Cluster analysis: theory, methods and applications

Plenary lecture

Rudolf Scitovski
Department of Mathematics
University of Osijek
scitowsk@mathos.hr

The given set of data $\mathcal{A} = \{a_i \in \mathbb{R}^n : i = 1,\dots,m\} \subset \mathbb{R}^n$, $|\mathcal{A}| = m \gg n$ should be partitioned into $1 \le k \le m$ nonempty disjoint subsets $\pi_1,\dots,\pi_k$. By introducing some distance-like function $d\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$, $\mathbb{R}_+ = [0, +\infty\rangle$, to each cluster π_j we can associate its center

$$c_j = c(\pi_j) := \operatorname*{argmin}_{x\in\operatorname{conv}(\pi_j)} \sum_{a_i\in\pi_j} d(x, a_i), \quad j = 1,\dots,k.$$

In this way, the optimization problem of searching for d-optimal partition $\Pi^\star = \{\pi_1^\star,\dots,\pi_k^\star\}$ on the set of all partitions $\mathcal{P}(\mathcal{A}; m, k)$ with objective function $\mathcal{F}\colon \mathcal{P}(\mathcal{A}; m, k) \to \mathbb{R}_+$

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a_i\in\pi_j} d(c_j, a_i),$$

can be established. This problem is equivalent to the following center-based clustering problem

$$\min_{z_1,\dots,z_k\in\mathbb{R}^n} \Phi(z_1,\dots,z_k), \qquad \Phi(z_1,\dots,z_k) = \sum_{i=1}^{m} \min_{1\le j\le k} d(z_j, a_i),$$

where $\Phi\colon \mathbb{R}^{kn} \to \mathbb{R}_+$. This is a global optimization problem, where the objective function Φ can have a great number of independent variables, it does not have to be either convex or differentiable and generally, it may have several local minima. Therefore this becomes a complex global optimization problem.

There are various applications of this problem. As an illustration, let us mention only some: detection of spatial and time locations of the center of seismic activities, forecast of hourly natural energy resources consumption, acceptable definition of constituencies, grouping students based on their academic records.

The choice of the distance-like function d depends on the nature of the problem. Especially in this lecture, we will talk about the applications of the Least Squares (LS) and the Least Absolute Deviations (LAD) distance-like functions. In case of LS-clustering, the stability problem and the problem of some data point occurring on the Voronoi diagram is considered. In case of LAD-clustering, an efficient method for searching for the locally optimal partition will be shown.

Finally, since the objective function Φ is a symmetric function, an efficient method for searching for the globally optimal partition as a special case of the well-known `DIRECT`-method will be mentioned.

MSC2010: 68T10, 62H30, 91C20, 90C26.

Keywords: Data mining, Clustering, k-means.

Section: 16,17.