11741 Reading Summary, Ch 16.3

Xin Qian (xinq@cs.cmu.edu)

Clustering has a goal of getting the clusters highly cohesive and inter-cluster dissimilar, which defines the quality of a clustering and yields an internal criterion. As an alternative, there's the direct evaluation from the information retrieval application. This section infroduces 4 external criteria to measure clustering quality. We resort to gold standard or evaluation benchmark in computing an external criterion.

Purity is calculated as the number of correct assignment divided by the total number of documents. Purity has a value from 0 to 1. It is easy to have a high purity when the number of clusters is large but we cannot increase the number of clusters without a limit at the cost of the clustering quality.

Normalized mutual information is the amount of class information acquired given the cluster divided by the arithmetic mean of entropy. A minimum value of mutual information is 0, which indicates we still cannot tell which cluster a document belongs to even when we know its class. Maximum mutual information happens when a clustering perfectly replays the original classes or further subdivides the class hierarchy. However, MI has zero penalty on large clustering cardinality. We thus normalize MI by the entropy as entropy increases with the number of cluster. NMI has a value between 0 and 1.

Random index measures the percentage of correct pairwise decisions. TP is a correct assignment for two similar documents. TN is a correct assignment to separate two dissimilar documents. FN is a wrong assignment to separate two similar documents. FP is a wrong assignment for two dissimilar documents to be put together. RI is calculated as (TP+TN)/(TP+FP+FN+TN). This assumes separating similar documents is of equal harm as putting paris of dissimilar documents in the same cluster. Random index is derived from the perspective of information theory. The process of clustering can be viewed as a series of dichotomy decision.

F measure revisits this formula by plugging in a boost factor $\beta > 1$. Due to its flexibility, it is advantageous to evaluate clustering with F.