In recent years, the data stream clustering problem has gained considerable attention in the literature. Clustering data streams requires a process capable of partitioning observations continuously while taking into account restrictions of memory and time. In the literature, most of existing algorithms (e.g. StreamKM++ (Ackermann et al., 2012), CluStream (Aggarwal et al., 2003), DenStream (Cao et al., 2006), or ClusTree (Kranen et al., 2011)) divide the clustering process in two phases. In this work we present MBG-Stream¹, a Micro-Batching version of the growing neural gas approach (Fritzke, 1994), aimed to clustering data streams by making one pass over the data. MBG-Stream allows us to discover clusters of arbitrary shapes without any assumptions on the number of clusters (cf. Figure 1). The proposed algorithm is implemented on a “distributed” streaming platform, the Spark Streaming API (Zaharia et al., 2012), and its performance is evaluated on public data sets.

Références


Micro Batch Data Stream Clustering Based on Growing Neural Gas

Figure 1: Evolution of graph creation of MBG-Stream on DS1 (data set and topological result). The intermediate graph after seeing the first window’s data points; the 1/3 of all windows; the 2/3 of all windows; and the final graph.