Data Analytics and Artificial Intelligence
Vol: 2(5), 2022
REST Publisher; ISBN: 978-81-948459-4-2
Website: http://restpublisher.com/book-series/data-analytics-and-artificial-intelligence/

A Study of Clustering Algorithmin Data Science

[*]R.AntoPriyadharshini, B.Nithya
Gonzaga College of Arts and Science for Women,Kathampallam, Elathagiri, Tamil Nadu, India.
[*]Corresponding author Email:priyamic25@gmail.com

Abstract. Data clustering is a process of partitioning data points into meaningful clusters such that a cluster holds similar data and different cluster hold dissimilar data. It is an unsupervised approach to classify into the following two categories: Firstly, hard clustering, where a data object can belong to a single and distinct cluster and secondly, soft clustering, where a data object can belong to different cluster. In this report we have made a comparative study of three major data clustering algorithm highlighting their merits and demerits. These algorithms are: k-means, fuzzy c-means and K-NN clustering algorithm. Choosing an appropriate clustering algorithm for grouping the data takes various factors into account for illustration one is the size of data to be partitioned.

1. Introduction

Many companies have recognized the strategic together by some set of customers (clustering in importance of the knowledge hidden in their database a data warehouse storing sales transaction). and, therefore, have built data warehouse. A data · Symptoms distinguishing disease A form B warehouse is a collection of data from multiple sources, (classification in a medical data warehouse) integrated into a common repository and extended by · Description of the typical WWW access summary information for a purpose of analysis. There are two categories of data warehousing. patterns (summarization in the data warehouse of an internet provider). Derived Information is present for the purpose of analysis. The environment is dynamic.

2. Clustering Algorithm in Datamining

In such an environment, either manual analysis Cluster analysis or clustering is the task of grouping set of objects in such a way that object in the same group (called as cluster) are more similar to each other supported by appropriated visualization tools or (semi) than to those in other group (cluster). It is main task of automatic data mining may be performed. Data mining exploratory data mining, and a common technique for has been defined as the application of data analysis and statistical data analysis, used in many fields, including discovery algorithm that – under acceptable machine learning, pattern recognition, image analysis, computational efficiency limitations-produce a particular information retrieval, bioinformatics, data compression, enumeration of patterns over the data. Several data and computer graphics. mining tasks have been identified, e.g., clustering, classification and summarization. Typical results of data the result of a cluster analysis shown as the coloring of mining are as follows: a square into three cluster. A clustering is essentially a set of such cluster, usually containing all object in the data set. Additionally, it may specify the relationship of the cluster to each other, for example, a hierarchy of figure 3: Single –linkage on density-based cluster. Hard clustering: Each object belongs to a 20 clusters extracted, most of which contain single Cluster embedded in each other. Clustering's can be roughly distinguished as: elements, and since linkage clustering does not have notion of "noise".

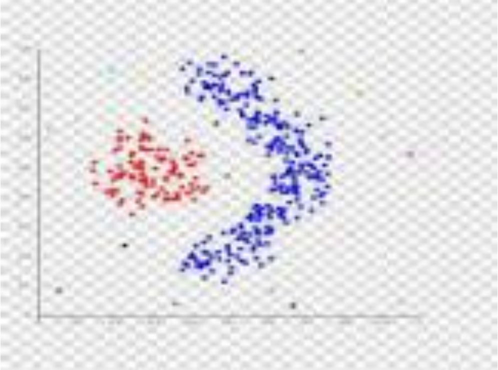
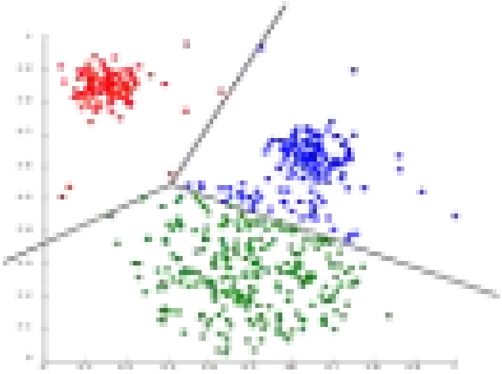


FIGURE 1.the coloring of the squares into three cluster.

