International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT22821

An Analysis of Clustering Techniques

Pradeep Bolleddu
Computer Science Department, Indian Institute of Technology Ropar, India

Article Info
Volume 8, Issue 2
Page Number : 52-57

Publication Issue :
March-April-2022

Article History
Accepted: 01 March 2022
Published: 06 March 2022

ABSTRACT

I analyzed different clustering methods of credit card customers data by some machine learning clustering algorithm i approached to classify the high dimensional data by using clustering algorithms in the real world ,It is a type of process where we divide the similar objects into one group The Objective of paper is to identify which method gives a general view of clustering techniques like K-means, algometrive and gaussian distribution ,to analyze Segmentation of customers can be used to define marketing strategies. Clustering comes under unsupervised learning, if we take data set it contains some input patterns, entities etc. The main aim of clustering is to make partition & to make clusters having similarity, so that we can get some useful insights and make further analyses .The process of finding closely connected and fair information from high dimensional datasets is complex and difficult to carry .In this paper we tried to analyze our model through different clustering models like K-means, DBSCAN, Agglomerative Hierarchical and gaussian mixture model) and I explained about all algorithm that I used unique approaches for clustering the high dimensional data.
Keywords: Machine Learning, Clustering, Data Analysis, Python

I. INTRODUCTION

In Clustering our task is to divide to Ith datapoint into similar groups or homogenous clusters such that the items in one class will be same and another class data points in unique cluster will be different. Several clustering algorithms have come into existence, many researchers and scientists had discovered many clustering techniques in early 20's century, those methods helped many to develop in this modern era clustering we can define, where we will compress the data, a huge quantity of samples will be formed into small subgroups or clusters. We have to identify the classes by using different types of similarity measures on the data. Where similarity measures identify how the clusters are made, these include distance, connection, density, intensity etc. It actually used to determine by using dissimilarity measures such as Euclidean distances between pairs of objects. We came to know that some algorithms are useful for only low dimensional data (traditional clustering) which is nothing but datasets with less No. of attributes. Whereas recent research focuses on many attributes to avoid missing major information related to the data. For high dimensional data not, all attributes are same and mandatory for the process of

clustering, thus grouping objects that are similar to other objects while we consider only a few subsets of relevant attributes by eliminating the irrelevant ones, some traditional methods for clustering high dimensional data these are some methods we can use 1. Dimensionality Reduction, 2. Subspace Clustering,3. Projected Clustering, 4.Hybrid clustering 5.Correlation Clustering ,6.Hypergraph-based Clustering and Grid Based Clustering has become a challenging task to the researchers.

1.1 Some challenges for clustering in high dimension data

In large dataset, it's hard to find clusters or patterns, it's a big task for researchers, the high dimensional data to find meaningful insights from the dataset, it's been complex here i mentioned 2 important reasons
Every dataset will produce equal density of data points, so when n dimensional space increases exponentially it will have the same dimension.

1. A large area should be present to fit all the points present in high dimensional space
2. In a high dimensional space, we bias to every point be closer in a group than to other sample point

II. EDA Process & Data Preprocessing

EDA is one of the crucial Step in machine learning, where we can discover patterns & outliers, so that we can form hypotheses, EDA is complementing to inferential statistics-getting conclusions

1. Firstly, we need to Preview data
2. We need to find all entries and column types in the dataset
3. Later find out all null values, Check and remove the unknown entries
4. By using mat plot library Plot the probability distribution of numeric data like example univariate & pair wise joint probability distribution
5. Again, visualize by plotting the count distribution of categorical data in this dataset daily, monthly and yearly frequencies, are Analyzed the time series of numeric data, by using mat plot library we tried to analyze

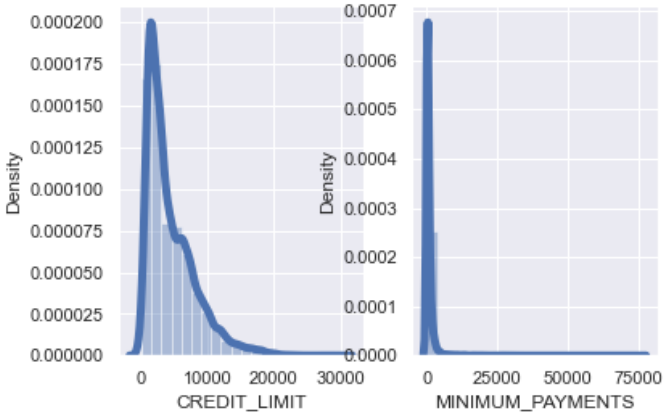

Fig 1 : Visualizing the credit limit and minimum payments

2.2 Outlier detection

We can eliminate outliers by this technique named 'Transforming variables.

