Holistic distributed stream clustering for smart grids

Pedro Pereira Rodrigues[1] and **João Gama**[2]

Abstract. Smart grids consist of millions of automated electronic meters that will be installed in electricity distribution networks and connected to servers that will manage grid supervision, billing and customer services. World sustainability regarding energy management will definitely rely on such grids, so smart grids need also to be sustainable themselves. This sustainability depends on several research problems that emerge from this new setting (from power balance to energy markets) requiring new approaches for knowledge discovery and decision support. This paper presents a holistic distributed stream clustering view of possible solutions for those problems, supported by previous research in related domains. The approach is based on two orthogonal clustering algorithms, combined for a holistic clustering of the grid. Experimental results are included to illustrate the benefits of each algorithm, while the proposal is discussed in terms of application to smart grid problems. This holistic approach could be used to help solving some of the smart grid intelligent layer research problems, thus improving global sustainability.

1 INTRODUCTION

The Smart Grid (SG), regarded as the next generation power grid, is an electric system that uses two-way digital information, cyber-secure communication technologies, and computational intelligence in an integrated fashion across heterogeneous and distributed electricity generation, transmission, distribution and consumption to achieve energy efficiency. It is a loose integration of complementary components, subsystems, functions, and services under the pervasive control of highly intelligent management-and-control systems [4].

A key and novel characteristic of smart grids is the *intelligent layer* that analyses the data produced by these meters allowing companies to develop powerful new capabilities in terms of grid management, planning and customer services for energy efficiency. The development of the market with a growing share of load management incentives and the increasing number of local generators will bring new difficulties to grid management and exploitation.

1.1 Research problems

Power and current balance is major goal of all electricity distribution networks, given its impact on the need to produce, buy or sell energy. Moreover, due to the fluctuating power from renewable energy sources and loads, supply-demand balancing of power system becomes problematic [17]. Several intelligent techniques have been proposed in the past that make use of the amounts of streaming data that is available. As an example, Pasdar and Mahne (2011) proposed to use ant collony optimization on smart meters data to improve the current balancing on low-voltage distribution network. Further research could even take more advantages from smart grids if consumption patterns could be extracted [14].

The energy market is changing to meet the global challenge of power consumption awareness even at the lower household level [3]. New energy distribution concepts and the advent of smart grids has changed the way energy is priced, negotiated and billed. We are now in a world of hourly real-time pricing [1] which make use of smart meters to overcome the need for demand prediction precision and, more important, demand prediction reliability [13]. Furthermore, with the advent of micro-generation at household level, the market expanded into multiplicity of energy buyers and energy sellers. In this setting, new techniques to efficiently auction in the market are required in order to make the smart grid smarter. Ramachandran et al. (2011) developed a profit-maximizing adaptive bidding strategy based on hybrid-immune-system-based particle swarm optimization.

1.2 Components and features

Smart grids are built on different sub-systems and present special features that need to be attended. The sources of energy are heterogeneous (power plants, wind, sun, sea, etc) and might be intermittent. A key characteristic of a SG is that it supports two-way flow of electricity and information: a user might generate electricity and put it back into the grid; electric vehicles may be used as mobile batteries, sending power back to the grid when demand is high, etc. This backward flow is relevant, mainly in *microgrids*, where parts of the system that might be *islanded* due to power failures. Following [4], the three major systems in SG are:

- Smart infrastructure system that supports advanced and heterogeneous electricity generation, delivery and consumption. Is responsible for metering information and monitoring, and information transmission among of systems, devices and sensors.
- Management systems providing advanced management and monitoring, grid topology and control services. The objectives are energy efficiency improvement, supply and demand balance, emission control, operation cost reduction, and utility maximization.
- Protection system providing grid reliability analysis, failure protection, security and privacy protection services.

