INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 5, Issue 3)
Available online at: www.ijariit.com

Clustering for the decision making in data and business analytics

Suhani Jain
suhani.jain71@gmail.com
Mody University, Lakshmangarh, Rajasthan

ABSTRACT

The term clustering is very important in the field of data analytics and in building the decision and for reaching to the conclusion to make the decision from the given data set. How clustering is used in the process of decision making with different methods and processes, we will get to learn by this paper. The clustering is used for the prescriptive model by which it will form the decision. The clustering is used for the process of predictive model analysis to the prescriptive model analysis.

Keywords— *Clustering, Predictive, Prescriptive, Datasets, Mining*

1. INTRODUCTION

As we know that in the world of data analytics, there are three types of models on which the data analyst works. These are the descriptive, predictive and prescriptive models. Here we are talking about the clustering that is also known as the mining of the data. The clustering is formed on the predictive and prescriptive models. How the prediction converts into predictive is done and shown by the clustering. Here we will get to know how the clustering works and help us in the prescription of any data set and the cases.

1.1 Data analytics

Data analytics is the process or the technique to manage the datasets in order to draw conclusions about the information they contain, increasingly with the aid of specialized systems and software

1.2 Business analytics

Business analytics is the practice of different iterative, methodical exploration of an organization's data, with an emphasis on statistical analysis. Business analytics is used by huge companies committed to data-driven decision-making.

2. CLUSTERING

Cluster analysis is one in every of those, so-called, data processing tools. These tools are usually a combination of predictive, however since they assist managers to build higher selections, they will even be thought of prescriptive. Cluster analysis refers to a way for distinguishing teams during which the weather of every cluster is similar. Thus essentially cluster is that the technique or the tactic that is employed for the grouping and distinguishing the similar information from the larger information set.

2.1 Motive and importance of the clustering

The approach is often used in marketing in order to understand differences among customers. The main idea of this application of cluster analysis is to create homogeneous market segments to target product offerings. Example of clustering-Let's understand this with an example. Suppose, you are the head of a rental store and wish to understand the preferences of your costumers to scale up your business. Is it possible for you to look at the details of each costumer and devise a unique business strategy for each one of them? Definitely not. But, what you can do is to cluster all of your costumers into say 20 groups based on their purchasing habits and use a separate strategy for costumers in each of these 20 groups. And this is what we call clustering.

2.2 Types of clustering

Broadly speaking, clustering can be divided into two subgroups:

2.2.1 Hard Clustering: In this type of clustering, the data are classified into only one group. For example, in the above example, each customer is put into one group out of the 100 groups.

2.2.2 Soft Clustering: In soft clustering, the data are classified and put into different data sets and the segments, a probability or likelihood of that data point to be in those clusters are assigned. For example, from the above scenario, each customer is assigned a probability to be in either of 100 clusters of the retail store.

2.3 Types of clustering algorithms

2.3.1 Connectivity models: The model itself is giving the meaning that in this the data points are connected to each other and in they are very close to each other. The distance between them is very small. This is called the connectivity models.

2.3.2 Centre of mass models: These square measure repetitious agglomeration algorithms during which the notion of similarity comes by the closeness of an information purpose to the centre of mass of the clusters. K-Means agglomeration algorithm rule may be a well-liked algorithmic rule that falls into this class.

2.3.3 Distribution models: These agglomeration models square measure supported the notion of however probability is it that everyone information points within the cluster belong to constant distribution (For example traditional, Gaussian). These models typically suffer from over fitting. A preferred example of those models is Expectation maximization algorithmic rule that uses variable traditional distributions.

2.4.4. Density Models: These models search the information area for areas of the assorted density of information points within the data area. It isolates numerous completely different density regions and assigns the information points within these regions within the same cluster. Well-liked samples of density models square measure DBSCAN and OPTICS.

2.4 Distance between clusters

2.4.1 Single linkage: In single-link or single linkage hierarchical clustering, the two clusters with the smallest minimum pairwise distance are there or we merge in each step the two clusters whose two closest members have the smallest distance.

2.4.2 Complete linkage: In complete-link or complete linkage hierarchical clustering, there are two clusters with the smallest maximum pairwise distance we merge in each step the two clusters whose merger has the smallest diameter.

2.4.3 Average linkage: it is a relationship between the sensitivity of complete-link clustering to outliers and the tendency of single-link clustering to form long chains that do not correspond to the intuitive notion of clusters as compact, spherical objects.

2.4.4 Average group linkage: In Average linkage clustering, the distance between two clusters is defined as the average of distances between all pairs of objects, where each pair is made up of one object from each group. $D(r,s) = T_{rs} / (N_r * N_s)$ Where T_{rs} is the sum of all pairwise distances between cluster r and cluster s.

