Data Stream Clustering

Elaine R. Faria*
Federal University of Uberlândia, Uberlândia, Brazil – elaine@ufu.br

Jonathan de A. Silva
Federal University of Mato Grosso do Sul, Brazil - jandrade@icmc.usp.br

Rodrigo C. Barros
Pontifícia Universidade Católica do Rio Grande do Sul , Brazil - rodrigo.barros@pucrs.br

Eduardo R. Hruschka
Big Data Brasil, Brazil - eduardo.hruschka@bigdata.inf.br

André C. P. L. F. de Carvalho
University of São Paulo, São Carlos, Brazil, andre@icmc.usp.br

João Gama
University of Porto, Portugal - jgama@fep.up.pt

Data stream mining is an active research area that has recently emerged to discover knowledge from large amounts of continuously generated data. The intrinsic nature of stream data requires the development of algorithms capable of performing fast and incremental processing of data objects, suitably addressing time and memory limitations. Extracting potentially useful knowledge from data streams is a challenge. One of the important tasks in this context is clustering. Essentially, the clustering problem can be posed as determining a finite set of categories (clusters) that appropriately describe a data set. Data stream clustering imposes several challenges to be addressed, such as dealing with non-stationary, unbounded data that arrive in an online fashion. In addition, important issues need to be addressed such as provide a model representation that is not only compact, but also does not grow with the number of objects processed; rapidly detect the presence of outliers and act accordingly; and deal with different data types, e.g., XML trees, DNA sequences.
The goal of this talk is to present the state-of-the-art in clustering data streams. We will present the main research questions and an overview of some important clustering algorithms for data streams. We will provide a taxonomy that allows to identify every work with respect to important aspects in data stream clustering. We will also present the experimental methodology usually employed in the literature as well as the main data sets and frameworks available. A series of reference and possible applications will be presented. Finally we will indicate challenges to be faced and promising future directions for the area. We believe that clustering data streams is a relevant and challenging research area in which much effort should be addressed in order to improve it.

Keywords: data stream; clustering; data mining.