3. Centroid-based clustering

Soft clustering (fuzzy clustering): Each object Centroid-based clustering, clusters are represented by a central vector, which may not belong to each cluster to a certain degree. Algorithms necessarily be a member of a data set. When the number of clusters is fixed to k, k-means clustering gives a Clustering algorithm can be categorized based on their formal definition as an optimization problem: find the cluster model, as list above. The following overview will cluster center and assign the objects to the earnest cluster, only list the most prominent examples of clustering such that the squared instance from the cluster are algorithms, as there are possibly over 100 published clustering algorithms minimized. Connectivity based clustering (Hierarchical clustering) K-means has a number of interesting theoretical properties. First, it partitions the data space in to a Connectivity based clustering, also known as Structure known as a Voronoi diagram. Second its hierarchical clustering, is based on the core idea of conceptually close to a nearest neighbor classification, object being more related to nearby object further away. and as such is popular in machine learning. Third, it can These algorithms connect “object” to form “cluster” be seen as variation of model-based classification, and based on their distance. In a dendrogram the y-axis Lloyd’s algorithm as a variation of the expectation- marks the distance at which the cluster merge, while the maximization algorithm for this model discussed below. object is placed along the x-axis such that the clusters don’t mix. At 35 clusters, the K-means separate data into Voronoi-cell, which assumes biggest cluster starts fragmenting into small parts, while equal-sized cluster(not adequate here) before it was still connected to the second largest due to the single-link effect.



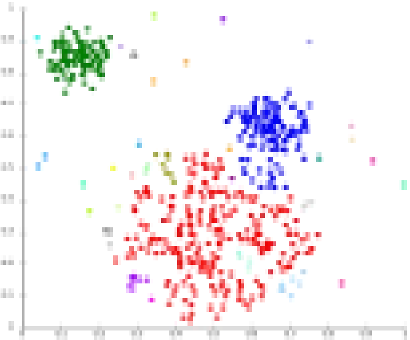
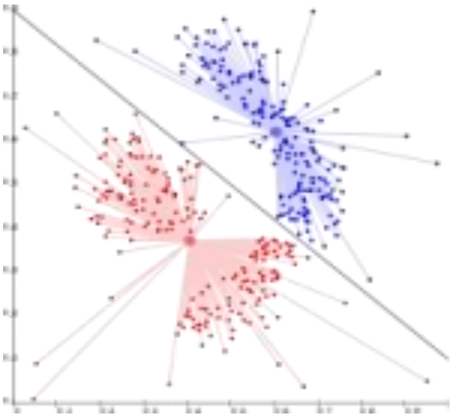


FIGURE 2. Linkage Clustering; **FIGURE 3.** K-means clustering Single-linkage on Gaussian data.



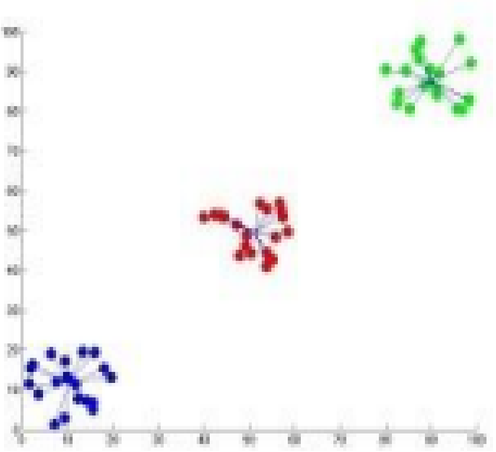


FIGURE 4. K-means cannot represent density-based; **FIGURE 5.** Showing the result of k-means for 'N' =60

