A Study of Improvement of Quality of the Clustering in Data Mining

B Hari Krishna[1], Dr. Vaibhav Bansal[2]

[1]Research Scholar of Sri Satya Sai University
[2]Research Supervisor of Sri Satya Sai University

Abstract

Clustering analysis is one of the main analytical methods in data mining. K-means is the most popular and partition based clustering algorithm. But it is computationally expensive and the quality of resulting clusters heavily depends on the selection of initial centroid and the dimension of the data. Several methods have been proposed in the literature for improving performance of the k-means clustering algorithm. Principal Component Analysis (PCA) is an important approach to unsupervised dimensionality reduction technique. This paper proposed a method to make the algorithm more effective and efficient by using PCA and modified k-means. In this paper, we have used Principal Component Analysis as a first phase to find the initial centroid for k-means and for dimension reduction and k-means method is modified by using heuristics approach to reduce the number of distance calculation to assign the data-point to cluster. By comparing the results of original and new approach, it was found that the results obtained are more effective, easy to understand and above all, the time taken to process the data was substantially reduced.

Keywords: *k-means, principal component analysis, dimension reduction*

1. INTRODUCTION

Data mining is a convenient way of extracting patterns, which represents knowledge implicitly stored in large data sets and focuses on issues relating to their feasibility, usefulness, effectiveness and scalability. It can be viewed as an essential step in the process of knowledge discovery. Data are normally preprocessed through data cleaning, data integration, data selection, and data transformation and prepared for the mining task. Data mining can be performed on various types of databases and information repositories, but the kind of patterns to be found are specified by various data mining functionalities like class description, association, correlation analysis, classification, prediction, cluster analysis etc. Clustering is a way that classifies the raw data reasonably and searches the hidden patterns that may exist in datasets. It is a process of grouping data objects into disjoint clusters so that thedata in the same cluster are similar, and data belonging to different cluster are differ. Many algorithms have been developed for clustering. A clustering algorithm typically considers all features of the data in an attempt to learn as much as possible about the objects. However, with high dimensional data, many features are redundant or irrelevant. The redundant features are of no help for clustering; even worse, the irrelevant features may hurt the clustering results by hiding clusters in noises. There are many approaches to address this problem. The simplest approach is dimension reduction techniques including principal component analysis (PCA) and random projection. In these methods, dimension reduction is carried out as a preprocessing step. K-means is a numerical, unsupervised, non-deterministic, iterative method. It is simple and very fast, so in many practical applications, the method is proved to be a very effective way that can produce good clustering results. The standard k-means algorithm [10, 14] is effective in producing clusters for many practical applications. But the computational complexity of the original k-means algorithm is very high in high dimensional data. Different methods have been proposed [4] by combining PCA with k-means for high dimensional data. But the accuracy of the k-means clusters heavily depending on the random choice of initial centroids. If the initial partitions are not chosen carefully, the computation will run the chance of converging to a local minimum rather than the global minimum solution. The initialization step is therefore very important. To combat this problem it might be a good idea to run the algorithm several times with different initializations. If the results converge to the same partition then it is likely that a global minimum has been reached. This, however, has the drawback of being very time consuming and computationally expensive.

Data mining is refers to “extracting or mining" knowledge from large amounts of data. There are many other terms carrying a similar or slightly different meaning to data mining, such as knowledge mining from databases, knowledge extraction, data/pattern analysis, data archaeology, and data dredging. Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery in Databases", or KDD. Alternatively, others view data mining as simply an essential step in the process of knowledge discovery in databases. Knowledge discovery as a process is depicted in Figure 1. and consists of an iterative sequence of the following steps:

- Data cleaning: To remove noise or irrelevant data.

- Data integration: multiple data sources are combined.

- Data Selection: Data relevant to the analysis task are retrieved from the database.

- Data Transformation: Data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.

- Data Mining: An essential process where intelligent methods are applied in order to extract data patterns.

- Pattern Evaluation: To identify the patterns representing knowledge based on measures.

- Knowledge Presentation: To visualization and knowledge representation techniques are used to present the mined knowledge to the user.

