Clustering Data Streams Based on Shared Density between Micro-Clusters

As more and more applications produce streaming data, clustering data streams has become an important technique for data and knowledge engineering. A typical approach is to summarize the data stream in real-time with an online process into a large number of so called micro-clusters. Micro-clusters represent local density estimates by aggregating the information of many data points in a defined area. On demand, a (modified) conventional clustering algorithm is used in a second offline step to re-cluster the micro-clusters into larger final clusters. For re-clustering, the centers of the micro-clusters are used as pseudo points with the density estimates used as their weights. However, information about density in the area between micro-clusters is not preserved in the online process and re-clustering is based on possibly inaccurate assumptions about the distribution of data within and between micro-clusters (e.g., uniform or Gaussian). This paper describes DB_STREAM, the first micro-cluster-based online clustering component that explicitly captures the density between micro-clusters via a shared density graph. The density information in this graph is then exploited for re-clustering based on actual density between adjacent micro-clusters. We discuss the space and time complexity of maintaining the shared density graph. Experiments on a wide range of synthetic and real data sets highlight that using shared density improves clustering quality over other popular data stream clustering methods which require the creation of a larger number of smaller micro-clusters to achieve comparable results.