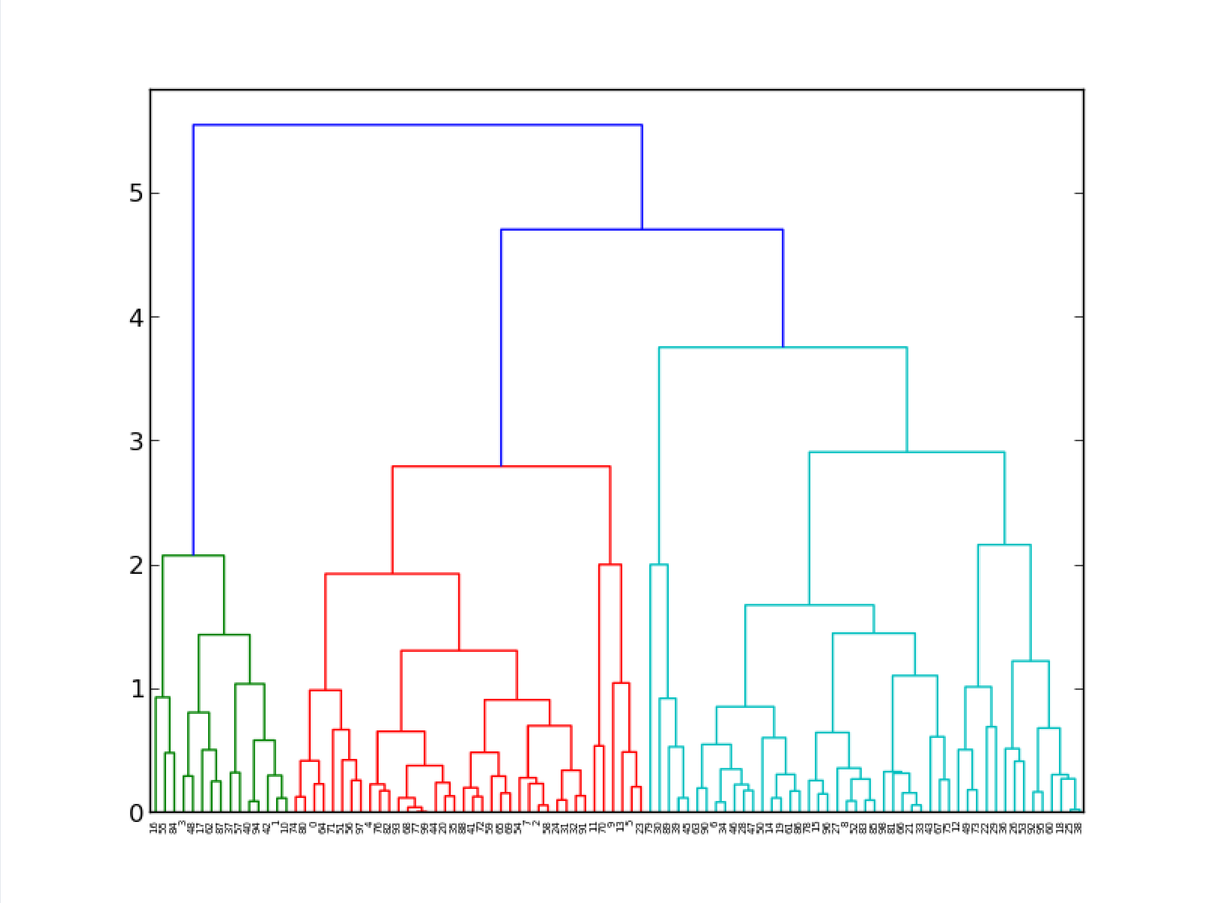
Hierarchical Clustering



Submitted By:

Nidhi Arun

Contents

Abstract

This project helps us to understand hierarchical in the case where the pair of points have dissimilarity scores(ex:distances) as a part of the input. The recently introduced objective for points with dissimilarity scores results in every tree being a ½ approximation if the distances form a metric. This shows the objective does not make a significant distinction between a good and poor hierarchical clustering in metric spaces.

Motivated by this, this project develops a new global objective for hierarchical clustering in Euclidean space. The objective captures the criterion that has motivated the use of divisive clustering algorithms : that when split happens, points in the same cluster should be more similar than points in different clusters. Moreover ,this objective gives reasonable results on ground truth inputs for hierarchical clustering.

Objective

Hierarchical Clustering is a recursive partitioning of a dataset into clusters at an increasingly finer granularity.Motivated by the fact that most work on hierarchical clustering was based on providing algorithms ,rather than optimizing a specific objective.

The main aim of this project is to understand the datasets more clearly in an efficient manner.

Das Gupta framed similarity based hierarchical clustering in one that minimizes a particular cost function .He showed that this cost function has certain desirable properties :to achieve optimal cost,disconnected components must be separated at higher levels of the hierarchy,and when the similarity between data elements is identical,all clusterings achieve the same cost.

