MODAL CLUSTERING FOR CATEGORICAL DATA

Noemi Corsini [1] and Giovanna Menardi[1]

[1] Department of Statistical Sciences, University of Padova,
(e-mail: noemi.corsini@phd.unipd.it, menardi@stat.unipd.it)

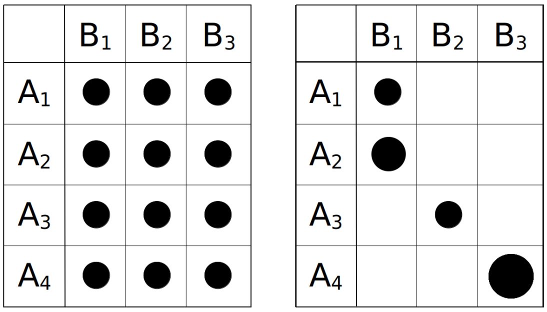
ABSTRACT: Despite the ill-posedness of the clustering task, in the continuous setting a broad consensus is overall acknowledged in defining the concept of cluster. Conversely, a general notion of cluster remains controversial in the presence of categorical data. We propose a novel notion of cluster hinging on the twofold concept of high frequency and association between variables. The former concept, in fact, complies with the cluster notion described by the modal formulation of the clustering problem, which we take advantage of to borrow some operational tools to propose an operational procedure.

KEYWORDS: association, contingency table, graph

1 Introduction

The importance of clustering in statistics has never been questioned over the years, thanks to the many fields in which it finds relevant applications. However, more than to its wide applicability, the proliferation of a voluminous amount of literature on this topic is perhaps due to the ill-posedness of the problem, which is inherent with its unsupervised nature. In fact, when numerical data are at hand, a general agreement is met across alternative notions of cluster, which collectively fall under the heading of groups of similar subjects. Even when more sophisticated density-based cluster formulations are considered, indeed, the underlying notion of cluster implies the observations to be somewhat close to each other.

Conversely, this does not apply to categorical data. While, in principle, a natural clustering gathers subjects within the observed cross-categories of the variables, such description turns out to lack parsimony when either the number of variable and/or the number of categories grows. On the other hand, the lack of a total order among categories makes somewhat controversial even the notion of distance, and increases the arbitrariness in the subsequent definition of cluster, which, in the literature about clustering categorical data, is usually left unspecified.



	B1	B2	B3
A1	●	●	●
A2	●	●	●
A3	●	●	●
A4	●	●	●

	B1	B2	B3
A1	●		
A2	●		
A3		●	
A4			●

Figure 1. *Graphical representation of two contingency tables where the diameter of each circle is set as proportional to the frequency of the cell.*

In this work we attempt the ambitious aim of filling this gap by proposing a novel notion of cluster within the setting of categorical data, along with an operational procedure to identify clusters.

Consider the two toy examples in Figure 1, which describe two alternative frequency patterns of a number of subjects observed with reference to two variables. Even in the lack of a cluster definition, we feel highly shareable to acknowledge that the left panel, where variables are independent, identifies a configuration without clusters (or formed by 12 clusters, as the number of cross-categories); on the other hand, the right panel, characterized by a strong association pattern between the two variables, aggregates the subjects in three clusters. This intuition leads us to build a novel notion of cluster hinging on the twofold concept of high frequency and association between variables, i.e. groups arise as highly populated (aggregations of) cross-categories of variables leading a large contribution of mutual information. The former concept, in fact, complies with the cluster notion described by the *nonparametric* or *modal* formulation of the clustering problem (see Menardi, 2016, for a review), which we shall use to borrow some operational tools to identify groups. Note that a similar idea is implicitly acknowledged by one of the most widespread approaches to clustering categorical data, i.e. the $k-$modes (Huang, 1998).

2 Method

According to the nonparametric formulation of the clustering problem, groups are intended as the domains of attraction of the modes of the density underlying data. In the continuous setting, such regions are operationally identified either as the set of points whose direction of the steepest gradient ascent path converges to the same mode, or as connected sets with density above a threshold.