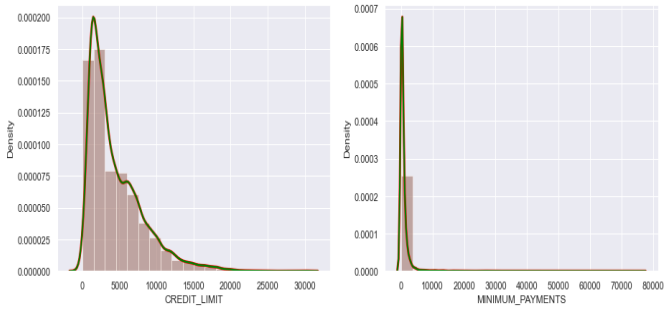

Fig 2 : Relationship between credit limit and minimum payments of customers

2.3 .Finding k through elbow method (for optimal K value)

The function of elbow method it utilizes sum of distances from the ith points & cluster centroid or mean value.

This method is a type of unsupervised algorithm, where we will find the optimal number of data points in which data will be clustered.

Distortion, we can say like Center of individual cluster to average squared distance from the cluster

Inertia : it can be defined as sum of squared distance from respective sample to its center

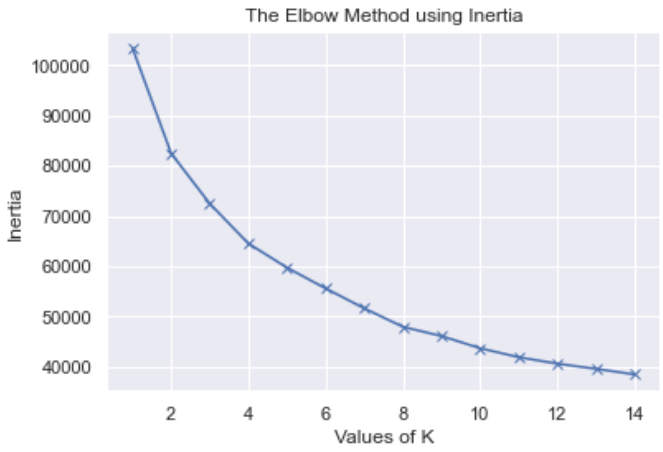

Fig 3 : Elbow method using interia

III. MODEL BUILDING

We can analyze the data and get some useful insights from it , for the purpose of future we can also make predictions by using algorithms

The exam of machine learning model qualify requires accurate selection , criteria

3.1 K-mean Clustering

It's an popular machine learning unsupervised method ,here in the iterative process makes All data points will form into K number of clusters Sum of squared distance from clustered centroid to data point is minimum,so in this position centroid of cluster is mean of data points that present in subclass.

How to implement k means algorithm

This clustering method undergoes through iterative process ,where it partition the dataset into K- data points of non-overlapping clusters ,it actually forms data points into sub group and helps to keep clusters as far as possible ,In this algorithm the main objective is to set data points as cluster ,if the sum squared distance b/w subgroup centroid and data points should be minimum ,here cluster centroid is the mean of data points which present in the subgroups,there will be low variation and we will get same data points within present in the subgroups.

Fig 4 : Clustering using K-means algorithm

3.2 Agglomerative Hierarchical Clustering

It is a one of the hierarchical clustering ,where we can use this method to form subgroups or clusters by analyzing their similarity. It also has another name i.e., AGNES (Agglomerative Nesting). This method is actually ready by taking a single object as a one cluster. Later, its objective is to form a pair of cluster data points which are joined until all clusters have formed into one big cluster which contains all objects. This results in visualization of the objects as tree-based, called dendrogram.

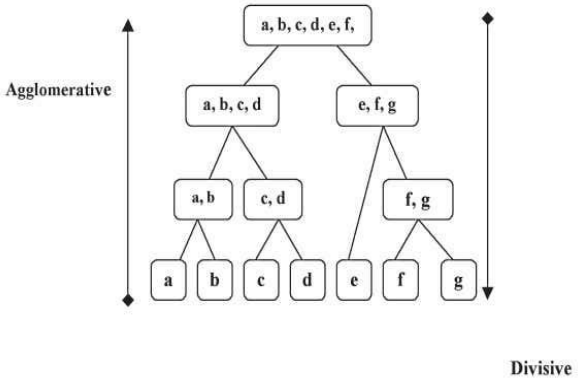

Fig 5 : Representation of dendrogram

I propose following Steps we can use this clustering algorithm

We have to proceed with the following steps below to perform agglomerative hierarchical clustering using python software:

1. Firstly we need to Prepare the data
2. we need to carefully Compute any dis-similarity information between each & every pair of objects present in the data set
3. Later we can Use the linkage function to group our objects into dendrograms. It can be done based on distance information formed at clusters that are very close to areas ,linked to each other so that we can use the linkage function.
4. Finally we have to Determine when and correct location to stop hierarchical trees into subclasses This finally results in the partition of the data.

