

# Data Stream Clustering Based on Micro-Batch Growing Neural Gas Using Spark Streaming

Mohammed Ghesmoune, Mustapha Lebbah, Hanene Azzag

LIPN-UMR 7030

University of Paris 13, Sorbonne Paris City – CNRS

99, av. J-B Clément – F-93430 Villetaneuse, France

{firstname.lastname}@lipn.univ-paris13.fr

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-Stream<sup>1</sup>, 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

- Ackermann, M. R., M. Märtens, C. Raupach, K. Swierkot, C. Lammersen, et C. Sohler (2012). *StreamKM++ : A clustering algorithm for data streams*. *ACM Journal of Experimental Algorithmics* 17(1).
- Aggarwal, C. C., T. J. Watson, R. Ctr, J. Han, J. Wang, et P. S. Yu (2003). A framework for clustering evolving data streams. In *In VLDB*, pp. 81–92.
- Cao, F., M. Ester, W. Qian, et A. Zhou (2006). Density-based clustering over an evolving data stream with noise. In *SDM*, pp. 328–339.
- Fritzke, B. (1994). A growing neural gas network learns topologies. In *NIPS*, pp. 625–632.
- Kranen, P., I. Assent, C. Baldauf, et T. Seidl (2011). The ClusTree : indexing micro-clusters for anytime stream mining. *Knowledge and information systems* 29(2), 249–272.
- Zaharia, M., T. Das, H. Li, S. Shenker, et I. Stoica (2012). Discretized streams : An efficient and fault-tolerant model for stream processing on large clusters. In *Proceedings of the*

---

1. This work has been accepted and presented in the INNS Conference on Big Data : Mohammed Ghesmoune, Mustapha Lebbah, Hanene Azzag. Micro-Batching Growing Neural Gas for Clustering Data Streams using Spark Streaming. 8-10 August 2015 - San Francisco, USA.

## Micro Batch Data Stream Clustering Based on Growing Neural Gas

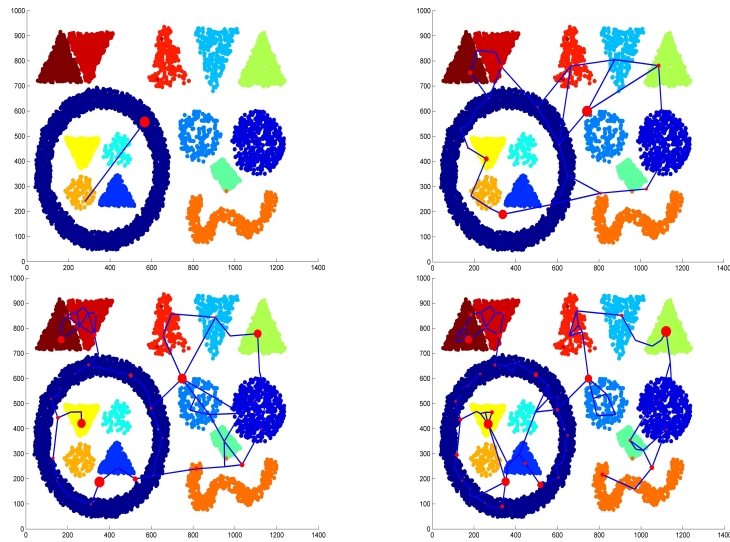


FIG. 1: Evolution of graph creation of MBG-Stream on DS1 (data set and topological result). The intermediate graph after seeing the first window's data points ; the 1/3 of all windows ; the 2/3 of all windows ; and the final graph.

*4th USENIX Conference on Hot Topics in Cloud Computing, HotCloud'12, Berkeley, CA, USA, pp. 10–10. USENIX Association.*