Evaluation of Clustering Algorithms

Derek Chiu

August 13, 2015

Contents

1 Introduction

Cluster analysis is the unsupervised learning method of assigning entities into different groups based on one or more of their attributes. The goal is to place similar objects together and separate dissimilar objects. For example, in genomics studies, we frequently try and cluster patient samples measured on a large number of molecular features. When we get a clustering assignment from an algorithm, we often want to evaluate its performance. Ideally, a good clustering algorithm is able to differentiate entities with no knowledge of the true class labels. In addition, we want the algorithm to arrive at a stable and optimal number of clusters. There are two main categories of clustering evaluation: **external evaluation** and **internal evaluation**.

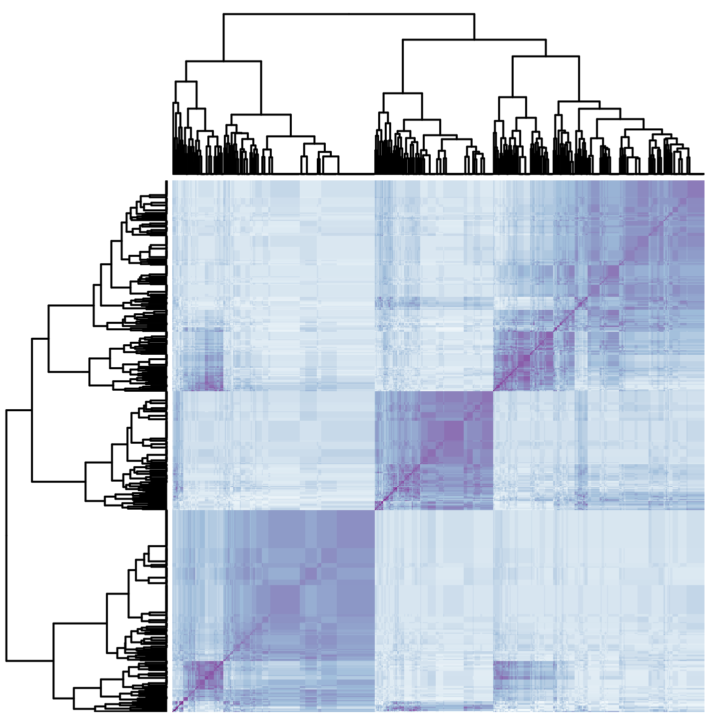

Figure 1:

2 External Evaluation

External evaluation usually refers to the case when we compare our clustering assignments to true class labels, or have some gold standard to compare to. In applications, this might be the published clustering result. The downside of using external evaluation is that the reference classes may not be correctly clustered themselves, and we are treating these as the norm. None the less, we can explore a few metrics.

2.1 Adjusted Rand Index

Algorithms	ARI
NMF (Divergence)	0.4799
NMF (Euclidean)	0.4435
PAM (Spearman)	0.427
Hierarchical (Diana)	0.4221
K-Means (Spearman)	0.4049
PAM (Euclidean)	0.3434
Hierarchical (Euclidean)	0.3275
K-Means (Euclidean)	0.2559
PAM (MI)	0.07465
K-Means (MI)	0.07369

2.2 Mutual Information

Algorithms	MI
NMF (Divergence)	0.6723
NMF (Euclidean)	0.6416
PAM (Spearman)	0.621
Hierarchical (Diana)	0.5976
K-Means (Spearman)	0.5889
PAM (Euclidean)	0.546
Hierarchical (Euclidean)	0.481
K-Means (Euclidean)	0.4598
PAM (MI)	0.1652
K-Means (MI)	0.1169