Katarzyna Stąpor, PhD DSc
Institute of Computer Science, Silesian University of Technology
Adrian Brueckner, MSc
Institute of Mathematics, Silesian University

Clustering in Hilbert space: kernel K-Means algorithm and its application in ophthalmology

This paper presents the kernel K-Means clustering algorithm and its application in the segmentation of cup region in digital fundus eye image (FEI). This is the first stage of the proposed automatic method supporting glaucoma diagnosis in ophthalmology. The remaining two stages of this method comprise cup feature selection based on genetic algorithms and classification using support vector machine (SVM) classifier.

Traditional K-Means clustering algorithm aims to partition the data set composed of N samples $x_1, \ldots, x_N$ into K clusters: $G_1, \ldots, G_K$, and then returns the center of each cluster: $c_1, \ldots, c_K$ as the representatives of the data set. The assumption behind the above algorithm is the belief that the data space consists of isolated elliptical regions. However, such assumption is not always held on specific applications. To tackle this problem, one idea is to apply a transformation $\Phi : R^d \to Q$, that maps each data x_i from the input space R^d to a new space Q, being a Hilbert space, where the given algorithm can be used. Such transformation is done implicitly by means of a kernel function k, satisfying:

$$k(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j) \tag{1}$$

for example a Gaussian one: $k(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{2r^2})$. Mercer's theorem guarantees that as long as the kernel function is positive definite, the algorithm implicitly operates in a higher dimensional space. This kernel trick saves the algorithm from the computational expense of explicitly representing all of the features in a higher-dimensional space.

The key issue extending traditional K-Means clustering algorithm to kernel K-Means is the computation of the Euclidean distance between $u_i = \Phi(x_i)$ and t_k, the cluster center in the transformed space:

$$D^2(u_i, t_k) = k(x_i, x_i) + h_1(x_i, G_k) + h_2(G_k), \tag{2}$$

where

$$h_1(x_i, G_k) = -\frac{2}{|G_k|} \sum_{j=1}^{N} \omega(u_j, G_k) k(x_i, x_j), \tag{3}$$

$$h_2(G_k) = \sum_{j=1}^{N} \sum_{l=1}^{N} \omega(u_j, G_k) \omega(u_l, G_k) k(x_j, x_l), \tag{4}$$

$$\omega(u_i, G_k) = \begin{cases} 1 & \text{if } \forall j \neq k \;\; h_1(x_i, G_k) + h_2(G_k) < h_1(x_i, G_j) + h_2(G_j), \;\; j = 1, \ldots, K, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

Since the cluster center in a transformed space cannot be expressed explicitly, we have to choose a pseudo center instead, for example the sample that is closest to the center:

$$c_k = \operatorname*{argmin}_{x_i : \omega(u_i, G_k) = 1} D(\Phi(x_i), t_k). \tag{6}$$

In the second stage, genetic algorithms are used to select the most significant features characterizing the shape of the segmented cup region. The last stage is the training and testing procedure of SVM classifier with Gaussian kernel.

The following results were obtained on the set composed of 200 segmented FEI: mean sensitivity: 93% and mean specificity: 97%.