Clustering Data Streams

Clustering datasets is one of the most important algorithms used in data mining. There are different traditional algorithms used for clustering such as K-means, hierarchical, and PAM clustering algorithms. Those algorithms are useful for datasets that are of fixed size, can fit easily in memory, and do not change through time. Most of the datasets nowadays are of large sizes that require large memory, arrive continuously, and might change over time. These kinds of data require a special kind of algorithms that can handle these special features. Datasets of large size that arrives continuously, in an ordered manner, and are of infinite size are called data streams.

The tutorial explains different algorithms used for clustering data streams, compares them, and provides an example using R tool to show how data stream algorithms work. The tutorial will concentrate on BIRCH and Clustream clustering algorithms.

The reason why I have chosen this topic is because a lot of research is going on this area and researchers are trying to improve the different algorithms used for clustering data streams. Furthermore, there are a lot of datasets available in different fields that require a special tools and algorithms to gather the hidden information within these data and provide useful output. Also, different algorithms have different advantages and disadvantages that could be improved and by discussing this topic, students will have a better understanding about applications of this kind and the reason why they are used. Finally, traditional clustering algorithms are not efficient to apply on many problems so the tutorial will help students know when to use such algorithms and how to simplify data streams to be able to apply clustering on them.