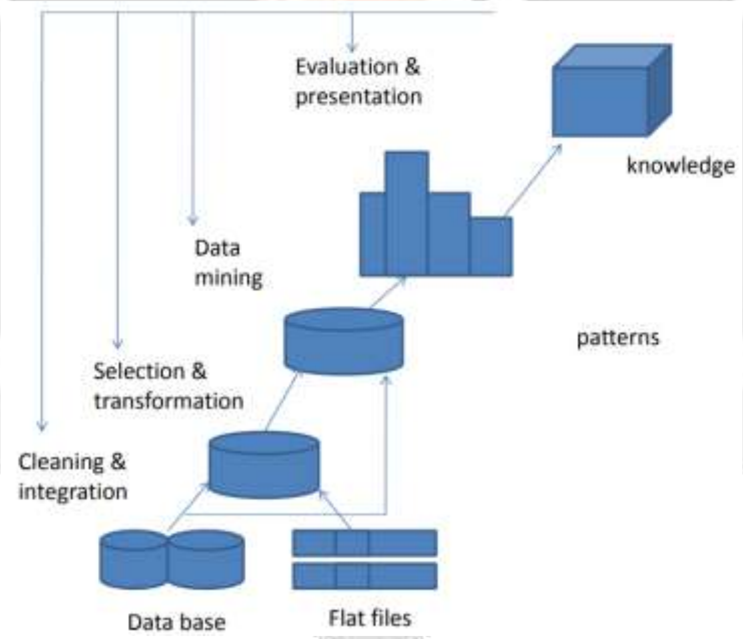

Figure 1: Knowledge Discovery in Database (KDD)

Data Mining consists of four classes of tasks [2].

1) Clustering: Clustering is the automatic learning technique in which division of the data elements into groups of similar objects takes place.

2) Classification: It is the supervised learning technique which is used to map the data into predefined classes.

3) Regression: It is the statistical technique which is used to develop a mathematical formula (like mathematical equations) that fits the dataset.

4) Association Rule Mining: It is the data mining technique which is used to identify relationships from a set of items in a database [1].

2. CLUSTER

A large dataset divides data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. Data clustering has its roots in a number of areas; including data mining, machine learning, biology, and statistics. Traditional clustering algorithms can be classified into two main categories: hierarchical and partitioned [2].

1 Principles of Clustering

The formed clusters need to follow and satisfy the following principles of clustering.

1) Homogeneity: elements of the same cluster are maximally close to each other.

2) Separation: data elements in separate clusters are maximally far apart from each other.

A superior clustering method will create high quality clusters with high intra-cluster similarity and low inter-cluster similarity. The quality of a result produced by clustering depends on both the similarity measure used by method and its implementation. The quality of a cluster produced by clustering method is also measured by its ability to discover some or all of the hidden patterns [1].

3. CLUSTERING METHODS

Clustering methods can be classified into the following categories –

- Partitioning Method
- Hierarchical Method
- Density-based Method
- Grid-Based Method
- Model-Based Method
- Constraint-based Method

1. Partitioning Method

Suppose to given a database of ‘n’ objects and the partitioning method constructs ‘k’ partition of data. Each partition will represent a cluster and k ≤ n. It means that it will classify the data into k groups, which satisfy the following requirements –

- Each group contains at least one object.
- Each object must belong to exactly one group.

For a given number of partitions (say k). The partitioning method will create an initial partitioning and it uses the iterative relocation technique to improve the partitioning by moving objects from one group to other [10]. In partitioning clustering method clustering creates the clusters in one step instead of creating several steps. Only one set of clusters is formed at the end of clustering, although several sets of clusters may be created internally. As we know that only one set of clusters will be formed then user must have to specify the input(the desired number of clusters). The most well-known and commonly used partitioning methods are k-means, k- medoids.

i). k-means method or centroid based method: The k-means method takes the input parameter,k, and partitions a set of n objects into k clusters so that the resulting intracluster similarity is high but the intercluster similarity is low.

"How k-means methods work?" The k- means method work as follows. Randomly k objects are selected; each object represents a cluster mean or center. object which is most similar or close to cluster mean based on the distance between the object and the cluster is assigned to the cluster .This process will remain continue until the criterion function meets.

ii). FCM - Fuzzy CMEANS algorithm: The algorithm is based on the K-means concept to partition dataset into Clusters. The algorithm is as follows: Calculate the cluster centroids and the objective value and initialize fuzzy matrix. Computer the membership values stored in the matrix. The paper presents list of all algorithms and their efficiency based on the input parameter to mine the Big Data as described below: If the value of objective is between consecutive iterations is less than the stopping condition then stop. This process is continuous until a partition matrix and clusters are formed [7].

iii). k-medoid method: A reference point or mean value of the cluster, we choose actual objects to represent the clusters, i.e one object per cluster. Each leftover object is clustered with the chosen object to which it is most similar. Then performed the partitioning method based on the principal of minimizing the sum of dissimilarities between each object and its corresponding reference point or mean value.

iv). CLARANS (Clustering Large Application Based upon Randomized Search) is partitioning method used for large database. Combination of Sampling technique and PAM is used in CLARANS. In CLARANS we draw random sample of neighbours in each step of search dynamically. CLARANS doesn‟t guaranteed search to localized area. The minimum distance between Neighbours nodes increase efficiency of the algorithm. Computation complexity of this algorithm is O (n^2) [1].