1) Randomly select 'c' cluster centers. 2) Calculate the distance between each data point and cluster centers.3) Assign the data point to the cluster center whose distance from the cluster center is minimumof all the cluster centers. 4) Recalculate the new cluster center using: clusters. Where, 'ci' represents the number of data pointsK-means is one of the simplest 5) Recalculate the distance between each data point and unsupervised earning algorithms that solve the well- new obtained cluster centers. known clustering problem. The procedure follows a simple and easy way to classify a given data set through 6) If no data point was reassigned then stop, otherwise a certain number of clusters (assume k clusters) fixed repeat from step (3). apriority. The main idea is to define k centers, one for each cluster. These centers should be placed n a Merits cunning way because of different location causes different result. So, the better choice is to lace 1) Fast, robust and easier to understand. 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 center. When to 2) Relatively efficient: O(tknd), where n is # objects, kispoint is pending, the first step is completed and an # clusters, d is # dimension of each object, and t is # early group age is done. At this point we eed to re- iterations. Normally, k, t, d << n. calculates k new centroids as barycenter of the clusters resulting from the previous step. After we have these k 3) Gives best result when data set are distinct or well new centroids, a new binding has to be separated from each other. done between the same data set points and the nearest new center. A loop has been generated. As a result.

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i \quad J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (\|x_i - v_j\|)^2$$

the following ways. of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more. Finally, this algorithm aims at minimizing an objective function know as squared error function given by: where, '||xi- vj||' is the Euclidean distance and 'c' = 3 between xi and vj. Note: For more detailed figure for k-means algorithm please refer to k-means figure subpage. 'ci' is the number of data points in ith cluster. 'c' is the number of cluster centers. Demerits 1) The learning algorithm requires apriorism pacification the number of cluster centers. Algorithmic steps for k-means clustering 2) The use of Exclusive Assignment - If there are two Let X = x1, x2, x3,…..,xn} be the set of data points and V = {v1,v2,…….,vc} be the set of centers. highly overlapping data then k-means will not be able to resolve that there are two clusters. 3) The learning algorithm is not invariant to non-linear Three-Tier data Warehouse Architecture transformations i.e.; with different representation of data, we get different results (data represented in form of Generally a data warehouse adopts a three-tier cartesian co-ordinates and polar co-ordinates will give architecture. Following is the three tiers of the data different results). warehouse architecture. · Bottom Tier- The bottom tier of the architecture4) Euclidean distance measures can unequally weight underlying factors. 5) The learning algorithm provides the local optima of the squared error function. is the data warehouse database server. It is the relational database system. We use the backend tools and utilities to feed data into the bottom tier. These back-end tools and utilities perform the extract, Clean, Load, and refresh functions. · Middle Tier – In the middle tier, we have the OLAP Server that can be implemented either of 6) Randomly choosing of the cluster center cannot lead us to the fruitful result. Pl. refer By Relational OLAP (ROLAP), which is an extended. 7) Applicable only when mean is defined i.e., fails for categorical data. 8) Unable to handle noisy data and outliers. 9) Algorithm fails for non-linear data set. relational database management system. The roadmaps the operations on multidimensional data to standard relational operations. By Multidimensional OLAP (MOLAP) model, which directly implements the multidimensional data and operations. · Top-Tier – This tier is the front-end client layer. This layer holds the query tools and reporting tools, analysis tools and data mining tools.

