Winnie Ngo and Eric Oh

Evaluation Analysis

k	SSE	AIC	Silhouette Index
2	2965.53	3281.54	0.28
3	2210.78	2684.78	0.22
4	2130.94	2762.94	0.19
5	1985.70	2775.70	0.13
6	1600.32	2548.32	0.19
7	1662.83	2768.83	0.42
8	1562.15	2826.15	0.23
9	1498.27	2920.27	0.21
10	1545.44	2967.44	0.20

The table values for SSE, AIC and Silhouette Index are averages of each metric using the results from 10 runs for each k value tested.

SSE is the measure optimized by k-means. In k-means, increasing k decreases SSE. Thus using SSE as the main validation metric, k=2 is the best k value. AIC measures the quality of the statistical model for a given set of data. It quantifies the trade-off between how well the model fits that data and how complex the model is. Thus using AIC as the main validation metric, k=6 is the best k value because it yields the smallest AIC value. The Silhouette Index metric assigns to each object describing how well it fits into its cluster. The object matches its cluster well when its silhouette index is close to one and poorly matches it if its index is close to zero or negative. Thus, k=7 is the best k value using the silhouette metric as the main validation metric. SSE, AIC and Silhouette Index are all measures of external evaluation for the quality of a clustering. But good scores on an internal criterion do not necessarily translate into good effectiveness on the clustering algorithm and on their own these are not great validation metrics. The is why we also need external criterion such as a description of each genes functional category. A good clustering algorithm will result in clusters of genes with similar functions. Looking for this, we saw some of the ‘best’ clusters when using k=5. In addition, we need to normalize our inputs to improve results so that the features that could have been measured on different scales are adjusted to be on a common scale. Therefore, it will make sense to compare the feature values for different genes once they are measured in a common way.

The problem biologists encounter is that given the same data set, different clustering algorithms can generate very different clusters. Eisen et al. (1998) applied a variant of the hierarchical average-link clustering algorithm to identify groups of co-regulated yeast genes. Eisen et al. validated their results by visual inspection using biological knowledge. The paper used the Euclidean distance to measure the similarity between two genes. The resulting data on the relationships among genes are illustrated by a tree whose branch lengths represent the degree of similarity between the objects.