Introduction

Hierarchical Clustering is another unsupervised machine learning algorithm,which is used to group the unlabelled datasets into a cluster and also known as hierarchical cluster analysis or HCA.

In this Project we will be looking into the types of clusters and the way they work on the codes and helps to sort the data sets with simple line of code.

Methodology

In this algorithm , we develop the hierarchy of clusters in form of a tree,this tree shaped structure is known as DENDOGRAM.

The hierarchical clustering technique has two approaches:

1.Agglomerative:

This approach is a bottom up approach,in which the algorithm starts to take all the data points as single clusters and merges all till one cluster is left.

2.Divisive:

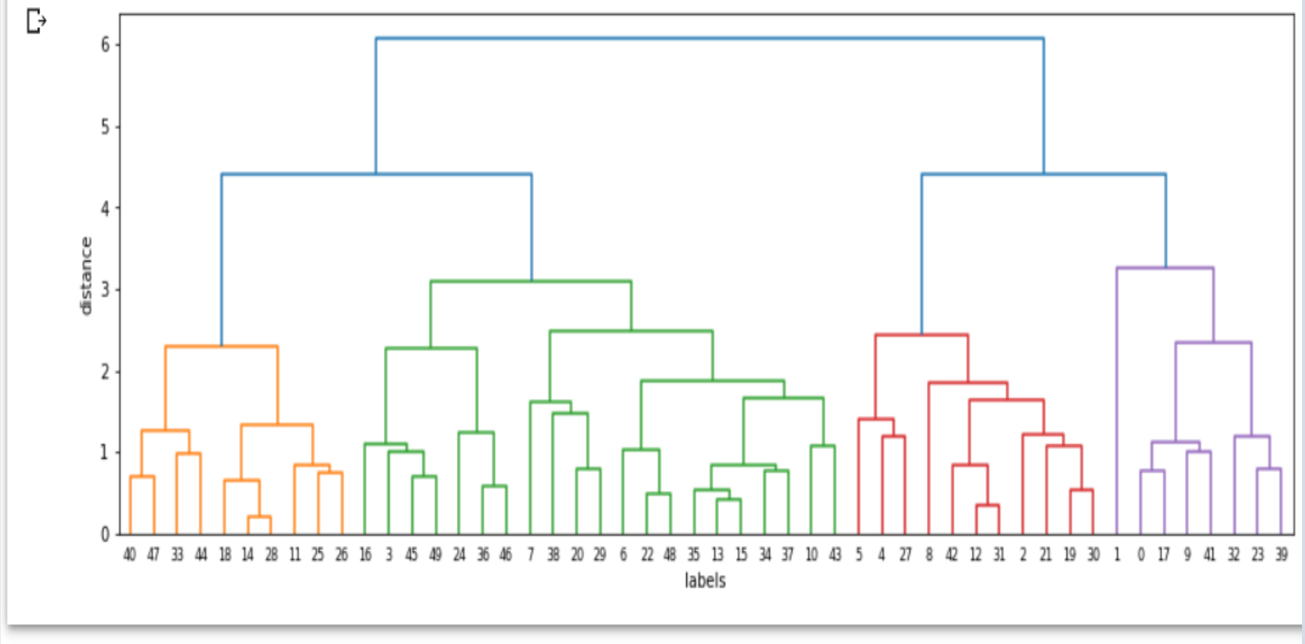
This approach is the reverse of the agglomerative approach as it is top down approach.

Code

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
crime_data=pd.read_csv(r'C:\Users\nidhi\Hierarchial-Clustering-using-Python\crime_data.csv')
crime_data.isnull().sum()
def norm_func(i):
    x=(i-i.mean())/i.std()
    return x
norm_data=norm_func(crime_data.iloc[:,1:])
#dendrogram#
from scipy.cluster.hierarchy import linkage
z=linkage(norm_data,method='complete',metric='euclidean')
plt.figure(figsize=(15,5));plt.xlabel('labels');plt.ylabel('distance')
import scipy.cluster.hierarchy as sch
sch.dendrogram(z,leaf_rotation=0.,leaf_font_size=8.)
plt.show()
#Agglomerative clustering#
from sklearn.cluster import AgglomerativeClustering
```

```
h_complete=AgglomerativeClustering(n_clusters=4,linkage='complete',affinity='euclidean').fit(norm_data)

type(h_complete)

cluster_labels=pd.DataFrame(h_complete.labels_)

final_data=pd.concat([cluster_labels,crime_data],axis=1)

final_data.rename(columns={0:'clusters'},inplace=True)

final_data.groupby(final_data.clusters).mean()

final_data.to_csv('crime_final.csv')
```

OUTPUT:



Conclusion

This project gives a new objective function for hierarchical clustering designed to mathematically capture the principle used to motivate most divisive algorithms.That is ,comparing inter vs intra clusters distances at splits in the tree.