Clustering - Unimodal, to Cluster Ensembling, to Multi-View Clustering

Captain Iain Cruickshank
icruicks@Andrew.cmu.edu
Summer Institute 2020

Community Detection and Clustering

- A common task in Network Science is to detect communities or meso-structures within our networks
  - Find groups of nodes who are more ‘similar’ to each other than to other groups of nodes
- This is done by clustering our networks
There are many difficulties in clustering data

- Clustering a data set often results in problems that are non-convex, NP-Hard, and has no (or many possible) labels

- Many clustering algorithms are stochastic
  - The initialization matters

- Clustering algorithms have many different types of losses that they are attempting to minimize
  - These loss function can capture different aspects of the data

Have you ever had this problem?

You cluster a data set using something like Louvain or Louvain Network Clustering to get out cluster labels...

... And then, later, when you run the same clustering technique on the same data set, you get a different set of labels than what you got the first time...
...Or these problems?

- There are multiple ways of clustering your data set and you are not sure which is the right one.
- There could also be multiple user-set parameters for any given clustering algorithm and you are not sure how they should be set.

A Better Way to Cluster Our Data

- What if we could combine all of the possible valid clusterings of our network to produce a robust and more accurate clustering of our network?
  - Able to take in any kind of clustering of our network.
  - Able to ameliorate the effects of stochastic algorithms.

- We can! By *Ensembling our Clusters*
**What is Ensembling?**

- Ensembling is the combining of many different methods to get better results than any one method can produce.

- Ensemble methods are meta-algorithms that combine several machine learning techniques into one model.

- Very successful in supervised learning
  - Highly used in Competitive Data Science Competitions.

**Introducing Cluster Ensembling**

- Combine a collection of cluster labels into one labeling scheme for the data.

  ![Diagram of Cluster Ensembling](Diagram)

- Ensemble clustering should be robust and more representative of the cluster structure in the data than any given clustering.
Cluster-based Similarity Partitioning Algorithm (CSPA)

- One of the original cluster ensembling algorithms proposed by Strehl and Ghosh in 2002 in the seminal work for the field

- Has seen many modifications over the years for things like link weighting, better graph clustering, and iterative refinement

- We will just focus in on the basic algorithm for today

Cluster-based Similarity Partitioning Algorithm (CSPA)

- Measure the similarity between every object being clustered by their ensemble clustering memberships, and then cluster this similarity matrix via suitable techniques

- There are many proposed ways of calculating this similarity
Ensembling of Multiplex Networks

- Cluster Ensembling can take in any clustering over the same objects

- So, what if we have more than one network defined over the same objects?
  - E.g. multiple social media accounts, online and in-person contacts, many different types of interactions
  - Multiplex and multilayers networks
Ensembling of Multiplex Networks

- We can even use cluster ensembling to combine *multiple views of our data* into *one clustering!*

- Can even be used to incorporate partial or incomplete views
  - You have labels for a population that were previously determined (i.e. user segments, previous clustering results, etc.) and want to combine those labels into one label
  - Some actors do not participate in certain actions (i.e. some Twitter users never re-tweet)

Time for Another Example!
Recap

- Clustering is the means by which we find communities in our networks
  - Find those individuals which are more ‘similar’ to each other than to other groups of individuals
- Cluster ensembling is a means of combining various clusterings over the same objects to get a better clustering of those objects
- Cluster ensembling can be used for standard networks as well as multiplex networks and even partially complete networks
- Cluster ensembling is an active area of research and has many useful techniques and strategies coming out