Kidney Q Paper

Evaluating Cluster Reproducibility for Putative Subtype Identification

"Are Clusters Reproducible and Consistent?"

November 27, 2024

David Lewis
lewis3d7@mail.uc.edu

Introduction

Identification of molecular subtypes is important for separating patients that present similar clinical profile but respond to treatment in significantly different ways. Examining molecular subtypes also allows identification of potential treatments that target the molecular basis for the subtype, such as gene expression. However, classification is not a solved problem. Reproducibility is a critical problem within subtype identification. Often, data is volatile enough to change how subtypes are identified between sample cohorts. As such, it is very important to realize the limitations of current methods for identification of molecular subtypes.

This paper hopes to answer questions about classification reproducibility and quality. Notably, much of this paper will delve into different unsupervised clustering methods and their ability to identify molecular subtypes within the context of the paper *Comprehensive Molecular Characterization of Clear Cell Renal Cell Carcinoma*;[1] This paper may be referred to as "the original paper" throughout this document. It should be noted that the subtypes identified in the original paper are not necessarily consistent with other literature.[2]

A note about "Consensus Clustering"

The clustering algorithm in the original paper uses a method of dimensionality reduction called nonnegative matrix factorization (NMF). The data is decomposed into a smaller matrix that is representative of the larger dataset. This is similar to PCA in practice; however, as part of NMF, one can determine the relative contribution of any column in the original matrix to the columns in the smaller matrix. This facilitates assignment of a sample (original matrix) to a cluster (column in small matrix). "Consensus clustering" typically refers to the process of aggregating clustering output for multiple clustering algorithms; however, in the original paper it refers to the average clustering over successive NMF runs. It should be noted that both NMF and k-means are similar in outputs and inputs. Both need to be given an ideal number of clusters and both are linear methods, they cannot determine non-linear relationships within the data. However, NMF clustering is much more computationally intensive, so evaluating it was not feasible for this paper.

Evaluating K-means clustering (Results + Methods)

Initial classification of molecular subtypes is typically done via unsupervised clustering. Many algorithms exist to do this, each with their pros and cons. One commonly used algorithms is called *k-means*. This algorithm is very common due to its ease of use and straight forward interpretation.

However, k-means clustering is only useful if one can determine the ideal number of clusters. Because k-means is an efficient algorithm, and the dataset is relatively small, it is possible to evaluate multiple numbers of clusters within a short amount of time while still forming a representative range of useful

clusters. For clustering methods that have user defined numbers of clusters, there are two methods that are commonly used to find the ideal number of clusters. Since the number of clusters corresponds to the number of subtypes, picking the right number of clusters is essential.

Generally, the ideal number of k-means clusters was consistent at 2. However, when the significantly differential data for the tumor and non-tumor conditions were evaluated (this is the signature defined in the group paper), the number of clusters was inconsistent and changed with random state.

The Elbow Method

One way to pick the ideal number of clusters is to use what's called the "elbow method". This method seeks to minimize the number of clusters as well as the inertia, a measure of "closeness" for data points and cluster centers. Lower inertia values represent more closely associated clusters. The elbow method is useful, but not the end all be all metric for determining the ideal number of clusters. In cases where the "elbow" is not immediately visible, it is difficult to assign the ideal number of clusters.