2.4.5 Wards method: In statistics, Ward's method is a criterion applied in hierarchical cluster analysis. Ward's minimum variance method is a special case of the objective function approach originally presented by Joe H.Ward, Jr.

2.5 Methods for clustering

2.5.1 K Means clustering: K means is an iterative clustering algorithm that aims to find local maxima in each iteration. This algorithm works in these 5 steps:

(a) Specify the required variety of clusters K: allow us to opt for k=2 for these five information points in 2-D points.
(b) Now assign every information to a cluster: Let's assign 3 points in cluster one shown victimization red color and 2 points in cluster two shown points in gray color.
(c) Work out cluster centre of mass: The centroid of knowledge points within the red cluster is shown victimization NGO and people in gray cluster victimisation gray cross.
(d) Re-assign every purpose to the nearest cluster centre of mass: Note that solely the info purpose at rock bottom is assigned to the red cluster albeit it's nearer to the centre of mass of the gray cluster. Thus, we tend to assign that information into the gray cluster.
(e) Re-compute cluster centroids: currently, re-computing the centroids for each the clusters. Repeat steps 4 and 5 until no improvements are possible
(f) Similarly, we'll repeat the 4^{th} and 5^{th} steps until we'll reach global optima. When there will be no further switching of data points between two clusters for two successive repeats. It will mark the termination of the algorithm if not explicitly mentioned.

Java code

```
// Load some data
Instances data = DataSource.read("data.arff");

// Create the model
SimpleKMeans kMeans = new SimpleKMeans();

// We want three clusters
kMeans.setNumClusters(3);

// Run K-Means
```



```
kMeans.buildClusterer(data);

// Print the centroids
Instances centroids = kMeans.getClusterCentroids();
for (Instance centroid: centroids) {
  System.out.println(centroid);
}

// Print cluster membership for each instance
for (Instance point: data) {
  System.out.println(point + " is in cluster " + kMeans.clusterInstance(point));
}
```

Python (Scikit-learn)

```
>>> from sklearn import cluster, datasets
>>> iris = datasets.load_iris()
>>> X_iris = iris.data
>>> y_iris = iris.target
>>> k_means = cluster.KMeans(n_clusters=3)
>>> k_means.fit(X_iris)
KMeans(copy_x=True, init='k-means++'
 >>> print(k_means.labels_[::10])
[1 1 1 1 1 0 0 0 0 0 2 2 2 2 2]
>>> print(y_iris[::10])
[0 0 0 0 0 1 1 1 1 1 2 2 2 2 2]
```

2.5.2 Hierarchical clustering: Hierarchical clustering, as the name suggests is an algorithm that builds a hierarchy of clusters. This algorithm starts with all the data points assigned to a cluster of their own. Then two nearest clusters are merged into the same cluster. In the end, this algorithm terminates when there is only a single cluster left. The results of hierarchical clustering can be shown using dendrogram. The dendrogram can be interpreted as:

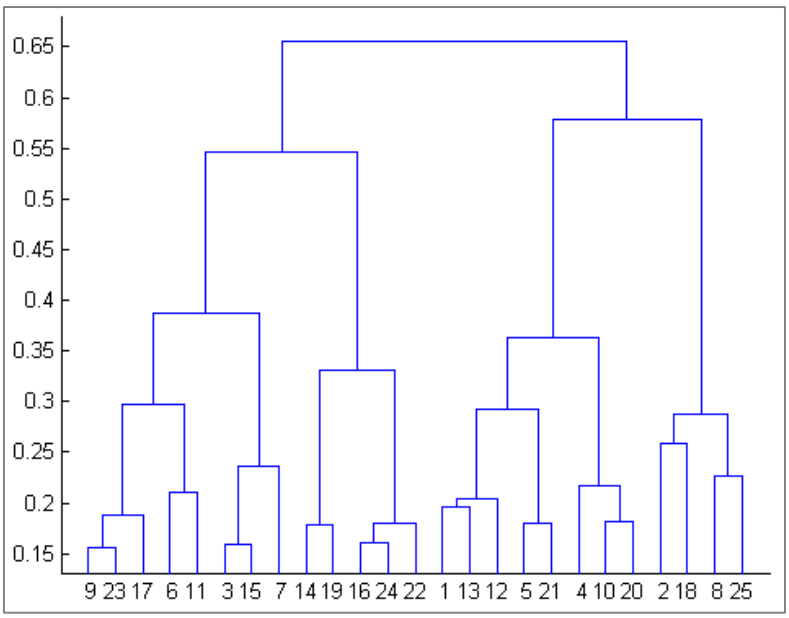

Fig. 1: Hierarchical clustering

At the lowest, we tend to begin with twenty-five knowledge points, every appointed to separate clusters. 2 nearest clusters square measure then incorporate until we've got only 1 cluster at the highest. The peak within the dendrogram at that 2 clusters square measure incorporate represents the between 2 clusters within the knowledge space.

