Bryson Li

CS 410 Technology Review

Topic: AGNES (AGglomerative NESting) Algorithm + Divisive Clustering Algorithms

The main purpose of clustering is to discover “natural structures” in the data by grouping similar structures together. This allows us to get an overall idea of what is in a collection of text. There are two types of clustering - agglomerative, which is a bottom-up approach where you gradually group similar objects into larger clusters. Another approach is divisive, which is a top-down approach where you gradually partition the data into smaller clusters. In this technology review I will be looking at a specific agglomerative clustering algorithm called AGNES and discussing its applications and performance comparisons with other clustering algorithms[1].

One specific application that I will be examining is a research study done on reducing the dimensions of phenotypes. First off, the researchers wanted to combine multiple phenotypes together because they noted that testing relationships between multiple phenotypes and variants was more effective than only testing relationships between a single phenotype and the variants. In this study, the researchers found that when multiple phenotypes are analyzed together, there are multiple groups of phenotypes that are very similar to each other, so they can be grouped together and reduced into “representative phenotypes” which “represent” multiple phenotypes with “one.” Through the research process, these researchers developed an “agglomerative nesting clustering algorithm for phenotypic dimension reduction analysis” or “AGNEP” for short (Liu). The steps of this process are as follows. First, AGNEP uses an agglomerative nesting clustering algorithm to group the phenotypes together (Liu). Next,

[1] most of the information in the introduction taken from cs 410 lecture videos

principal component analysis is performed on these groups to make these representative phenotypes (Liu). Finally, they also performed multivariate analysis on these to test association between the representative phenotypes and some genetic variants (Liu). They ran 3 simulations of this experiment on different datasets (genetic structures) and calculated the Silhouette coefficient of them in order to evaluate their results (Liu). They ended up finding out that the reduced dimension represents the phenotypes well (they are linear combinations of the phenotypes) (Liu).

Another group decided to develop a divisive clustering algorithm for biological data, since there is an increasing amount of biological data on cancers and diseases, and doctors always want to find sub-conditions, which divisive clustering algorithms fit. They wanted to develop a divisive clustering algorithm to fit their need that was not as computationally expensive (since divisive clustering algorithms tend to be more computationally expensive compared to agglomerative clustering algorithms) (Sharma). They called this method DRAGON (short for Divisive hieRArchical maximum likelihOod clusteriNg) (Sharma). There are a lot of math/ statistics and biology related concepts/assumptions being used here, but essentially there is a likelihood function that is being maximally increased each time a sample is removed (Sharma). Of course, this requires a specific log likelihood function to be defined (which they did in their paper) (Sharma). They found that the search for c clusters was to the order of $O(n^2c)$ which is actually much better than how divisive clustering algorithms usually are (which is to the order of 2^n) (Sharma). They even found that this method performed better than agglomerative clustering methods for the same data (Sharma).

In this brief review of two papers that discuss the use of clustering - agglomerative in one, and divisive in another, I thought it was interesting that each of these clustering algorithms has their place, depending on where it was used. In one paper, agglomerative clustering algorithms were used for clustering “representative phenotypes”, and in another, divisive clustering algorithms were developed to cluster sub-condition of cancers and other diseases. This shows that a lot of work has been done in this field, and there continues to be more work done each day on these clustering algorithms (which are simple in general but can be improved a lot).

Citations

1. Hierarchical Clustering in R: The Essentials
https://www.datanovia.com/en/lessons/agglomerative-hierarchical-clustering/#

2. Liu, Fengrong et al. “AGNEP: An Agglomerative Nesting Clustering Algorithm for Phenotypic Dimension Reduction in Joint Analysis of Multiple Phenotypes.” Frontiers in genetics vol. 12 648831. 26 Apr. 2021, doi:10.3389/fgene.2021.648831
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8107386/

3. Sharma, A., López, Y. & Tsunoda, T. Divisive hierarchical maximum likelihood clustering. BMC Bioinformatics 18 (Suppl 16), 546 (2017).
https://doi.org/10.1186/s12859-017-1965-5