In the categorical setting, defining both a density or its gradient is precluded and, at the same time, there is no obvious method to define the connectedness of a region. Nevertheless, we shall define a procedure that jointly extends both these ideas, if not formally, at least conceptually. To this aim, we build a directed weighted graph where each node represents a cross-category. The idea of steepest gradient ascent path is translated into a sequence of links between nodes driving in the direction of the node with the locally highest (estimated) probability. On the other hand, the connectedness of a region, intended as a set cross-categories, is evaluated by the weight of the links.

Consider again the example in Figure 1: two nodes identified by the cross-categories (A_r, B_c) and (A_s, B_c) shall be considered as highly connected not only because they share the same level for variable B but also when, given that $B = B_c$, both A_r and A_s become more likely, that is, when

$$\frac{P(A_r|Bc)}{P(A_r)} \quad \text{and} \quad \frac{P(A_s|B_c)}{P(A_s)} \tag{1}$$

are high. This results in providing the link with a weight set to the minimum between the probabilities (1), or, for example, to their mean, a choice selected hereafter to avoid ties. The direction of the link will point toward the maximum between the two probabilities. In fact, note that

$$\frac{P(A_r|Bc)}{P(A_r)} < \frac{P(A_s|B_c)}{P(A_s)} \Leftrightarrow \frac{P(A_r,Bc)}{P(A_r)P(B_c)} < \frac{P(A_s,B_c)}{P(A_s)P(B_c)},$$

that is, the path of each node moves toward the direction where the ratio between the joint probability of the cell and the expected probability under the hypothesis of independence is the maximum.

With this toolkit at hand, clusters can be formed as high density upper-level sets or, at the same time, as domains of attractions of the density modes, where the concept of density is here intended as a measure of how each cross-category occurs more frequently than it would do if the variables were independent. The outlined ideas easily extend to an arbitrary number of variables.

3 Application

Figure 2 outlines a synthetic illustrative example that cross-classifies 460 individuals according to their religion and geographic area of origin. The two

	Christianity	Islam	Eastern Religions
Europe	85	4	1
America	110	2	3
Africa	20	30	4
Asia-Pacific	1	120	80
Europa	42.3	30.5	17.2
America	54.0	39.0	22.0
Africa	25.4	18.3	10.3
Asia-Pacific	94.4	68.2	38.5

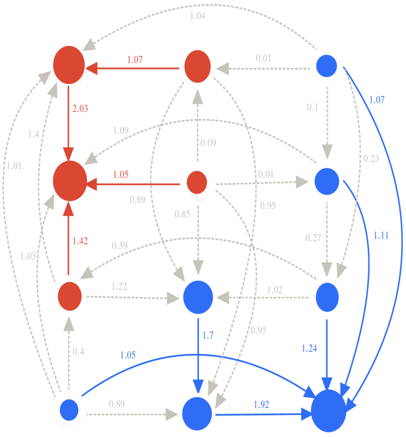


Figure 2. *Contingency table of religion and geographic area of origin for 460 observations (top), associated table expected under the hypothesis of independence (bottom), and graph built based on the proposed method, with highlighted the resulting clusters in different colors.*

variables exhibit a high dependency structure, with certain cross-categories presenting a higher frequency than expected under the assumption of independence. We built the graph according to the presented procedure, by estimating the involved joint probabilities by their empirical counterpart and the expected ones under the hypothesis of independence, by a suitable log-linear model. The graph, displayed on the right side of Figure 2 reports the direction addressed by the nodes, along with the intensity of the connections between them (shaded colors describe outgoing links whose weight is not the highest). While, for example, the cross-categories in the first column share a common level for the religion variable, they are not connected with the same strength, because being Christian increases the probability of coming from America and Europe, whereas the same does not apply to the Asia-Pacific region. By following the path of each node, two clusters are revealed, attracted by subjects of Asiatic origin of Eastern religions and by Christians from America.

References

Huang, Z. 1998. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data min. Know. Disc.*, **2**, 283–304.

Menardi, G. 2016. A review on modal clustering. *Int.Stat.Rev.*, **84**, 413–433.