DATA STREAM MINING FOR UBIQUITOUS ENVIRONMENTS

João Gama
LIAAD-INESC TEC, and FEP, University of Porto
R. Ceuta 118-6; 4050-190 Porto; Portugal
Tel: +351 223392094
e-mail: jgama@fep.up.pt

1 INTRODUCTION

In the data stream computational model examples are processed once, using restricted computational resources and storage capabilities. The goal of data stream mining consists of learning a decision model, under these constraints, from sequences of observations generated from environments with unknown dynamics. Most of the stream mining works focus on centralized approaches. The phenomenal growth of mobile and embedded devices coupled with their ever-increasing computational and communications capacity presents exciting new opportunities for real-time, distributed intelligent data analysis in ubiquitous environments. In domains like sensor networks, smart grids, social cars, ambient intelligence, etc. centralized approaches have limitations due to communication constraints, power consumption, and privacy concerns. Distributed online algorithms are highly needed to address the above concerns. The focus of this presentation is on distributed stream clustering algorithms that are highly scalable, computationally efficient and resource-aware. These features enable the continued operation of data stream mining algorithms in highly dynamic mobile and ubiquitous environments.

2 DISTRIBUTED CLUSTERING

One of the most popular knowledge discovery techniques is clustering, the process of finding groups in data such that data objects clustered in the same group are more alike than objects assigned to different groups [1]. On top of clustering algorithms, several tasks can be computed: profiling, anomaly and event detection, outliers detections, trends, deviations, etc. Sensor networks, smart grids, social cars are paradigmatic examples of ubiquitous streaming data sources. The quality of these clusters is usually called clustering validity, and can be measured in several different ways [2]. But classical methods tend to become obsolete for application in streaming and (especially) ubiquitous settings, due to their high time and space complexity. Hence, new machine learning algorithms are being developed to cope with this new demanding scenario, and different quality indices are being considered (e.g. computation and communication load). There are two different clustering problems in ubiquitous and streaming settings: clustering sensor streams and clustering streaming sensors. The former problem searches for dense regions of the data space, identifying hot-spots where sensors tend to produce data, while the latter finds groups of sensors that behave similarly through time [3]. In the first setting, a cluster is defined to be a set of data points (Figure 1) generated by multiple sources. In the second setting a cluster is defined to be a set of sensors (Figure 2) .

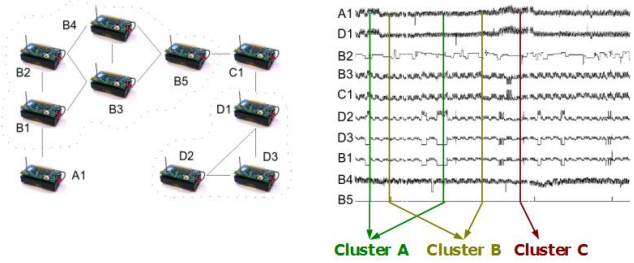

Figure 1: *Clustering Data Points*.

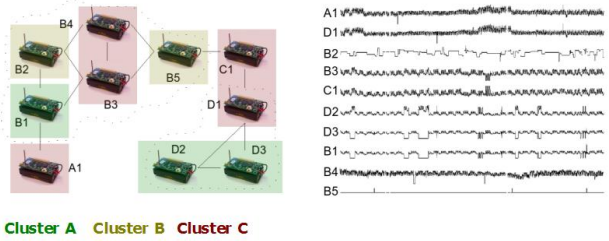

Figure 2: *Clustering Data Sources*.

2.1 GRID CLUSTERING OF DISTRIBUTED STREAMS

Clustering data points is probably the most common unsupervised learning process in knowledge discovery. In ubiquitous settings, however, there aren't many tailored solutions to try to extract knowledge in order to define dense regions of the sensor data space. Clustering examples in sensor networks can be used to search for hot-spots where sensors tend to produce data. In this settings, grid-based clustering represents a major asset as regions can be, strictly or loosely, defined by both the user and the adaptive process [3]. The application of clustering to grid cells enhances the abstraction of cells as interval regions which are better interpreted by humans. Moreover, comparing intervals or grids is usually easier than comparing exact

points, as an external scale is not required: intervals have intrinsic scaling. The comprehension of how sensors are interacting in the network is greatly improved by using grid based clustering techniques for the data examples produced by sensors.

