SYRTO

Systemic Risk Tomography
Signals, Measurements, Transmission channels and Policy Interventions

Finding number of groups using a penalized internal cluster quality index

Marica Manisera, Marika Vezzoli

SYRTO WORKING PAPER SERIES
Working paper n. 9 | 2013

This project is funded by the European Union under the 7th Framework Programme (FP7-SSH/2007-2013) Grant Agreement n° 320270

Finding number of groups using a penalized internal cluster quality index

Marica Manisera and Marika Vezzoli

Abstract In cluster analysis, the identification of number of groups is a non trivial question. Many papers investigated this issue and several criteria have been introduced. The objective of this study is to propose a new method that automatically identifies the optimal number of groups in a hierarchical cluster algorithm. Starting from the idea of pruning, introduced in the context of classification and regression trees, we propose to use a penalized internal cluster quality index in order to identify the best cut in the dendrogram able to provide a partition easily interpretable. In this paper, we show the results obtained by applying our procedure on simulated data with known structure.

Key words: hierarchical cluster analysis, internal cluster quality index, optimal number of clusters, penalized score function, pruning

1 Introduction

Identifying the optimal number of groups is of central importance in cluster analysis ([3]). Many authors handled this issue by exploring several criteria (among others, see [2], [4]), based on the trade-off between a low inter-cluster similarity and a high intra-cluster similarity. It is reasonable to expect that the optimal partition using this trade-off criterion is obtained when the number of groups equals the number of subjects analyzed. However, this type of partition is useless.

The objective of this study is to propose a new method that automatically identifies the optimal number of groups k^* in a hierarchical cluster analysis. In detail, we

Marica Manisera
University of Brescia, c.da S. Chiara, 50, 25122 Brescia, Italy, e-mail: manisera@eco.unibs.it

Marika Vezzoli
University of Brescia, viale Europa, 11, 25123 Brescia, Italy, e-mail: marika.vezzoli@med.unibs.it

want to overcome the subjective cutting of the dendrogram, which appears to be a common choice in practice.

We identify k^* by optimizing an internal cluster quality index, penalized by the number of clusters in order to take account of the interpretability of the resulting groups. This idea was inspired by the pruning ([1]), used in classification and regression trees as a method to avoid the overfitting problem that, in the extreme case, arises when each leaf of the tree contains only one subject. Indeed our idea is conceived along the same line and has the same aim: in a hierarchical cluster analysis, we grow the dendrogram by imposing a penalty depending on the k number of clusters so as to stop the procedure up to identify a reduced, and therefore interpretable, number of groups. In this study, we penalize the intrinsic index proposed in [2], suitable for quantitative data. However, a penalization can be proposed on alternative cluster quality indices, whenever they show a behaviour leading to choose k^* equal to the number of subjects in the analysis (for example, the error measure W_k in [4]).

The paper is organized as follows. Section 2 describes the proposed penalized internal quality index. Section 3 shows an illustrative example based on simulated data with known structure discussing results and future research.

2 Methodology

Starting from the $n \times p$ data matrix $\mathbf{X}$ with n subjects and p quantitative variables, cluster analysis aims at partitioning subjects into k clusters. Many criteria identify the optimal number k^* of groups on the basis of the trade-off between a low inter-cluster similarity and a high intra-cluster similarity, where similarity is usually defined starting from a chosen distance function. In this study, we focus on the Calinski and Harabasz (CH) index ([2]), suitable for quantitative data, which measures the internal cluster quality for a given k as

$$\mathrm{CH}(k) = \frac{\mathrm{BGSS}/(k-1)}{\mathrm{WGSS}/(n-k)}.$$

WGSS (Within-Group Sum of Squares) summarizes the intra-cluster similarity and is given by $\mathrm{trace}(\mathbf{W})$, where $\mathbf{W}$ is a $k \times k$ matrix whose generic element $\{w_{ht}\}_{h,t=1,\cdots,k}$ is the distance of the subjects belonging to group h from the centroid $\mathbf{c}_t$ of group t ($\mathbf{c}_t$ is a p-dimensional vector containing the means $m_j^{(t)}$, $j = 1, \cdots, p$, computed on subjects of group t). BGSS (Between-Group Sum of Squares) summarizes the inter-cluster similarity and is given by $(\mathrm{trace}(n\Sigma) - \mathrm{WGSS})$ where Σ is the variance-covariance matrix of $\mathbf{X}$. When Euclidean distance is used, the concepts of inter-cluster and intra-cluster similarities are related to the dispersion between and within the clusters in the analysis of variance. The best k is given by $k^* = \underset{k=2,\cdots,n-1}{\mathrm{argmax}}\ \mathrm{CH}(k)$. Whenever CH increases as k increases, the optimal partition is expected for $k^* = n$. However, this result is useless and does not comply with the aim of a cluster analysis. In order to identify an interpretable partition, k should be

reasonably small and this is commonly achieved by subjective choices. In hierarchical clustering, that is the focus of this study, this corresponds to a subjective cutting of the dendrogram. In order to avoid such arbitrariness, we propose to identify k^* as:

$$k^* = \underset{k=2,\cdots,n-1}{\operatorname{argmax}} \; \mathrm{Q}(k|\lambda). \tag{1}$$

where $\mathrm{Q}(k|\lambda) = \mathrm{CH}(k) - \lambda \times k$ is obtained by introducing the penalty $\lambda \in \mathbb{R}_+$ on the number k of groups, in order to keep k^* reasonably small and find it automatically (equation (1) holds for internal cluster quality indices, alternative to CH, that have to be maximized and increase with k). If $\{0\}$ is included in the domain of λ, for $\lambda = 0$ we have $\mathrm{Q}(k|\lambda) = \mathrm{CH}(k)$ and no penalization is imposed. The larger the values of λ, the stronger the penalty (and *viceversa*). The effect of a fixed λ on k^* depends on the magnitude of the chosen cluster quality index.

