varun Chandola, and Vipin Kumar. "Similarity measures for categorical data: A comparative evaluation." red 30, no. 2 (2008): 3.

Shashi Sharma received her Bachelor of Engineering from University of Rajasthan. Now she is doing her M. Tech. from Kautilya Institute of Technology & Engineering, Jaipur under **RTU** affiliation. She has several national & International publications.

Ram Lal Yadav received his first degree from MaharshiDayanand University, Rohtak in Computers, after that he received his Master of Engineering from BITS-Pilani, India. Now he is doing his Ph. D. from JKLU, Jaipur. He is now working as Associate Professor in Deptt. of Computer Science & Engineering at Kautilya Institute of Technology & Engineering, Jaipur. His research interest includes N/W, Data-mining & Software Engineering. He has guided 3 scholars at Masters Level. He is a member of Computer Society of India.