Research Article

Survey performance approach k-Mean and k-Mediod clustering algorithm

Anjani Pandey¹, Mahima Shukla²

*Corresponding author:

Anjani Pandey

¹Dept of Computer Engineering V.I.T.S, Satna (M.P)
²Dept of CSE (Software System) V.I.T.S, Satna (M.P)

Abstract

K-means clustering algorithms are widely used for many practical applications. Original k-mean algorithm select initial centroids and medoids randomly that affect the quality of the resulting clusters and sometimes it generates unstable and empty clusters which are meaningless. The original k-means algorithm is computationally expensive and requires time proportional to the product of the number of data items, number of clusters and the number of iterations.

The new approach for the k-mean algorithm eliminates the deficiency of exiting k mean. It first calculates the initial centroids k as per requirements of users and then gives better, effective and good cluster without sacrificing Accuracy. It generates stable clusters to improve accuracy. It also reduces the mean square error and improves the quality of clustering. We also applied our algorithm for the evaluation of student’s academic performance for the purpose of making effective decision by the student councilors.

Keywords: Cluster analysis, Centroids, K-mean, k-mediod

Introduction

Unsupervised learning is the part of machine learning whose purpose is to give the ability to machine to find some hidden structure within data. Typical task in unsupervised learning include the discovery of “natural” clusters present in the data, finding a meaningful low dimensional representation of the data or learning explicitly a probability function that represents the true distribution of the data. The clustering problem is classical problem of database, knowledge discovery, artificial intelligence and theoretical literature is use to find similar groups of record from very large datasets [6]. Given a training data set, the goal of a clustering algorithm is to group similar data points in the same cluster while putting dissimilar data points in different clusters. Clustering is used in a wide variety of fields: biology, statistics, pattern recognition, information retrieval, machine learning, psychology, and data mining. For example, it is used to group related documents for browsing, to find genes and proteins that have similar functionality, to find the similarity in medical image database, or as a means of data compression. Clustering is an important branch of pattern recognition, and it aims at modeling fuzzy (i.e., ambiguous) unlabeled pattern efficiently [1].

There are a number of clustering methods which can be classified into following categories: Partitioning methods, Hierarchical methods, Density-based methods, Grid-based methods, Model-based methods [10]. Each of these methods handles some issues related to clustering but, there is not a single universal clustering algorithm that can handle all the issues related to it [9]. With regard to the problem of partitioning N objects into k classes, to get the best clustering is a NP-hard problem. It is a well-known fact that the standard k-means algorithm gets easily trapped in a local minimum.

Procedure of Cluster Analysis

Cluster analysis is mainly divided into four basic steps as shown in Figure:1[3].

Feature Selection or Extraction

Feature selection is the process of discovering the most relevant attribute of a dataset to the data mining task. It is commonly used and powerful technique for reduction the dimensionality of a problem to more manageable task. Feature extraction utilizes
some transformations to generate useful and novel features from original ones. It does not remove any of the original attribute from further consideration. This technique is best suited to dataset where most of the dimensions are relevant to the clustering task, but may are highly correlated or redundant. Generally the ideal features should be of use in distinguishing patterns belonging to different clusters, immune to noise, easy to extract and interpret [2].

Clustering Algorithm design and selection

In this step, the proximity (similarity or dissimilarity) measure and criterion function is selected. Proximity measure greatly affects the resulting clusters. Almost all clustering algorithm are explicitly or implicitly connected to some definition of proximity measure. Once the proximity measure is chosen, the criterion function is selected in order to optimize clustering problem, which is well defined mathematically (e.g. square error function). There are lots of clustering algorithms has been developed to solve different issues related to clustering in variety of fields, but there is no clustering algorithm that can be universally used to solve all problems. Therefore, it is important to carefully select and design the clustering algorithm which satisfies the characteristics of the specified problem.

Cluster Validation

It is difficult to identify that whether the clusters generated are of meaningful or just an artifact of an algorithm. Each clustering algorithm divides the given dataset into number of partition, without worrying about whether there exists any structure or not. Moreover, different clustering algorithm generates different result for the same dataset, and even some algorithm generates different result for different result for different set of parameters or different order of input data. Therefore there must be some evaluation standards and criteria to provide the user with the degree of confidence for the clustering results derived from the used algorithm.

There are three methods of validating criteria: [5]

**External indices:** based on prior knowledge and used as a standard to validate clustering solutions.

**Internal indices:** independent or prior knowledge. They examine the clustering structure directly from the original data.

**Relative criteria:** compares different clustering structure to decide which one may best reveal the characteristics of the objects.

Result Interpretation

The goal of the clustering algorithm is to extract the important hidden information from the original dataset and to provide user with meaningful insights. The result should be easily interpretable and usable by the user.

The above Figure: 1 shows the feedback pathway, because it is possible that clustering algorithm may iterate for several times to find the optimal solution, or to find optimal value of parameters or select appropriate features.

Review of Existing K-Mean Clustering

Distance Calculation

The distance between two points is taken as a common metric to assess the similarity among the components of a population. The most commonly used distance measure is the Euclidean metric which defines the distance between two points \( p = (p_1, p_2, \ldots) \) and \( q = (q_1, q_2, \ldots) \) as

\[
d = \sqrt{\sum (p_i - q_i)^2}
\]  

(1)

Cluster Seed

First document or object of a cluster is defined as the initiator of that cluster i.e. every incoming object’s similarity is compared with the initiator. The initiator is called the cluster seed.

Existing K-mean

K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume \( k \) clusters) fixed a priori. The main idea is to define \( k \) centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result [2]. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group age is done. At this point, this method needs to re-calculate \( k \) new canroidcs as bar centers of the clusters resulting from the previous step. After these \( k \) new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the \( k \) centroids change their location step by step until no more changes are done. In other words centroids do not move any more. Finally, this algorithm aims at minimizing an objective function, in this case a squared error function [9].

Existing K-medoid

The \( k \)-medoids algorithm is a clustering algorithm related to the \( k \)-means algorithm and the medoidshf algorithm. Both the \( k \)-means and \( k \)-medoids algorithms are partitional (breaking the dataset up into groups) and both attempt to minimize the distance between points labeled to be in a cluster and a point designated as the center of that cluster. In contrast to the \( k \)-means algorithm, \( k \)-medoids chooses data points as centers (medoids or exemplars) and works with an arbitrary matrix of distances between data points instead of \( \ell_2 \). This method was proposed in 1987 for the work with \( \ell_1 \) norm and other distances. It is more robust to noise and outliers.
as compared to \(k\)-means because it minimizes a sum of pair wise dissimilarities instead of a sum of squared Euclidean distances. A medoid can be defined as the object of a cluster, whose average dissimilarity to all the objects in the cluster is minimal i.e. it is a most centrally located point in the cluster. The most common realisation of \(k\)-medoid clustering is the Partitioning Around Medoids (PAM) algorithm.

### Conclusion

A New \(k\)-mean and \(k\) mediod algorithm which in new Approach of classical partition based clustering algorithm to be analysis improve both in the execution time of \(k\)-means and \(k\) mediod algorithm, with no miss of clustering quality in most cases. From our result we conclude that, the second proposed implementation of the \(k\)-means and \(k\) medio algorithm is the best one. From experiment we observe that proposed algorithm give more accuracy for dense dataset.

### References


[7]. MacQueen J. “Some method for classification and analysis of multi variate observation”, University of California, Los Angeles, 281 – 297.


[14]. Zhexue Huang. “A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining”.


[17]. Wei-Yin loh. “Regression trees with unbiased variable selection and interaction detection”, University of Wisconsin–Madison.