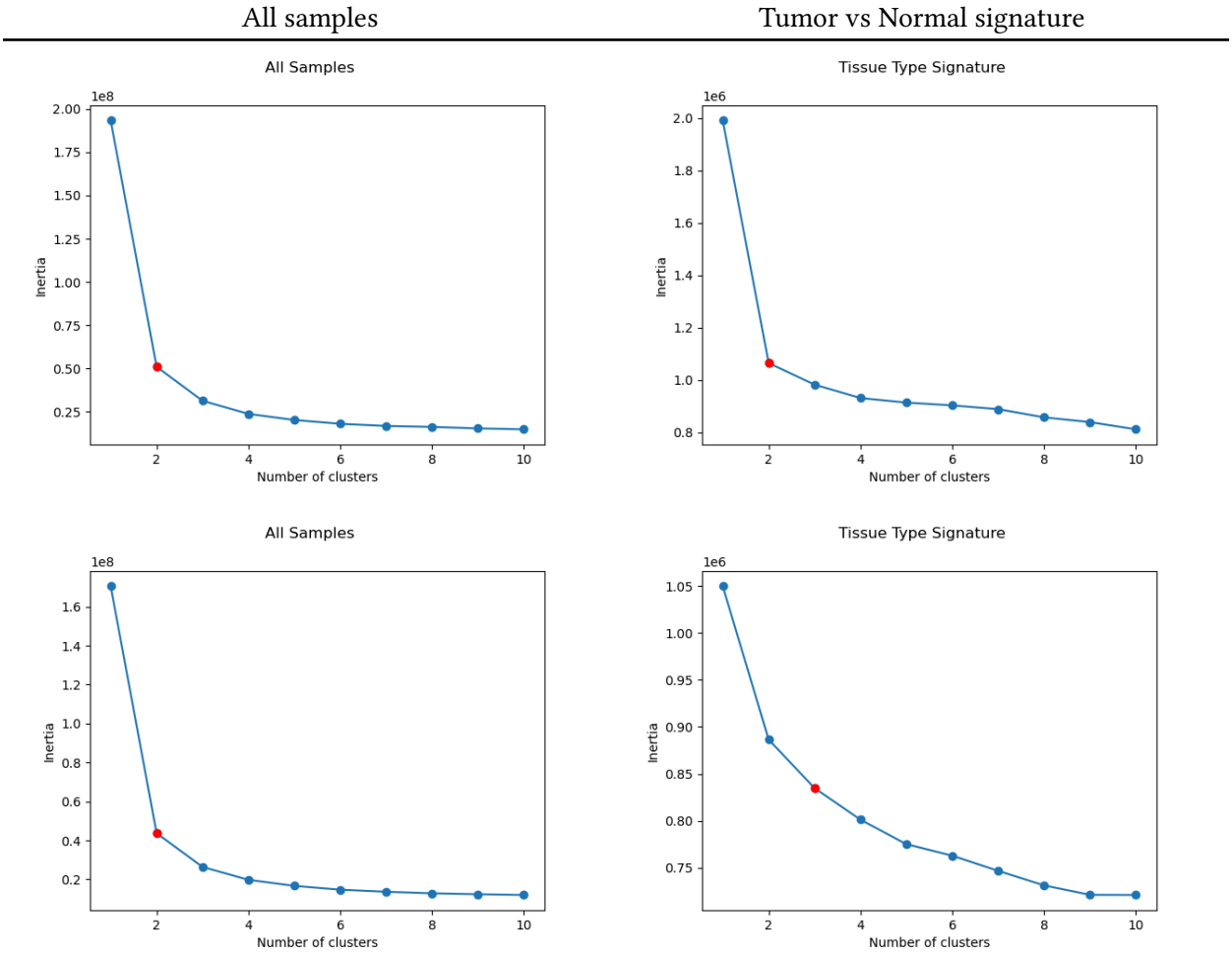

Figure 1: Elbow method for determining ideal number of clusters. The ideal number of clusters is shown in red. Top row of plots contains both the normal tissue and tumor samples. Bottom row of plots only contains tumor samples. Y axis is the inertia value, low values of inertia correspond to better clustering; however, low numbers of clusters are generally better.

The elbow for the bottom-right plot in Figure 1 is difficult to identify. Unfortunately, this is the plot that is the most important to the analysis. These are the genes that are significantly relevant to the relationship between non-tumors and tumors. Additionally, upon further testing, the ideal number of clusters is not stable for the condition in the bottom-right plot when the seed, or random state is changed.