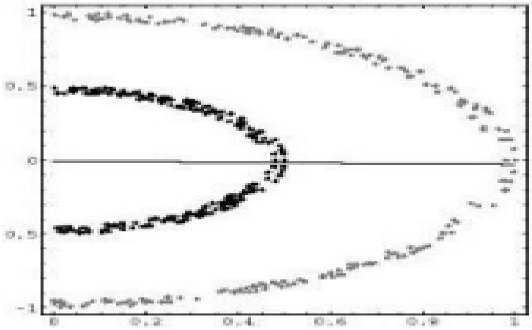


FIGURE 6. Three tier Architecture of data warehouse

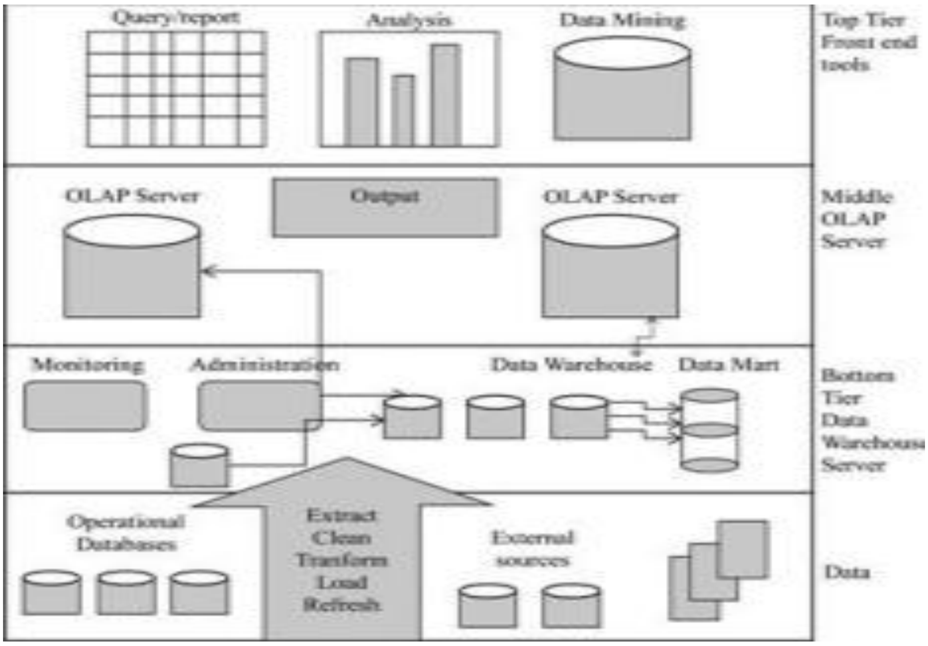


FIGURE 7. Showing the non-linear data set where k- means algorithm fails

4. Clustering Algorithm in Data Warehousing

Data Warehouse The business analyst gets the information from the warehouse to measure the performance and make critical adjustments in order to win over other business holders in the market. Having a data warehouse offers the following advantages: · Since a data warehouse can gather information Figure 1: Three tier Architecture of data warehouse quickly and efficiently, it can enhance business productivity. · A data warehouse provides us a consistent Fuzzy C-Means Clustering view of customers and items, hence, it helps us manage customer relationship. The Algorithm Fuzzy c-means (FCM) is a methodic · A data warehouse also helps in bringing down clustering which allows one piece of data to elongto the costs by tracking trends, patterns over a two or more clusters. This method (developed by Dunn long period in a consistent and reliable manner. in 1973 and improved by Bezdek in 1981) is frequently used in pattern recognition. It is based on minimization mono-dimensional example. Given a certain data set, of the following objective function supposed to represent it as distributed on an axis. The figure below shows this: Looking at the picture, we may identify two clusters in where m is any real number greater than 1, uij is the proximity of the two data concentrations. We will refer degree of membership of xi in the cluster j, xi is the ith of to them using 'A' and 'B'. In the first approach showman d-dimensional measured data, cj is the d-dimension this tutorial - the k-means algorithm - we associated each center of the cluster, and ||*|| is any norm expressing the datum to a specific centroid; therefore, this membership similarity between any measured data and the center. function looked like this: Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership uij and the cluster centers cj by: This iteration will stop when , where is a termination criterion between 0 and 1, whereas k are the iteration steps. This procedure converges to a local minimum or a saddle point of Jm. The algorithm is composed of the following steps: In the FCM approach, instead, the same given datum does not belong exclusively to a well-defined cluster, but 1. Initialize U=[uij] matrix, U(0) 2. At k-step: calculate the centers C(k)=[cj]