The Distributed Grid Clustering (DGClust) algorithm was proposed for clustering data points produced on wide sensor networks [4]. The rationale is to use: a) online discretization of each single sensor data, tracking changes of data intervals (states) instead of raw data (to reduce communication to central server); b) frequent state monitoring at the central server, preventing processing all possible state combinations (to cut computation) [9]; and c) online clustering of frequent states (to keep high validity and adaptivity) [5]. Each local sensor receives data from a given source, producing a univariate data stream, which is potentially infinite. Therefore, each sensor's data is processed locally, being incrementally discretized into a univariate adaptive grid. Each new data point triggers a cell in this grid, reflecting the current state of the data stream at the local site. Whenever a local site changes its state, that is, the triggered cell changes, the new state is communicated to a central site. Furthermore, the central site keeps the global state of the entire network where each local site's state is the cell number of each local site's grid. Nowadays, sensor networks may include thousands of sensors. This scenario yields an exponential number of cell combinations to be monitored by the central site. However, it is expected that only a small number of this combinations are frequently triggered by the whole network, so, parallel to the aggregation, the central site keeps a small list of counters of the most frequent global states. Finally, the current clustering definition is defined and maintained by an adaptive partitional clustering algorithm applied on the frequent states central points.

2.2 DISTRIBUTED CLUSTERING OF GRID NODES

Clustering streaming data sources has been recently tackled in research, but usual clustering algorithms need the data streams to be fed to a central server [6]. Considering the number of sensors possibly included in a smart grid, this requirement could be a bottleneck. A local algorithm was proposed to perform clustering of sensors on ubiquitous sensor networks, based on the moving average of each node's data over time [8]. L2GClust has two main characteristics. On one hand, each sensor node keeps a sketch of its own data. On the other hand, communication is limited to direct neighbors, so clustering is computed at each node. The moving average of each node is approximated using memoryless fading average [7], while clustering is based on the furthest point algorithm [5] applied to the centroids computed by the node's direct neighbors. This way, each sensor acts as data stream source but also as a processing node, keeping a sketch of its own data, and a definition of the clustering structure of the entire network of data sources.

3 CONCLUDING REMARKS

In this talk we discussed approaches to ubiquitous data mining were both data sources and processing devices are distributed. Algorithms process local data and are able to communicate and interact with other agents to collaboratively construct a global solution. Ubiquitous data mining implies new requirements to be considered: i) the algorithms will have to use limited computational resources (in terms of computations, space and time); ii) the algorithms will have only a limited random access to data and may have to communicate with other agents; iii) answers will have to be ready in an anytime protocol [10, 11]. Ubiquitous data mining is in the core of next generation of data mining systems.

References

[1] Guha, S.; Meyerson, A.; Mishra, N.; Motwani, R.; and O'Callaghan, L. *Clustering data streams: Theory and practice*. IEEE Transactions on Knowledge and Data Engineering 15(3):515–528, 2003.

[2] Halkidi, M.; Batistakis, Y.; and Varzirgiannis, M. *On clustering validation techniques*. Journal of Intelligent Information Systems 17(2-3):107–145, 2001

[3] Rodrigues, P. P.; Gama, J.; and Lopes, L. *Knowledge discovery for sensor network comprehension*. In Cuzzocrea, A., ed., Intelligent Techniques for Warehousing and Mining Sensor Network Data. IGI Global. chapter 6, 118–135, 2010

[4] Gama, J.; Rodrigues, P. P.; and Lopes, L. *Clustering distributed sensor data streams using local processing and reduced communication*. Intelligent Data Analysis 15(1):3, 2011

[5] Gonzalez, T. F. *Clustering to minimize the maximum inter-cluster distance.* Theoretical Computer Science 38:293–306, 1985.

[6] Rodrigues, P. P., and Gama, J. *Clustering techniques in sensor networks. In Learning from Data Streams*. Springer Verlag. chapter 9, 125–142, 2007.

[7] Gama, J.; Sebastião, R.; and Rodrigues, P. P. *Issues in evaluation of stream learning algorithms*. In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 329–337. Paris, France: ACM Press, 2009.

[8] Rodrigues, P. P.; Gama, J.; Araújo, J.; and Lopes, L. *L2GClust: Local-to-global clustering of stream sources*. In Procs. 26th ACM International Symposium on Applied Computing, 1011–1016, 2011.

[9] Metwally, D. , A. Abbadi, *Efficient Computation of Frequent and Top-k Elements in Data Streams*, ICDT 2005.

[10] Gama, J. *Knowledge Discovery from Data Streams*, CRC Press, 2010

[11] M. May, L. Saitta, Ubiquitous Knowledge Discovery, LNAI 6202, Springer, 2010.