The Silhouette Method

Another method for determining the ideal number of clusters using k-clustering methods is the silhouette method. The silhouette method scores each point in the cluster on a scale from −1 to 1 where higher values represent a better assignment to a particular cluster. The silhouette is generally considered more reliable than the elbow method, but is more costly to calculate, which may be problematic for large datasets. Luckily, the dataset that is used in this paper is not particularly large, so both the elbow method and silhouette method can be repeatedly calculated.

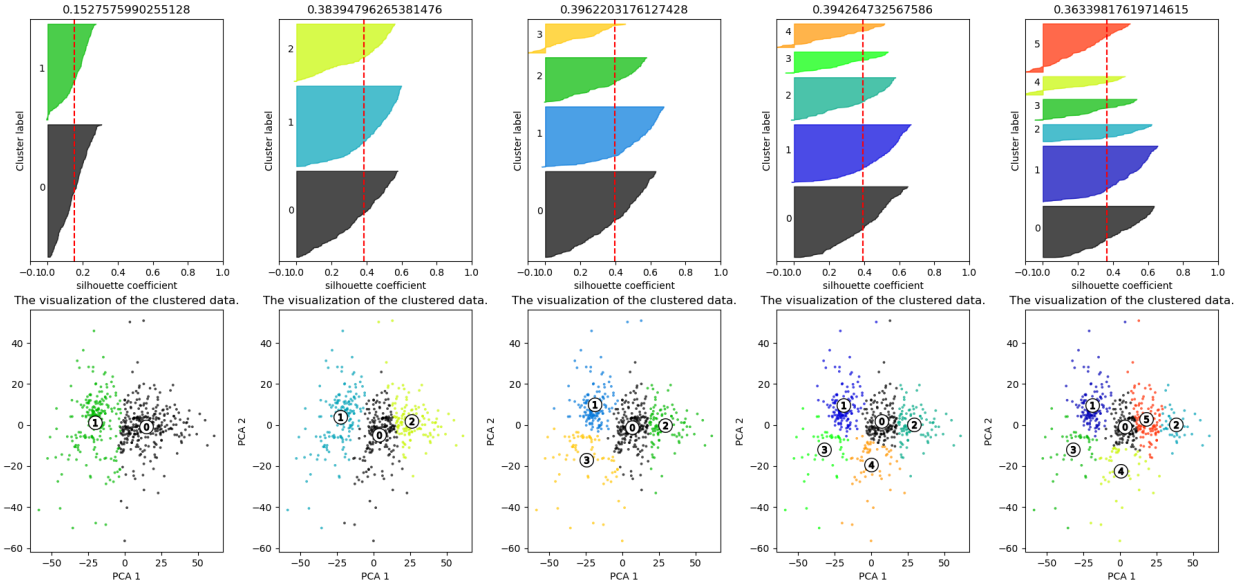

Figure 2: Top row shows Silhouette depicting cluster performance, bottom row shows the clusters plotted using the first 2 principle components from PCA. Data represents the signature for tumor vs normal tissue.

Performing silhouette analysis on the hard to interpret plot from Figure 1 also shows difficulty in selecting the correct number of clusters for the data. The silhouette scores are very similar for 3, 4, and 5 clusters, and the ideal clustering is similarly not stable based on random state.

Discussion

Given the unreliability of k-means clustering for this data, it may be a reasonable assumption that the data is not linearly separable, or does not conform to the assumptions of k-means. If the data does have non-linear associations, then the clustering in the paper would also not be valid. Non-linear clustering methods should be evaluated for this data, and typical consensus clustering should be performed to provide a more balanced picture of the putative subtypes present in the data.

References

1. Creighton C J, Morgan M, Gunaratne P H, et al. Comprehensive Molecular Characterization of Clear Cell Renal Cell Carcinoma. *Nature.* 2013;499(7456):43-49. doi:10.1038/nature12222

2. Brannon A R, Reddy A, Seiler M, et al. Molecular Stratification of Clear Cell Renal Cell Carcinoma by Consensus Clustering Reveals Distinct Subtypes and Survival Patterns. *Genes & Cancer.* 2010;1(2):152-163. doi:10.1177/1947601909359929