2. Hierarchical Methods

This method creates a hierarchical decomposition of the given set of data objects. It can classify hierarchical methods on the basis of how the hierarchical decomposition is formed. There are two approaches here –

- Agglomerative Approach

- Divisive Approach

Agglomerative Approach: This approach is also known as the bottom-up approach. In this, method starts with each object forming a separate group. It keeps on merging the objects or groups that are close to one another. It keep on doing so until all of the groups are merged into one or until the termination condition holds.

Divisive Approach: This approach is also known as the top-down approach. In this, we start with all of the objects in the same cluster. In the continuous iteration, a cluster is split up into smaller clusters. It is down until each object in one cluster or the termination condition holds. This method is rigid, i.e., once a merging or splitting is done, it can never be undone. Approaches toImprove Quality of Hierarchical Clustering, Here are the two approaches that are used to improve the quality of hierarchical clustering

- Perform careful analysis of object linkages at each hierarchical partitioning.

- Integrate hierarchical agglomeration by first using a hierarchical agglomerative algorithm to group objects into micro-clusters, and then performing macro-clustering on the micro-clusters[10].

3. Density-Based Method

This method is based on the notion of density. The basic idea is to continue growing the given cluster as long as the density in the neighborhood exceeds some threshold, i.e., for each data point within a given cluster, the radius of a given cluster has to contain at least a minimum number of points[10]. Density Based method can be classified into three parts that are as follows:

i). DBSCAN: density based clustering method based on connected regions It is density based clustering method for handling spatial data with noise in application or database. It uses the high density region for making the cluster, and the other regions which have low density are kept outside the cluster by marking as outlier. There is no need to define the number of clusters in advanced. By using the “Minpt” parameter it is able to find out the cluster which is totally different. “Density reachability” and “Density connectability” are the two concepts which are used during making the cluster which in turn have asymmetric and symmetric relation. “Minpt” and “ e “are the two parameters ,if point k contains more “Minp”t than the e-neighborhood then a new cluster with core object will be created, then the DBSCAN will gather the density reachable object from these core objects. When there are no new points that can be further added into the cluster than the DBSCAN process is turned off.

ii). OPTICS: ordering point to identify the clustering structure Optics creates the liner ordering of objects in the database. Like the DBSCAN it use two parameter “e “and “Minpt” where e define the maximum distance and “Minpt” define the number of points or objects required to make a cluster. For making clustering automatic and itrative augment ordering of objects in the database is created. Core distance and Nagar,Reachability distance are needed define to ordering of objects in to the database. It is similar to DBSCAN but overcome one of the major weakness i.e density meaningful cluster in data of varying density.

iii). DENCLUE: (Clustering based on density distribution functions) DENCLUE use the density distribution function for making the clusters. It use the influence function which wedges the data point along with its neighborhood points. The points are arranged in the hill climbing manner where the points having the same local maximum are placed together into the cluster. But this hill climbing can create some error or problem as it may never coincide exactly to the maximum, just come close. DENCLUE have strong mathematical foundation and good properties which perform the arbitrarily shaped cluster in high dimensional data set with large amount of noise. Grid cell are used to maintain the data points information in tree like structure for faster performance [4].

4. Grid-Based Method

This method, the objects together form a grid and object space is quantized into finite number of cells that form a grid structure. The main pro is fast processing time and Its dependent only on the number of cells in each dimension in the quantized space [10]. There are two types of Grid based method as follows:

i. STINGS: statistical information grid STINGS break the whole spatial area into rectangular cells. These rectangular cell promote tree like structure which give in return to other different level of resolution. Every cell is break into other cells at a high level to make the next lower level. This algorithm assumes that a query can be answered from the stored statistical information which is reciprocated in the tree. The upper part of the tree consists the entire space and the lower area or level have one leaf for each smallest cells. In this algorithm only vertical and horizontal boundaries are built. Scanning is done one time and all the parameters like, mean, variance, distribution are determined for each cell which makes it more efficient. Due to its grid like structure it perform incremental and parallel processing. Quality of clustering only depends on the granularity of the lowest level of the grid if lowest level is brutish then quality will decrease.

ii. WAVECLUSTER: clustering using wavelet Transformation In this approach every grid cell encapsulate the information of points that is mapped into the cell. This pruned knowledge /information is then applied into the multiresolution wavelet transform for the cluster analysis. This multiresolution property helps in recognizing the varying level of accuracy. The relative distance between the points at different resolution is reciprocated into more distinguishable form for preservation by transforming the data through the wavelet transform. It uses the filters to find the frequency of signal or regions and automatically remove the outliers [4].

