Linear Pre-Conditioning in Clustering for Distinguishing Treatment Effects

Kate, Thad, Eva

2019

Keywords:

Abstract

We look at methods to optimize linear transformations of a set of covariates for clustering which best distinguish different treatments.

1 Introduction

The idea of transforming a data set prior to performing a particular statistical analysis is sometimes known as "pre-conditioning", (e.g., Jie and Rohe, 2015). In this paper we investigate linear transformations of a set of baseline variables in order to optimize the results of clustering outcome data. It is well known that results from running a clustering algorithms, such as the k-means algorithm, can depend quite strongly on how the data is transformed. This was illustrated in the context of clustering curves in Tarpey (2007).

In this work, we tie the problem of linear transformations for clustering to that of precision medicine and sub-group identification.

Here is the basic idea. We have a set of baseline covariates $\boldsymbol{x} = (x_1, \ldots, x_p)'$. We also have a set of outcome data, in our case, the outcome measures will be longitudinal trajectories from different treatment groups, e.g., active drug and placebo. In a disease such as depression, there is often a very high-degree of overlap in outcome measures between different treatment groups.

Our objective is to find a linear transformation of $\boldsymbol{x}$ so that when the transformed data is clustered, it corresponds to groups that are as homogeneous as possible in terms of the different treatment groups.

We shall compare the results of clustering to the treatment labels using the variation of information (VI) criterion Meilă (2007). Smaller VI values correspond to a better correspondence between two different partitionings of a data set.

In order to clarify the basic idea, consider a best base scenario where we cluster $\boldsymbol{x}$ into k groups. Suppose every individual in a given cluster received treatment A and every individual in another cluster received treatment B. In this ideal scenario, we achieve perfect separation of the outcome measures based on baseline characteristics. This will not happen in practice, and the best we can hope for is that some clusters will be as homogeneous as possible with respect to one or the other outcome measures (while other clusters may be very heterogeneous with respect to treatment received).

2 Approaches

Here are some initial approaches to the problem.

1. Choose a continuous baseline variable, cluster it into k clusters and computer the VI with treatment labels (drug vs placebo). Now do this for several continuous baseline variables and see if any baseline variables lead to a smaller VI compared to other baseline variables.

2. Repeat step one for pairs of variables. For example, cluster $(x_1, x_2)'$ and compute the VI when comparing this clustering to the treatment labels. Now do this for other pairs of variables $(x_j, x_h)'$. Is there a pair of variables that gives a substantially smaller VI?

3. For the best pair of variables found in part (2), can we linearly transform those two variables prior to clustering in order to achieve a lower VI? That is, suppose we are clustering the pair $\boldsymbol{x}^* = (x_j, x_h)'$. Let's look at clustering $\boldsymbol{A}\boldsymbol{x}^*$ for some matrix $\boldsymbol{A}$. How can we determine $\boldsymbol{A}$ to obtain a smaller VI?

4. Repeat this type of analysis for arbitrary combinations of variables $(x_{i_1}, \ldots, x_{i_h})'$ say.

References

Jie, J. and Rohe, K. (2015). Preconditioning the lasso for sign consistency. *Electronic Journal of Statistics* **9**:1935–7524.

Meilă, M. (2007). Comparing clusterings: an information based distance. *Journal of Multivariate Analysis* **98**:873–895.

Tarpey, T. (2007). Linear transformations and the k-means clustering algorithm: Applications to clustering curves. *The American Statistician* **61**:34–40.