

March 2017

A nature-inspired, anytime and parallel algorithm for Big Data stream clustering

Giandomenico SPEZZANO and Andrea VINCI ¹

*CNR - National Research Council of Italy
Institute for High Performance Computing and Networking (ICAR),
Rende (CS) - Italy— {spezzano, vinci}@icar.cnr.it*

Abstract. In the context of time-critical applications there exists the need of clustering big data streams so as to provide approximated solutions in the shortest possible time, in order to capture in real-time the evolution of physical or social phenomena. In this work, a nature-inspired algorithm for clustering of evolving big data stream is presented, which is designed to be executed on many-core GPU architectures.

Keywords. Clustering, Data Stream, General Purpose GPU Computing, anytime,

1. Introduction

With the successful computerization of most human activities, the amount of available digital data has increased dramatically. This results in a crescent need for tools capable of analyzing large amounts of data (Big Data) so as to extract knowledge, exploitable in areas ranging from market surveys to scientific research.

Research in the field of data mining was developed precisely to address this need, producing techniques and methodologies capable of performing different types of tasks. The present work relates to the clustering task, that is, the discovery of groups of objects as similar as possible to each other and dissimilar from the objects belonging to other groups. Considering, for example, a database that describes the purchases of each customer at a supermarket, it is possible to categorize the latter to propose to each different group of clients a custom marketing campaign.

Besides “static” contexts, such as that described in the previous example, in which cluster analysis is carried out on the basis of data with a low update rate and renewal of information, there are other contexts in which the data is produced at higher rates, such as those generated by wireless sensor networks in Smart Cities, atmospheric sensors, astronomical observations, bank transactions, click on websites flows. In these more dynamic contexts, there is the need for tools which can analyze *evolving data streams*[1], addressing constraints related to scalability and responsiveness which are more restrictive than in the previous contexts.

¹Corresponding Author: Andrea Vinci, PH.D., CNR - National Research Council of Italy - Institute for High Performance Computing and Networking (ICAR), Via P. Bucci, Cubo 7/11c, 87036, Rende (CS) - Italy, email: andrea.vinci@icar.cnr.it