ID de aportación : **50** Tipo: **Oral**

max_clique: the project and development of a new package for robust clustering

The clique partitioning problem is a classic integer linear programming model: Problem data are a set of units U, characterized by a relation of similarity/dissimilarity d(i,j) between every pair of units i,j in U. Set U must be partitioned in such way that the sum of the similarities of the pairs (i,j) that are in the same group is as high as possible. When we compare the clique partitioning problem to other computational tools for clustering, like the k-means for example, we can recognize some advantages.
1) Clique partitioning can be defined for both qualitative and quantitative data, as only an appropriate similarity measure must be defined. Conversely, the k-means, requiring averages, can work only with quantitative data.
2) The number of clusters is not defined by the user, as it is a problem output. The clique model itself selects the optimal group of clusters. Conversely, the k-means is usually run with varying values of k, then the correct k is selected by some rule-of-thumb.
3) Consequence of point 2, outliers are detected by singletons of the partition. Conversely, outliers of the k-means model can be detected only by some preliminary analysis.
4) The clique partitioning model is flexible enough to be used for community detection too. That is, a clustering model in which units are connected by links. In this framework, the model is called modularity maximization.
5) Clique partitioning and k-means are both NP-complete. However, clique is integer programming, while k-means is non-convex programming. It is easier to find the optimal solution of an ILP, rather than to a non-convex problem.
Even though there are many reasons to prefer a clique model to a k-means, still the latter is more known and popular. One of the reasons is the lack of a reliable procedure and package for clique partitioning in the most important platforms: R, Julia or Python. In our communication, after revising the combinatorial and statistical properties of the clique partitioning model, we will illustrate our computational experience with it and describe the structure of new R package, under development for our research projects. Our package is composed of heuristic and exact procedures, some new and some already known in the literature. Nevertheless, an original package feature is the use of the set partitioning formulation of the clustering model. This feature is a unifying framework that can be used to decompose various models into a master and a pricing problem. While the master problem is the same for many clustering models, pricing is peculiar. It implies that users can test new clustering model and interact with the package defining only their own pricing problem.

¿Presentas la comunicación a premio?

Afiliación (del autor)

Dipartimento di Sociologia e Ricerca Sociale, Università di Trento

Autor primario: BENATI, Stefano (Dipartimento di Sociologia, Università di Trento)

Coautor: Prof. AVELLONE, Alessandro (Dipartimento di Statistica e Metodi Quantitativi)

Clasificación de pistas: Matemáticas