Graph Clustering Assignment

Akash G   CB.EN.U4CSE17105

Clustering:

Clustering is a type of unsupervised learning method.

An unsupervised learning method is a method in which we draw references from datasets consisting of input data without labelled responses. It is used as a process to find meaningful structure, explanatory underlying processes, generative features, and groupings inherent in a set of examples.

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups.

Clustering Analysis:

The process of dividing a set of input data into possibly overlapping, subsets, where elements in each subset are considered related by some similarity measure.

Graph Clustering:

Graph clustering refers to clustering of data in the form of graphs. Two distinct forms of clustering can be performed on graph data. Vertex clustering seeks to cluster the nodes of the graph into groups of densely connected regions based on either edge weights or edge distances.

Graph-Based Clustering:

- Graph-Based clustering uses the proximity graph.
- Start with the proximity matrix
- Consider each point as a node in a graph
- Each edge between two nodes has a weight which is the
- proximity between the two points
- Initially the proximity graph is fully connected
- MIN (single-link) and MAX (complete-link) can be viewed
- as starting with this graph

Information networks – An application of graph clustering:

In any communication network, graph clustering serves as a tool for analysis, modelling and prediction of the function, usage and evolution of the network. Applications include business analysis, marketing, improving the infrastructure, and identifying anomalous use.

In computer networks, clustering may be used to identify relevant substructures and to analyse the connectivity for purposes of modelling or structural optimization; the canonical example is the Internet and the structure of the Autonomous Domains.

In the World Wide Web, clustering of hypertext documents–representing each web page by a vertex and each hyperlink by an edge–helps to identify topics and other entities formed by several interconnected documents.

Clustering algorithms are also used in the structural design and operation of ad hoc and sensor networks. For networks with a dynamic topology, with frequent changes in the edge structure, local clustering methods prove useful, as the network nodes can make local decisions on how to modify the clustering to better reflect the current network topology.

Some other applications:

In the business world, stock market data can be clustered: represent each stock by a vertex and place weighted edges to represent the correlations of the valuations of the stocks in the stock market.

Such a representation allows for the identification of clusters of stocks that either all gain or lose value together, or alternatively–varying the cluster definition–stocks that appear to behave independently of each other. Such knowledge is useful in portfolio management when one wishes to distribute and/or concentrate investments.

In logistics, the hub-location problem and other kinds of facility-location problems are of interest. Several clustering-based heuristic solutions have been proposed to the hub-location problem and heuristics and approximation algorithms that rely on a graph representation and a clustering computation exist for the facility-location problem as well.

A related problem of sales territory design also has proposed solutions building on the use of clustering methods.