3 An illustrative example

We applied the proposed procedure on an artificially generated data described in [5] and referred to 5 interval-type variables on 75 subjects clustered into 5 groups. We performed a hierarchical cluster analysis, using the `hclust` function in `R`, with complete linkage. The cluster dendrogram, shown in Figure 1 (left), does not provide strong evidence for the simulated 5-class structure, which is instead confirmed by CH in [5], but only because the authors considered small k's ($k = 2,\cdots,6$, see p. 12). We replicated their procedure computing CH for $k = 2,\cdots,50$, as shown in Figure 1 (right). CH increases as k increases and is maximized for $k = 50$ (see the black circle in the right part of Figure 1). However, if we want a partition with an interpretable number of groups, CH suggests $k^* = 8$. This choice requires the inspection of the dendrogram and a subjective reasoning by the researcher.

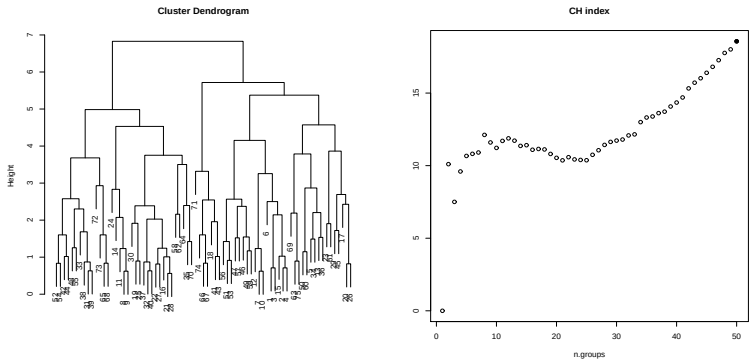

Fig. 1 Cluster dendrogram (left) and CH index for $k = 2,\cdots,50$ (right) - artificial data

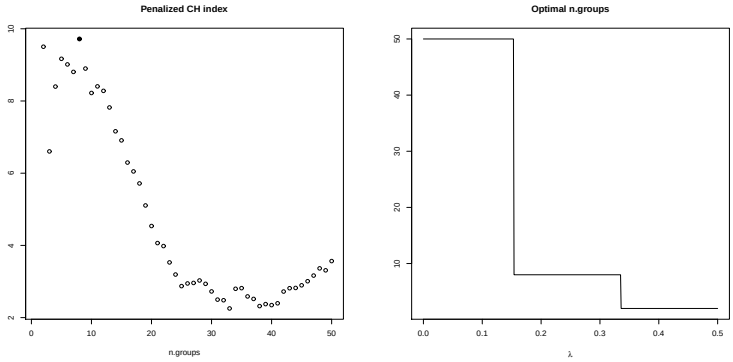

Fig. 2 Penalized CH index for $k = 2, \cdots, 50$ and $\lambda = 0.3$ (left) and optimal number k^* of groups based on the penalized CH index for $\lambda = 0, \cdots, 0.5$

Instead, the use of the penalized CH index makes the choice automatic. Figure 2 (left) shows the penalized CH index computed for fixed $\lambda = 0.3$ and $k = 2, \cdots, 50$. The penalization introduced in the cluster quality index automatically leads to the choice of the 8-class structure as the best solution (see the black circle in the left part of Figure 2). The maximization of the penalized CH index identifies 8 as the best number of groups for a wide range of fixed λ ($[0.154, 0.335]$), while outside that range the best number of groups is always 2 or 50, as shown in Figure 2 (right). Therefore, in this data set we can fairly choose $k^* = 8$.

Concluding, results obtained in this illustrative example show that the proposed procedure is able to reach the objective of automatically identifying the best number of clusters in a data set by taking account of the interpretability of the resulting groups. Current research is being devoted to refine the optimization algorithm, especially with reference to the choice of λ. Simulation studies and the analysis of real data sets, involving several internal cluster quality indices suitable for different data types, could confirm the validity of our proposal.

Acknowledgments: This research was partially funded by a grant from the European Union Seventh Framework Programme (FP7-SSH/2007-2013); *'SYstemic Risk TOmography: Signals, Measurements, Transmission Channels, and Policy Interventions' - SYRTO* - Project ID: 320270.

References

1. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and regression trees. Wadsworth International Group, Belmont(1984)
2. Calinski, T., Harabasz, J.: A dendrite method for cluster analysis. Commun. Stat. **3**, 1–27 (1974)
3. Everitt, B.S., Landau, S., Leese, M., Stahl, D.: Cluster Analysis. Wiley, Chichester (2011)
4. Tibshirani, R., Walther, G., Hastie, T.: Estimating the number of data clusters via the Gap statistic. J. Roy. Stat. Soc. B **63**, 411-423 (2001)
5. Walesiak, M., Dudek, A.: clusterSim: Searching for optimal clustering procedure for a data set. R package version 0.41-8. http://CRAN.R-project.org/package=clusterSim (2012)