QUANTIFYING VARIABLE IMPORTANCE IN CLUSTER ANALYSIS

Christian Hennig [1] and Keefe Murphy[2]

[1] Department of Statistical Science "Paolo Fortunati", University of Bologna, Italy (e-mail: christian.hennig@unibo.it)

[2] Hamilton Institute, Maynooth University, Ireland (e-mail: keefe.murphy@mu.ie)

ABSTRACT: We propose to measure the importance of variables when running a cluster analysis by measuring the similarity of a clustering using all variables with a clustering applying the same method leaving out one variable. If the resulting clustering is very similar, the left out variable does not have much impact. An alternative is to replace the variable by randomly permuted values. Beyond variable selection (on which we will not focus), variable importance measurement is useful for interpreting and understanding a clustering. Also we will use variable importance measurement to discuss whether clustering methods appropriately balance the impact of different variables in mixed type variables clustering

KEYWORDS: variable importance, adjusted rand index, permutation, mixed type variables clustering.

1 Introduction

The quantification of variable importance in cluster analysis is of interest in order to interpret and understand the impact of the variables on a clustering, and potentially also for variable selection.

Consider a data set of n observations $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ with $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$, where $x_{ij} \in \mathcal{X}_j$. $j = 1, \ldots, p$, where $\mathcal{X}_j$ is the sample space for variable j, with potentially different $\mathcal{X}_j$ for different j. Let $X_j = (x_{1j}, \ldots, x_{nj})$ denote variable j. As clusterings, we consider partitions $\mathcal{C} = (c_1, \ldots, c_n)$ of a data set with n observations, where $c_i \in \{1, \ldots, k\}$ indicates the cluster to which $\mathbf{x}_i$ belongs, with $k = \max \mathcal{C}$ the number of clusters. k is not necessarily known or fixed. Let C denote a general clustering (partitioning) method so that $C(\mathbf{X}) \in \{1, \ldots, k\}^n$.

2 Variable importance by leaving a variable out

In a case study regarding socioeconomic stratification based on mixed type (i.e., continuous, ordinal, categorical) variables, in order to assess the importance of the various variables for clustering, Hennig & Liao, 2013 re-ran their clusterings with each variable left out, and they computed the adjusted Rand index (ARI; Hubert & Arabie, 1985) between the clustering with one variable left out and the clustering based on the full data. The ARI takes the value of 1 if clusterings are identical and a value around 0 (that can in principle be negative) if clusterings behave like unrelated random draws of cluster labels; the closer to 1 the ARI, the more similar the clusterings.

Formally: Let $\mathbf{X}^{-j} = (\mathbf{x}_1^{-j}, \ldots, \mathbf{x}_n^{-j})$, where $\mathbf{x}_i^{-j} = \mathbf{x}_i$ with x_{ij} left out. Let $I_{C,\mathbf{X}}(j) = \text{ARI}(C(\mathbf{X}), C(\mathbf{X}^{-j}))$ be the inverse variable importance of j. The interpretation regarding variable importance is that if $I_{C,\mathbf{X}}(j)$ is large, i.e., close to 1, the variable importance is low (therefore "inverse"), because it means that leaving out variable j reproduces pretty much the same clustering. A low value of $I_{C,\mathbf{X}}(j)$ means that leaving out variable j changes the clustering a lot, i.e., X_j has a large impact.

This principle of measuring variable importance can be applied to general clustering methods, and in fact clusterings generated by different methods on the same data can be compared regarding the importance they give to the different variables. This can be particularly interesting when clustering mixed type variables data, as it is a known issue with methods for mixed type variables that they may balance the different variable types, particularly continuous and categorical variables, against each other systematically in different ways, arguably giving too much (or too little) influence to categorical variables in certain situations (Foss *et al.*, 2019).

