mgr Marek Śmieja[1]
Instytut Informatyki Wydziału Matematyki i Informatyki
Uniwersytetu Jagiellońskiego

Spherical Clustering in Metric Space

One of the most challenging problems of clustering is to create an efficient method which can be applied in general metric spaces and groups data with respect to specified families of probability distributions. The main difficulty is the fact that metric space does not introduce a vector structure but only provides the distances between particular elements.

Several attempts were undertaken to overcome this problem. General Wards approach allows to replace the expressions containing mean m of data-set X by rewriting:

$$\sum_{x \in X} \|x - m\|^2 = \frac{1}{2} \sum_{y \in X} \sum_{x \in X} \|x - y\|^2.$$

On the other hand, kernel methods introduce a Gaussian model for the case inner product space. However, the greatest challenge is to determine the proper value of kernel radius.

Our approach relies on applying the cross-entropy clustering method (CEC) which is a new alternative for classical methods to the clustering with respect to the families of probability distributions as EM (Expectation Maximization). Let us assume that we divide data-set into groups $\mathcal{U} = (U_1, \ldots, U_n)$ where each one is represented by the optimal spherical Gaussian distribution. For this reason CEC minimizes the total cross-entropy between $\mathcal{U}$ and the family of spherical Gaussian distributions:

$$E(\mathcal{U}) = \sum_{i=1}^{n} (\# U_i) \cdot \left(-\log(\# U_i) + \frac{N}{2} \log \left(\frac{\sum_{x \in U_i} \|x - m_i\|^2}{\# U_i} \right) \right),$$

where N denotes the dimension of data-set. We use Wards approach to eliminate the means from the above expression. Another advantage of our method is that the correct number of clusters can be computed automatically with use of modified Hartigan approach.

[1] Praca naukowa finansowana ze środków budżetowych na naukę w latach 2013–2015, nr projektu: IP2012 055972.