The Variable Selection Problem for Clustering: An Integer Linear Programming approach

Stefano Benati
Sergio García Quiles
Justo Puerto

Finding structure in high-dimensional data is a difficult task that is even harder if data contains variables with no relevant information. If those variables are not detected and discarded form the analysis, the sequent analysis can be blurred or biased by their presence. The problem becomes more and more important as the the number of variables increases: Nowadays standard data base can comprise hundredths and even thousands of covariates, and therefore researchers from different disciplines need tools to detect and discard what are called noisy or masking variables.

There are three ways to address the problem of the masking variables. The most simple relies on peculiar indexes to detect variables that do not show evident patterns. The methodology begins with applying a form of standardization to data, calculate a clusterability index, add variables to data set in a greedy fashion, finally apply a clustering algorithm like k-means to the reduced data set. Conversely, the most sophisticated methodology assumes a multivariate probability distributions, whose parameters must be estimated through applications of the EM algorithm. One EM iteration complexity is at most quadratic, and numerical stability is an issue, therefore there are application in which this methodology is unreliable.

As the third way, some authors proposed a mathematical programming approach to detect masking variables. Here we present our new methodology, that is based on the formulation of the problem as Integer Linear Programming. We formulate the problem of the variable selection as follows: A set $U = \{1, \ldots, n\}$ of statistics units and a set $R = \{1, \ldots, r\}$ of cluster centers, or prototypes, are given. For every $i \in U$ and every $j \in R$, a set $V = \{1, \ldots, m\}$ of statistics variables are measured, so that the for every triple $i \in U, j \in R, k \in V$, a distance d_{ijk} is calculated, that is the difference measured through variable k between the statistic unit i and the cluster center j. If a subset $Q \subset V$ of variables is selected, the i, j distance using Q is the sum: $d_{ij}(Q) = \sum_{k \in Q} d_{ijk}$. Distances Q are used to determine the partition of U into clusters $G_j, j = 1, \ldots, r$: For a fixed $Q \subseteq V$, a unit i is assigned to cluster $j(i)$ if $d_{i,j(i)}(Q) = \min\{d_{ij}(Q) | j = 1, \ldots, r\}$. The total distance between units and clusters is the sum: $D(Q) = \sum_i d_{i,j(i)}(Q)$. The Variable Selection Problem is to select the subset $Q \subseteq V$ for which the objective

function $D(Q)$ is minimized.

We show that the problem is NP-hard, but we found two Integer Linear Programming formulations that allow the application of standard software as Cplex or LpSolve to its solution, and the two formulations are compared in term of time and quality of their linear relaxation. We developed two heuristic algorithms that are based on the ILP formulations, that resemble to two approaches to p-median problem, one is the selection-allocation heuristic, one is the add-swap interchange heuristic. Again, both are compared in term of time and solution quality. The main feature of a possible Lagrangian relaxation heuristic will also be sketched. Finally, we show how the model can be plugged-in to classic clustering algorithms like the k-mean, the p-median, or to EM model-based clustering to reduce the data dimension and improve the quality of the clustering. Finally, we show how the model can be applied to data coming from the World Value Survey.