It is important to note here that variable importance, measured in this way, applies to the empirical result of a clustering method. It can be informative not only about the "true" importance of the variables regarding any supposedly "true" clustering, but also about the way the different clustering methods treat the variables. The downside of this is that these two interpretations may be confounded with each other. This does not seem to be a problem with the proposed method in particular, but rather a general issue with defining and measuring variable importance in clustering. The user therefore needs to be careful when using variable importance measurements for variable selection. More generally, variable selection in clustering is a hard problem, because in general the clustering problem is not well defined, and various clusterings can be legitimate, for potentially different clustering aims, on the same data

set (Hennig, 2015). This means that there is no unique true set of relevant variables, rather the user's choice of involved variables determines the way the resulting clustering can be interpreted.

3 Variable importance by permutation

Breiman, 2001 proposed a scheme for measuring variable importance in random forests. The idea there was to replace a variable by a permutation of its values. This constitutes an alternative approach for measuring variable importance in clustering. For a permutation π on $\{1,\ldots,n\}$, let $\mathbf{X}^{j\pi} = (\mathbf{x}_1^{j\pi},\ldots,\mathbf{x}_n^{j\pi})$, where $\mathbf{x}_i^{j\pi} = \mathbf{x}_i$ except that X_j is replaced by $X_{j\pi} = (x_{\pi(1)j},\ldots,x_{\pi(n)j})$. As this depends on the specific permutation, it is advisable to run m random permutations (say $m = 100$) $\pi_1,\ldots,\pi_m$, and then average ARI-values over the permutations, i.e., define $I^*_{C,\mathbf{X}}(j) = \frac{1}{m}\sum_{h=1}^{m}\mathrm{ARI}(C(\mathbf{X}),C(\mathbf{X}^{j\pi_h}))$.

Both of these approaches (leave a variable out, "I", and permute its values, "I^*") have advantages and disadvantages. Advantages of I are:

- The approach is deterministic, fully reproducible, and computationally simpler.
- It is easy to think of a data set that has a variable left out as "realistic", whereas permuting values of variable X_j may lead to combinations with values of other variables that are unrealistic, due to potential dependence between variables. It may therefore be seen as irrelevant, in a real situation, what would be the effect of a permutation of the values.

Advantages of I^* are:

- Running the clustering method on $\mathbf{X}^{j\pi}$ is the same as running it on $\mathbf{X}$ in the sense that the variables are the same, whereas for I, C has to be run on a data set that has a variable fewer.
- We ran many simulations in which data were generated from Gaussian mixture models, with some variables intentionally generated as noise uninformative for clustering. The results show that I^* is clearly better at distinguishing informative from uninformative variables, i.e., I^*-values will be larger for the uninformative than for the informative variables with clearly larger probability than I-values, consistently over a fairly large number of simulation setups.

This indicates that I^* is preferable for variable selection and interpretation in terms of meaningful vs. noise variables, although it may not be preferable for

investigating the way different methods balance different variables. The most plausible explanation for the empirically superior performance of I^* is that for an informative variable it is worse to be permuted than to be left out, as permuting will replace good information with bad misinformation that can potentially (if a variable is clearly clustered on its own) actively indicate a wrong clustering. Therefore permutation makes more of a difference for variables with strong clustering information than leaving out the variable.

In the presentation we will show simulation results and examples and will discuss them in some detail.

References

BREIMAN, L. 2001. Random forests. *Machine Learning*, **45**, 5–32.

FOSS, ALEXANDER H., MARKATOU, MARIANTHI, & RAY, BONNIE. 2019. Distance Metrics and Clustering Methods for Mixed-type Data. *International Statistical Review*, **87**, 80–109.

HENNIG, C., & LIAO, T. F. 2013. Comparing latent class and dissimilarity based clustering for mixed type variables with application to social stratification. *Journal of the Royal Statistical Society, Series C*, **62**, 309–369.

HENNIG, CHRISTIAN. 2015. Clustering strategy and method selection. *Pages 703–730 of:* HENNIG, CHRISTIAN, MEILA, MARIAN, MURTAGH, FIONN, & ROCCI, ROBERTO (eds), *Handbook of Cluster Analysis*. CRC Press.

HUBERT, LAWRENCE, & ARABIE, PHIPPS. 1985. Comparing partitions. *Journal of Classification*, **2**, 193–218.