The Global Kernel k-Means Algorithm for Clustering in Feature Space
Grigorios Tsotras and Aristidis Likas
Department of Computer Science, University of Ioannina

Abstract
To overcome the cluster initialization problem associated with the popular kernel k-means algorithm [1], we propose the global kernel k-means algorithm that:
• Constitutes a deterministic and incremental approach to kernel-based clustering.
• Locates near optimal and nonlinearly separable solutions, avoiding poor local minima.
• Does not depend on cluster initialization.

Method
Global kernel k-means maps the instances \( X = \{x_1, x_2, \ldots, x_N\}, x_i \in \mathbb{R}^{d} \) from input space to a higher dimensional feature space through a nonlinear transformation \( \phi \) (in practice \( \phi \) is implicitly defined by a kernel function \( K(x_i, x_j) = \phi(x_i) \phi(x_j) \) and the corresponding kernel matrix) and minimizes the clustering error in feature space:

\[
E = \sum_{i=1}^{N} \sum_{j}^{C} \left| \phi(x_i) - \phi(c_j) \right|^2 = \sum_{i=1}^{N} \sum_{j}^{C} K(x_i, c_j)
\]

Deviating a partitioning with \( M \) clusters, global kernel k-means deterministically solves all intermediate problems with \( 1, \ldots, M \) clusters:
• The optimal solution to the 1-clustering problem is trivial, as all data points are assigned to the same cluster.
• For the \( k \)-clustering problem let \( \{ C_1, \ldots, C_k \} \) denote the solution with \( k \)-clusters and assume that \( x_i \in C_j \). We perform \( k \) executions of the kernel k-means algorithm, with initial clusters for the \( n \)-th run being \( \{ C_1, \ldots, C_n \} \), and keep the one resulting in the lowest clustering error as the solution with \( k \) clusters.
• The above procedure is repeated until \( k = M \).

The rationale behind the proposed method is based on the assumption that a near optimal solution with \( k \) clusters can be obtained through local search, from an initial state with \( k = 1 \) near optimal clusters and the \( k \)-th cluster initialized appropriately.

Extensions
Two schemes are developed to accelerate global kernel k-means:
• The fast variant, for each intermediate problem, locates the dataset point that guarantees the greatest reduction in clustering error and executes kernel k-means only once from this initialization.
• The CMM variant, locates a set of exemplars in the dataset, by fitting a convex mixture model [2], and tries only the exemplars as possible initializations for the newly added cluster. Also, the proposed algorithms are extended to handle weighted data points, which enables their application to graph partitioning in place of the spectral methods [3].

Indicative Results
The methods are extensively tested on several datasets, ranging from artificial data to handwritten digits, face images, graphs, and MRI images. The results show that global kernel k-means and its variants compare favorably to kernel k-means with random restarts.

References

Document Clustering Using Synthetic Cluster Prototypes
Argyris Kalogeratos and Aristidis Likas
Department of Computer Science, University of Ioannina
Data & Knowledge Engineering, vol. 70, pp. 384–396, 2011

Abstract
Centroid prototypes is not always the best choice for text clusters for the k-means family of clustering methods due to the high dimensionality and sparsity of text data. We propose:
• synthetic cluster prototypes computed by first selecting a subset of cluster objects, then computing a representative for these objects and finally selecting important features.
• a robust clustering method called k-synthetic prototypes (k-sp) that incorporates synthetic prototypes to the generic spherical k-means procedure (spk-means).

Method
Each text document \( d \) is represented as a \( |d| \)-dimensional vector, where \( |d| \) the vocabulary of the unique terms.
• Synthetic cluster prototypes implement a dynamic selection scheme (Fig. 1). The computation requires the definition of
  1. a reference prototype, an initial representative of the cluster constructed by a subset of its objects, and
  2. a feature selection procedure on prototypes in order to select features from the reference cluster prototype.
• As clustering proceeds the information progressively produced in the formed clusters is further exploited.

Spk-means is a special case of k-sp where all cluster objects are selected and no terms are pruned. Thus, we propose an optional self-refinement strategy after the basic k-sp phase:
• A k-sp with Centroids is used to fine-tune the formed clusters produced with MedoidKNN.
• It assists in reducing the sensitivity of k-sp to its parameter definition.
• Helps in choosing the best clustering solution among those obtained for different k-sp parameter settings by comparing the values of the objective function after the refinement.

Indicative Results
Comparative experimental evaluation of k-sp against spk-means (various initialization techniques were tested), spectral clustering [1], hierarchical agglomerative and soft subclustering (k-means like) clustering methods, ewk-means [2] and fswk-means [3], shows its superior performance in artificial and real-world text datasets (e.g. 20-Newsgroups).

References