with(k) 3. Update U(k) U(k+1) 4. If || U(k+1) - U(k)||< then STOP; other wisereturn to step 2. it can be placed in a middle way. In this case, the membership function follows a smoother line to indicate vectors U Remarks that every datum may belong to several clusters with different values of the membership coefficient. In the figure above, the datum shown as a red marked spot belongs more to the B cluster rather than the Acluster. The value 0.2 of ‘m’ indicates the degree of As already told, data are bound to each cluster by means membership to A for such datum. Now, instead of using of a Membership Function, which represents the fuzzy a graphical representation, we introduce a matrix behavior of this algorithm. To do that, we simply have to whose factors are the ones taken fromthe membership build an appropriate matrix named U whose factors are functions: numbers between 0 and 1, and represent the degree of membership between data and centers of clusters. For a better understanding, we may consider this simple performed yet so that clusters are not identified very well. Now we can run the algorithm until the stop conditions verified. The figure below shows the final condition reached at the 8th step with m=2 and =0.3: (a) (b) The number of rows and columns depends on how many data and clusters we are considering. More exactly we have C = 2 columns (C = 2 clusters) and N rows, where C is the total number of clusters and N is the total number of data. The generic element is so indicated: uij. In the examples above we have considered the k-means (a) and FCM (b) cases. We can notice that in the first case (a) the coefficients are always unitary. It is so to indicate the fact that each datum can belong only to one cluster. Other properties are shown below: Is it possible to do better? Certainly, we could use An Example higher accuracy but we would have also to pay for abigger computational effort. In the next figure we cansee a better result having used the same initial conditions and =0.01, but we needed 37 steps! Here, we consider the simple case of a mono- dimensional application of the FCM. Twenty data and three clusters are used to initialize the algorithm and to compute the U matrix. Figures below (taken from our interactive demo) show the membership value for each datum and for each cluster. The color of the data is that of the nearest cluster according to the membership function. It is also important to notice that different initializations cause different evolutions of the algorithm. In fact, it could converge to the same result but probably with a different number of iteration steps.

5. Conclusion

In this paper the authors have reviewed the three major clustering algorithms, namely, the k-means, the fuzzy c-means and the KNN clustering algorithm and have made a comparative study taking account their In the simulation shown in the figure above we have used a fuzzy ness coefficient m = 2 and we ave also merits and demerits and all other factors which may imposed to terminate the algorithm influence the criterion in choosing an appropriate when . The picture shows clustering algorithm for partitioning a given dataset. In the initial condition where the fuzzy distribution depends k-means algorithm, the main principle is to minimize the on the particular position of the clusters. No step is sum of squares of distances between data and Ortega, Ma. DelRocio Boone Rojas and Maria J. corresponding clustering centroids. Hence, the distance Somodevilla Garcia. measure is the quintessential criterion in classification

References

[1] An Efficient k-means Clustering Algorithm: Analysis and Implementation by Tapas Kanungo, David M. .Mount,Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman and Angela Y. Wu.

[2] Research issues on K-means Algorithm: An Experimental Trial Using Matlab by Joaquin Perez

[3] The k-means algorithm - Notes by Tan, Steinbach, and formation of clusters. The KNN algorithm uses

[4] Kumar Ghosh. neighborhood classification as the predication value of

[5] http://home.dei.polimi.it/matteucc/Clustering/tutorial_ the new query instant. The fuzzy c means algorithm, on html/kmeans.html the other hand, is a soft clustering algorithm. It uses k-means clustering by kechen. fuzzy sets to cluster data, such that each data point may belong to two or more clusters (overlapping clustering)

[6] J. C. Dunn (1973): "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact with different degrees of membership. The data objectsWell-Separated Clusters", Journal of Cybernetics 3: 32- on the boundaries between several cluster or not forced 57 to fully belong to one of the clusters, but are rather.

[7] J. C. Bezdek (1981): "Pattern Recognition with Fuzzy assigned membership degrees between 0 and 1Objective Function Algoritms", PlenumPress, New indicating their partial membership. Based on the various York constraints and conventions, it was observed that the

[8] Tariq Rashid: “Clustering”fuzzy c-means algorithm is the most widely used.