5. Model-Based Methods

In this method, a model is hypothesized for each cluster to find the best fit of data for a given model and locates the clusters by clustering the density function. It reflects spatial distribution of the data points and also provides a way to automatically determine the number of clusters based on standard statistics, taking outlier or noise into account to yields robust clustering methods [10]. It contains two method which are as follows:

i). EXPECTATION-MAXIMIZATION: EM is the most preferred iterative refinement method that is used to figure out the parameter estimates. Each cluster is defined by parametric probability distribution. Objects are assigned to cluster according to their mean value with some weight associated with objects. EM start with initial assumption of the parameter vector which is randomly choose on the basis of clusters mean value and then the expectation step and maximization step are applied for the distribution of the given data. EM is simple and easy to implement.

ii). CONCEPTUAL METHOD: Conceptual method is a unsupervised machine learning method for the classification of unknown classification. Concept based structure is used to separate the generated classes from the ordinary data. This concept based method is similar to decision tree in which a hierarchy is generated. Various conceptual clustering method have been proposed like COBWEB, WITT, GCF, GALOSS, CYRUS etc. Among all these methods COBWEB is the most prevailing method, which is simple and incremental approach. Categorical attribute values are used to define the objects and these objects are enact by the binary values in a hierarchy manner. COBWEB automatically adjust the number of classes in partition . Merging and splitting parameters makes the COBWEB less sensitive for input order but it is not scalable for the large data bases.

6. Constraint-Based Method

In this method, the clustering is performed by the merging of user or application-oriented constraints. A constraint refers to the user expectation or the properties of desired clustering results. Constraints provide us with an interactive way of communication with the clustering process. Constraints can be specified by the user or the application requirement [10].

4. CONCLUSION

The main study of applying PCA on original data before clustering is to obtain accurate results. But the clustering results depend on the initialization of centroid. In this paper, we have proposed a new approach to initialize the centroid and reducing the dimension using principal component analysis to improve the accuracy of the cluster results and the standard k-means algorithm also modified to improve the efficiency by reducing the computation complexity of the algorithm. The experiment results show that the substantial improvement in running time and accuracy of the clustering results by reducing the dimension and initial centroid selection using PCA. Though the proposed method gave better quality results in all cases, over random initialization methods, still there is a limitation associated with this, i.e. the number of clusters (k) is required to be given as input. Evolving some statistical methods to compute the value of k, depending on the data distribution is suggested for future research. In the future, we plan to apply this method to microarray cancer datasets.

5. REFERENCES

[1] Adam schenker, mark last, horst bunke, Abraham kandel (2003): Comparison of two noval algorithm for clustering using web documents, WDA.
[2] Arthur D., vassilvitskii S.(2007): K-means++ the advantages of careful seeding, on discrete algorithms (SODA).
[3] Babu G. and Murty M. (1993): A near Optimal initial seed value selection in k-means algorithm using a genetic algorithm, Pattern Recognition Letters Vol.14,1993, PP, 763-769.
[4] Chris Ding and Xiaofeng He (2004) : k-means Clustering via Principal component Analysis, In Proceedings of the 21st international conference on Machine Learning, Banff, Canada.
[5] Deelers S. S. and Auwatanamongkol S. (2007): Enhancing K-Means Algorithm with initial cluster centers Derived from Data partitioning along the Data axis with the highest Variance, Proceedings of world Academy of Science, Engineering and Technology Volume 26, ISSN 1307-6884.
[6] Fahim A.M,Salem A.M, Torkey A and Ramadan M.A (2006) : An Efficient enchanced k-means clustering algorithm,Journal of Zhejiang University,10(7): 1626-1633,2006.
[7] Fahim A.M,Salem A.M, Torkey F. A., Saake G and Ramadan M.A (2009): An Efficient k-means with good initial starting points, Georgian Electronic Scientific Journal: Computer Science and Telecommunications, Vol.2, No. 19,pp. 47-57.
[8] Huang Z (1998): Extensions to the k-means algorithm for clustering large data sets with categorical values, Data Mining and Knowledge Discovery,(2):283-304.
[9] Ismail M. and Kamal M. (1989): Multidimensional data clustering utilization hybrid search strategies,Pattern Recognition Vol. 22(1),PP. 75-89.
[10] Jiawei Han M.K (2006): Data mining Concepts and Techniques, morgan Kaufmann publishers, An imprint of Elsevier.
[11] Jolliffe I.T. (2002): Principal Component Analysis, Springer, Second edition.