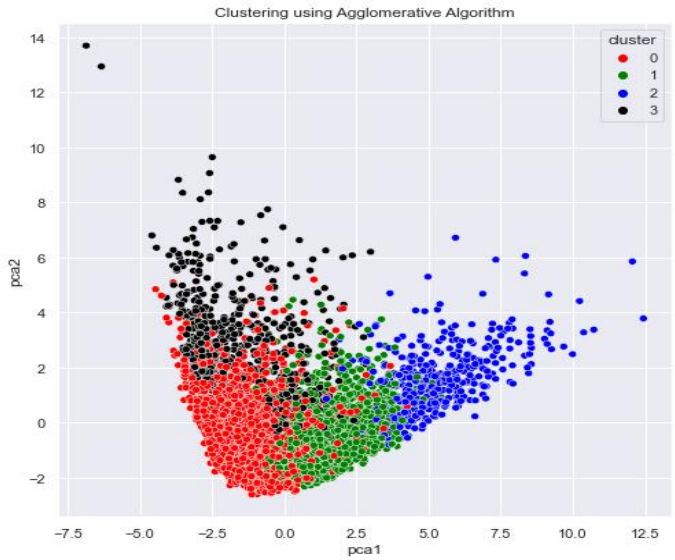

Fig 5 : Clustering using Agglomerative algorithm

3.3 GaussianMixture Model based clustering

These models can be used to form clusters ,where unlabeled data will be equal to k-means. There are, there are a some advantages to use Gaussian mixture model based clustering over k-means

Another difference b/w k-means & Gaussian mixture models which are univariate or multivariate they first prefer to perform hard classification , soft classification.

a Gaussian mixture model try to form the mixture i.e. superposition of multiple Gaussian distributions

So we can conclude that the major difference between k-mean & Gaussian mixture models validates for variance and gives output for the probability of a particular data point which actually belongs to each of the k clusters.

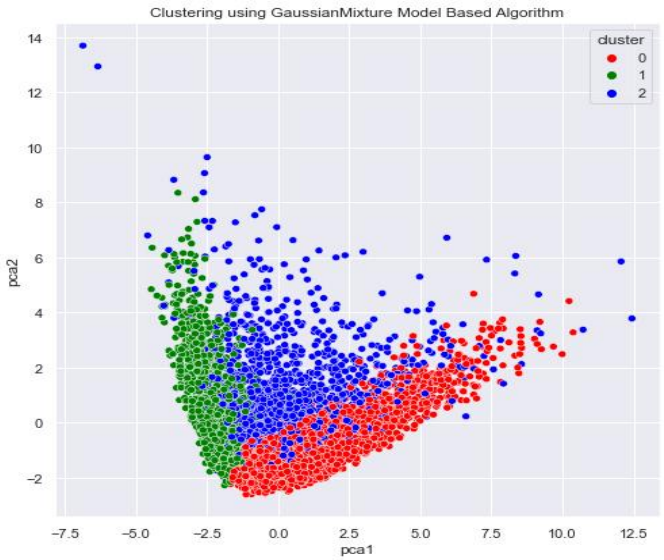

Fig 6 : Clustering using gaussian Mixture

3.4 DBSCAN method or Density based clustering

Like algorithms explained in the above, all clustering methods use the unique approach ,firstly we need to find out the similarities & it can be used to form the data points into sub-groups . We mainly focused on clustering of applications with noise that is the DBSCAN clustering algorithm . Subgroups will be present like huge regions in current data space, which are separated by regions like lower density of points. This method will be based on the notion of “clusters” &“noise”. The main idea of this method is For every data point in the cluster ,corresponding data point should be of equal radius which contains small no of points

What is the purpose of using this algorithm?

dividing methods like 1. k-means, 2.pam and 3.hierarchical clustering are very useful for determining the spherical shaped clusters or we can say convex clusters [9]. So this algorithm may work only on compact and well-separated clustersThis algorithm is also affected by presence of noise & some errors in our dataset.

data that we take may contain outliers and noise like Subgroups or present like random shape and Data may contain irregular noise.

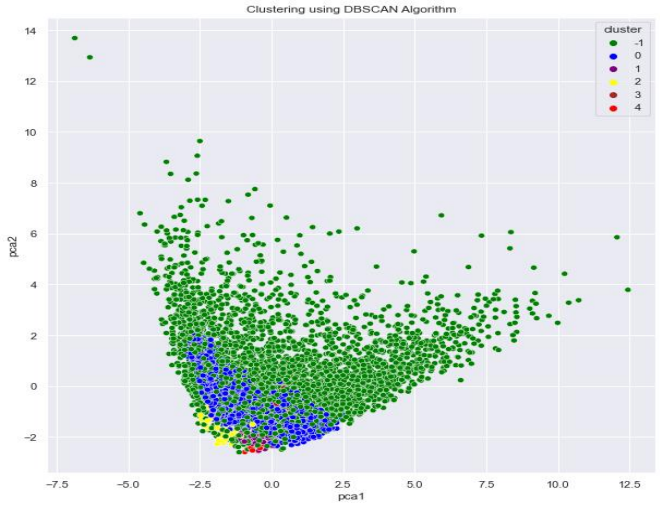

Fig 7 : Clustering using gaussianMixture Model based clustering

Drawbacks of K-algorithm:How does it differ from DBSCAN ?

