EXperimental Learning

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Big Data and Social Analytics certificate course

MODULE 4 UNIT 1
Video 6 Transcript

MIT BDA Module 4 Unit 1 Video 6 Transcript

Speaker key

XD: Xiaowen Dong

HY: Hapyak

XD: Hello, everyone. In this class we are going to talk about graph clustering, which is one of the most popular problems in graph-based data processing. I will first motivate our studies through a classical example, we will then look at the main philosophies behind three main approaches to graph clustering, which are spectral graph partitioning, hierarchical clustering and modularity maximization.

The general goal of graph clustering is to, based on edge structure, partition the vertices of the graph into different groups such that there are many edges within the same group but only a few between different groups. Therefore graph clustering can reveal sub-groups of vertices in the graph which we sometimes call communities; detecting such communities can be of great importance in sociology, biology, and in many other related fields.

Here is a classical example on graph clustering called Zachary's Karate Club. In the early 70s when Zachary studied the social network of a karate club that consisted of 34 vertices representing the members of the club and 78 edges representing the friendships between them.

00:01:26

The network is illustrated on the right. During the studies a conflict between vertex 1, the instructor, and vertex 34, the president, led to the splitting of the members into two groups. Some went with the instructor and others with the president, as indicated by the colors in the figure. Interestingly, based on the topology of the network, Zachary was able to partition the vertices into two clusters that almost perfectly predicts the splitting with only one mistake. Zachary's Karate Club is an example of how graph clustering techniques can be used to understand social relationships between a group of people.

In general we are interested in studying graph clustering because it provides insights into the structure of the network. First it reveals sub-groups or clusters of vertices in the network. Second, it can also reveal the positions or the rows of particular vertices within their clusters. For example, we see in the example of karate club, vertices 1 and 34 play a central part in their respective clusters, where vertices 3, 9 and 31 can be considered as lying on the boundaries of the two clusters. Third, graph clustering can also reveal a hierarchical structure possibly present in the network. And finally, graph clustering has a wide range of practical applications.

As a remark, since graph clustering has been studied in many different fields, sometimes different terminologies have been used, for example, both graph partitioning and community detections can be considered as graph clustering problems. However, the difference is that in graph partitioning the number and the size of the clusters are usually specified, at least roughly, where in comparison, in community detection such information is not specified beforehand.

HY: What is the difference between graph partitioning and community detection?

A. **The number and size of the subgroups is specified for graph partitioning, while this is not the case in community detection.**
B. The number and size of the subgroups is specified for community detection, while this is not the case in graph partitioning.
C. The number and size of the subgroups is specified for graph partitioning and community detection.
D. The number and size of the subgroups is specified for neither graph partitioning nor community detection.

00:03:22

Spectral graph partitioning

XD: We first look at a graph clustering method called spectral graph partitioning; say we want to partition the graph into k clusters of more or less equal size, in the example shown here k = 3. This can be thought of as a graph cut problem where we want to find a partition by cutting as few edges as possible in the graph given the constraint that the sub-groups are of more or less similar sizes. The problem is called a normalized graph cut where we want to find a set of clusters C1, C2 to Ck such that we minimize the graph cut objective.

The solution can be approximated by looking at the leading eigenvectors of the so-called graph Laplacian matrix, and because we are using eigenvectors of matrices, this is called spectral graph partitioning. We will discuss this is more detail in additional notes. Since we usually do not know beforehand the number and the size of the clusters, graph partitioning approaches such as spectral graph partitioning can be limiting in some cases. In the following we introduce community detection algorithms that do not have such a limitation

HY: Hierarchical clustering

XD: so specifically the second method in this video is hierarchical clustering, which is a popular clustering method in social network analysis.

00:04:42

Here we look at single linkage clustering, which is one type of hierarchical clustering algorithms that follow a bottom-up approach. Here is how it works; we start with each vertex in a graph being individual clusters and at each step we merge the closest pair of clusters into a bigger cluster and we repeat the process. In single linkage clustering the distance between a pair of clusters is defined as the minimum distance between vertices in each cluster.

Here is an example where we have six vertices and their pairwise similarities are measured according to euclidean distance. Let's say the closest pair of clusters are b and c, therefore we merge them into one cluster, bc. The next closest pair are d and e which we also merge into one cluster, de. Now we have four clusters, a, bc, de and f. The next closest pair are de and f because the minimum distance between vertices in each cluster is the smallest for this pair. We repeat the process until we have only

one cluster in the end which includes all the vertices. The whole process can be illustrated by a so-called dendrogram.

Hierarchical clustering does not need the number and the size of the clusters as input, and it can give a hierarchical structure of the network. However, it is not clear where we should cut the dendrogram to output a final set of clusters. These algorithms can also get quite expensive for large graphs.

HY: With Hierarchical clustering you need to know the number and size of clusters.

True

Incorrect. With hierarchical clustering you do not need to know the number and size of clusters, you just need to know the distance between clusters.

False

Correct, well done. With hierarchical clustering you do not need to know the number and size of clusters, you just need to know the distance between clusters.

XD: The third method that we look at is based on a different philosophy. We call that in spectral graph partitioning, we use the graph cut objective as a quality function to measure the goodness of the partition. Here we introduce another quality function called the modularity. Conceptually, for a group of vertices the modularity q is the difference between the actual number of edges in that group and the expected number of edges according to a random graph structure.

00:06:49

Intuitively, the larger the modularity, the more different the given structure is from a random structure. Therefore, a partition of the graph can be obtained by maximizing modularity over all the sub-groups in that partition. Modularity maximization based clustering methods are popular because modularity provides a principled understanding of the graph clustering problem with clear definition of communities and the strengths of communities and partitions.

HY: Louvain method

XD: A popular graph clustering method based on modularity maximization is the Louvain method. This is a greedy algorithm based on local modularity maximization. Here is an illustration of this method; similarly to the hierarchical clustering algorithms that we have seen before, the Louvain method iterates between two main steps. In the modularity optimization step vertices are put into local clusters according to a modularity maximization procedure. In the community aggregation step, vertices in the same local cluster are merged into super vertices for the next iteration.

00:07:55

The algorithm stops at some iteration when modularity cannot increase anymore. The Louvain method is a fast graph clustering method even on relatively large graphs, therefore it has been widely applied in many practical problems.

HY: Which of the following two approaches is the most resource efficient when working with large graphs.

1. Hierarchical clustering
2. **Louvain method**

The Louvain method is cost effective, while graph clustering is expensive when working with large graphs.

XD: As a recap, in this video we have talked about graph clustering problems and mentioned three types of approaches, namely spectral graph partitioning, hierarchical clustering and the modularity maximization based methods such as the Louvain method. I hope you understand the main philosophies behind these methods.