The decision of the no. of clusters which will best depict completely different teams will be chosen by perceptive the dendrogram. the most effective selection of the no. of clusters is that the no. of vertical lines within the dendrogram cut by a horizontal line which will the pairwise the most distance vertically while not deviate a cluster.

In the on top of example, the most effective selection of no. of clusters are going to be four because the red horizontal line within the dendrogram below covers most vertical distance AB.

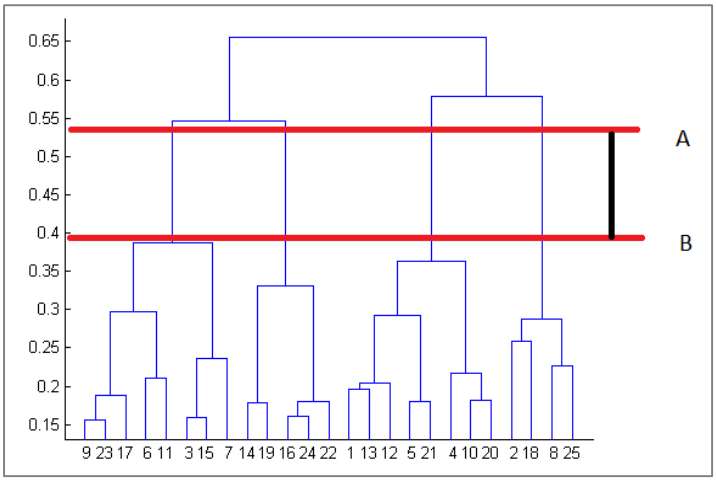

Fig. 2: Hierarchical clustering

Algorithmic steps for Agglomerative Hierarchical clustering

Let X = {x_1, x_2, x_3... x_n} be the set of data points.

Step 1: Begin with the disjoint clustering having level L (0) = 0 and sequence number m = 0.

Step 2: Find the least distance pair of clusters in the current clustering, say pair (r), (s), according to d [(r),(s)] = min d[(i),(j)] where the minimum is over all pairs of clusters in the current clustering.

Step 3: Increment the sequence number: m = m +1.Merge clusters (r) and (s) into a single cluster to form the next clustering m. Set the level of this clustering to L (m) = d [(r), (s)].

Step 4: Update the distance matrix, D, by deleting the rows and columns corresponding to clusters (r) and (s) and adding a row and column corresponding to the newly formed cluster. The distance between the new cluster, denoted (r,s) and old cluster(k) is defined in this way: d[(k), (r,s)] = min (d[(k),(r)], d[(k),(s)]).

Step 5: If all the data points are in one cluster then stop, else repeat from step 2).

Divisive Hierarchical clustering - It is just the reverse of Agglomerative Hierarchical approach.

3. APPLICATIONS OF CLUSTERING

Clustering has a large no. of applications which we see regularly in our daily life. Some of the popular applications of clustering are:

- Market segmentation
- Social network analysis
- Search result grouping
- Recommendation engines
- Medical imaging
- Image segmentation
- Anomaly detection

4. TOOLS USED FOR THE CLUSTERING

4.1 Excel Solver

The solver is an excel tool which is also used for the what-if analysis. The solver is used to find the optimal that is a maximum and minimum value for the formula that is in one cell also called as an objective cell.

4.2 XLMiner

XLMiner is a comprehensive data mining add-in for Excel. Data mining is a discovery-driven data analysis technology used for identifying patterns and relationships in data sets. With overwhelming amounts of data now available from transaction systems and external data sources, organizations are presented with increasing opportunities to understand their data and gain insights into it. Data mining is still an emerging field and is a convergence of fields like statistics, machine learning, and artificial intelligence.

5. CONCLUSION

Here we have studied the clustering formation in the decision making and how the evaluation will be done and made by the different methods, algorithm and formulas. The clustering is one very important concept in data analytics as well as in business analytics. Clustering is used for the prescriptive model and used under supervised learning.

6. REFERENCES

[1] https://www.solver.com/introduction-xlminer
[2] https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/
[3] https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68
[4] https://searchbusinessanalytics.techtarget.com/resources
[5] https://www.researchgate.net/post/Latest_Research_Topics_on_Big_Data_Data_analytics
[6] https://homepages.inf.ed.ac.uk/rbf/BOOKS/JAIN/Clustering_Jain_Dubes.pdf

[7] https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1
[8] https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/
[9] http://www.cad.zju.edu.cn/home/zhx/csmath/lib/exe/fetch.php?media=2011:presentation_ml_by_ibrar.pdf
[10] https://www.toptal.com/machine-learning/clustering-algorithms