In the K-Means algorithm It actually formed spherical clusters only but k-means method is failed when data isn't spherical(having equal variance in all directions)

1. This K-Means Method is sensitive towards the outliers,because they can damage the clusters in K-Means to a huge extent.
2. This algorithm requires specifying the total number of subgroups priory etc.

So we come to know that this DBSCAN algorithm finds all the disadvantages of the K-Means algorithm. This Method recognizes the large region by forming clusters of all data points that are much close to each other on the basis of distance measurement.

IV.CONCLUSION

We performed different Cluster analysis ,we tried to analyze and draw some meaningful insights. We saw that all the data points are clustered nicely with very less errors by using k-means clustering as compared to other clustering algorithms. So we'll use the K-Means model for clustering in this dataset.

The paper analyzes various methods that exist for clustering high dimensional data. Some algorithms (Traditional) will be useful to only low dimensional data, that is, datasets which contain a small number of attributes. Whereas Hard clustering algorithms are based on "classical set theory" and they require an object that belong to subgroup or not related to cluster I want to conclude that When we make comparison of some clustering methods which are traditional and based on fuzzy clustering methods ,these will be to be good for clustering the huge dimension dataset.

V. ACKNOWLEDGMENTS

This paper is based on the analysis of clustering and I would like to Appreciate my institute professors for teaching machine learning in the best manner and my seniors for their guidance and suggestions.

VI.REFERENCES

[1] Cheng-Far Tsai and Tang-Wei Huang (2012) DBSCAN: A Quick Density-Based Clustering Technique Idea International Symposium on Computer, Con-sumer and Control, pp. 638-641.

[2] Amandeep Kaur Mann and Navneet Kaur, "Review Paper on Clustering Techniques", Global Journal of Computer Science and Technology, Software and Data Engineering (0975-4350), Volume 13 Issue 5 Version 1.0 Year 2013

[3] Mihika Shah and Sindhu Nair, A Survey of Data Mining Clustering Algorithms, International

Journal of Computer Applications (0975 – 8887) Volume 128 No.1, October 2015

[4] S. Anitha Elavarasi and Dr. J. Akhilandeshwari (2011) A Survey on Partition Clustering Algorithms, International Journal of Enterprise Computing and Business Systems.

[5] S.Vijayalakshmi and M Punithavalli (2012) A Fast Approach to Cluster-Ing Datasets using DBSCAN and Applications (0975 – 8887) Vol 60– No.14, pp.1-7.

[6] Adela Tudor, Adela Bara, Romania and Juliana Botha (2011). "Solutions for Analyzing CRM Systems-data Mining Algorithms", International Journal of Computers, Vol. 5, No. 4, pp. 485-493.

[7] Preeti Baser and Dr. Jatinderkumar R. Saini, A Comparative Analysis of Various Clustering Techniques used for Very Large Datasets, International Journal of Computer Science Communication Networks, Vol 3(4),271-275

[8] Sunita Jahirabadkar and Parag Kulkarni (2013) Clustering for High Dimensional Data: Density based Subspace Clustering Algorithms, International Journal of Computer Applications (0975 – 8887) Vol 63 – No.20, pp. 29-35

[9] Agrawal, R and Srikant, R (1994). "Fast algorithms for mining association rules in large databases", In Proceedings of 20th International Conference on Very Large Databases, Santiago de Chile, Chile, Morgan Kaufmann Publishers Inc., pp. 487–499.

[10] Agrawal, R., Mannila, H., Srikant, R., Toivonen, H and Verkamo, A (1996). "Fast discovery of association rules", Advances in Knowledge Discovery and Data Mining, American Association for Artificial Intelligence. pp. 307–328.

[11] Ada Wai-Chee Fu and E, Ng Ka (2002). "Efficient Algorithm for Projected Clustering", Proceedings of the 18th International Conference on Data Engineering, DOI: 10.1109/ICDE.2002.994727.

Cite this article as :

Pradeep Bolleddu, "An Analysis of Clustering Techniques", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 8 Issue 2, pp. 52-57, March-April 2022.
Available at doi :
https://doi.org/10.32628/CSEIT22821
Journal URL : https://ijsrcseit.com/CSEIT22821