Abstract

Data clustering is an unsupervised learning activity which is a process of finding natural groups (clusters) present in the given dataset (i.e., the given set of patterns). It has several applications, like image segmentation, video analysis, bio-informatics, intrusion detection, outlier detection, etc. New application domains and amassed data poses new challenges in the area of data clustering. Different types of data clustering methods have been evolved to cater these upcoming challenges.

Among the existing clustering methods, the simplest and efficient clustering method is the k-means clustering method. It has been shown to produce good clustering results in various applications. The time complexity of the k-means method linearly grows with respect to the size of the dataset. In the iterative process of the k-means method, the entire dataset has to be scanned once in each iteration, which is a time consuming process in case of large datasets. Hence, the k-means method is not a suitable one to work with large datasets which do not fit in the main memory. Further, the method fails in identifying non-convex shaped and linearly inseparable clusters in the input space.

The kernel k-means clustering method is a nonlinear extension of the k-means method. By implicitly mapping data points to a higher-dimensional feature space (induced space) using a non linear transformation, the kernel k-means method can discover clusters that are linearly inseparable in the input space. But, the time complexity of this method grows quadratically with respect to the size of the dataset. Hence, the kernel k-means clustering method is also not a suitable one for large datasets. The present thesis is about speeding-up the k-means and kernel k-means clustering methods to work with large datasets.

In order to speed-up the k-means method, the thesis proposes two prototype based
hybrid approaches, which give the same result as that obtained by using the conventional k-means method. Similarly two prototype based methods are proposed to speed-up the kernel k-means method. But, same result as the conventional kernel k-means method is possible only in some favorable cases, which is theoretically shown. Empirically, it is shown to produce a similar result (that is, a very close result as that of the conventional method) using several datasets. Towards the end of the thesis, a single pass kernel k-means method is presented which is motivated from the existing methods viz., the single-pass k-means method and the two-step kernel k-means method. This is much faster, but, at the same time, could produce a much deviated result (when compared to the above mentioned prototype based methods).

The key idea of the proposed prototype based hybrid approaches is to reduce the size of the dataset that is used with k-means or kernel k-means clustering method, by selecting some prototypes instead of the entire dataset. Here, each prototype is a representative pattern of a grouplet (a small cluster) in the dataset. The overall scheme of the proposed hybrid approaches is as follows. First, the dataset is partitioned into a finite number of grouplets using a fast clustering method viz., the leaders clustering method. Later, these grouplets are again partitioned into k clusters using k-means or kernel k-means method to get a partition of the entire dataset.

The leaders clustering method is used to derive a set of grouplets from the dataset because it is a single scan and fast clustering method (takes linear time with respect to the size of the dataset). It takes a threshold value, which is the size of the grouplet, as an input parameter and partitions the dataset into a set of grouplets. The leaders clustering method is modified such that for the given threshold value, the grouplets are formed either in the input
space or in the kernel induced feature space in order to work with k-means or kernel k-means clustering method, respectively. Further, the leaders method is modified such that without the need of any input parameter, the grouplets of different sizes are formed either in the input space or in the kernel induced feature space.

The methods proposed in the thesis are as follows.

1. \textit{lk}-means-CMFT: Uses prototypes (leaders) derived by using a fixed threshold value (in other words, grouplets derived are of same size).

2. \textit{lk}-means-CMVT: Uses leaders of varying thresholds \textit{(i.e.,} grouplets are of various sizes).

3. Kernel-\textit{lk}-means-CMFT: Prototype based kernel k-means method that uses leaders of fixed threshold. Note, leaders are represented in the input space, but, grouplets in the kernel induced feature space.

4. Kernel-\textit{lk}-means-CMVT: Similar to above, but uses leaders of various thresholds.

5. Single-pass kernel k-means method: Scans the dataset only once and works similar to an existing method called, the two-step kernel k-means method.

One of the important contributions of the thesis is that to find a representative pattern (in the input space) for the cluster center in the kernel induced feature space, a gradient descent method is proposed. This is in contrast to the approximate cluster center which is a pattern in the dataset that is close to the center, as mentioned in some of recent similar methods.

The effectiveness of the proposed methods are shown both theoretically and experimentally. The proposed methods are scalable, hence, are suitable to work with large datasets like those in data mining applications.