1.3 Advantages and challenges

Some of the anticipated benefits of a SG include [4]:

- improving power reliability and quality;
- optimizing facility utilization and averting construction of back-up (peak load) power plants;

[1] LIAAD - INESC TEC & Faculty of Medicine of the University of Porto, Portugal, email: pprodrigues@med.up.pt
[2] LIAAD - INESC TEC & Faculty of Economics of the University of Porto, Portugal, email: jgama@fep.up.pt

- enhancing capacity and efficiency of existing electric power networks, hence improving resilience to disruption;
- enabling predictive maintenance and self-healing responses to system disturbances;
- facilitating expanded deployment of renewable energy sources;
- accommodating distributed power sources, while automating maintenance and operation;
- reducing greenhouse gas emissions by enabling electric vehicles and new power sources, thus reducing oil consumption by reducing the need for inefficient generation during peak usage periods;
- presenting opportunities to improve grid security;
- enabling transition to plug-in electric vehicles and new energy storage options;
- increasing consumer choice, new products, services, and markets.

All these jointly lead to massive research problems that might be tackled by artificial intelligence techniques. Some challenges where machine learning can play a relevant role, include:

- The reliability of the system supports itself on millions of meters and other devices that require online monitoring and global asset management [2].
- Real-time simulation and contingency analysis of the entire grid have to be possible. However, not all operations models currently make use of real-time data [8].
- Interoperability issues that arise from the integration of distributed generation and alternate energy sources [17].
- The heterogeneity and volatility of smart grids require mechanisms to allow islanding [9] and self-healing [2].
- Finer granularity in management leads to strong demand response requirements [7] and dynamic pricing strategies [1].

2 THE DATA MINING POINT OF VIEW

Present SG monitoring systems suffer from the lack of machine learning technologies that can adapt the behavior of monitoring systems on the basis of the sequence patterns arriving over time. From a data mining point of view, a smart grid is a network (eventually decomposable) of distributed sources of high-speed data streams.

Smart meters produce streams of data continuously in real-time. A data stream is an ordered sequence of instances that can be read only once or a small number of times [6, 10], using limited computing and storage capabilities. These sources of data are characterized by being open-ended, flowing at high-speed, and generated by non stationary distributions.In smart grids the dynamics of data are unknown; the topology of network changes over time, the number of meters tends to increase and the context where the meter acts evolves over time.

In smart grids, several knowledge discovery tasks are involved: prediction, cluster (profiling) analysis, event and anomaly detection, correlation analysis, etc. However, different types of devices present different levels of resources and care should be taken in data mining methods that aim to extract knowledge from such restricted scenarios. All these characteristics constitute real challenges and opportunities for applied research in ubiquitous data mining. Generally, the main features inherent to ubiquitous learning algorithms are that the system should be capable of process data incrementally, evolving over time, while monitoring the evolution of its own learning process and *self-diagnosis* this process. However, learning algorithms differ in the extent of self-awareness they offer in this diagnosis. .

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 [6]. 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 [15]. We identify two different settings for clustering problems in smart grids. In the first setting a cluster is defined to be a set of sensors (meters, households, generators, etc.). In the second setting, a cluster is defined to be a set of data points (demand, supply, prices, etc.) generated by multiple sources.

2.1 Research on clustering electrical networks

Several real-world applications use machine learning methods to extract knowledge from sensor networks. The case of electricity load demand analysis is a paradigmatic one that has been (and continues to be) studied. Sensors distributed all around electrical-power distribution networks produce streams of data at high-speed. Three major questions rise: a) can we define consumption profiles based on similar sensors? b) can we find global patterns in network consumption? and c) can we manage the uncertainty in sensor data?

To efficiently find consumption profiles, clustering techniques were applied to the streams produced by each sensor, either hierarchically at a central server [16] or distributed in the network [15]. Although the problem is still very hard to model, given the dimensionality of the networks at stake, the incremental systems evolved and adapt to changes in the data, bridging the gap to future paths of research. Regarding global network patterns, related research has resulted in a system that distributes the clustering process into local and central tasks, based on single sensor data discretization and centralized clustering of frequent states [5]. But data and models are both uncertain. For example, if a sensor reads 100, most of times it could be 99 or 101. This uncertainty has been tackled by reliability estimators and improved predictions using those estimates [13], but reliability for clustering definitions is still uncharted territory.

2.2 Clustering as a smart grid problem solver

In this work we argue that major smart grids problems previously enunciated can and should be addressed as unsupervised machine learning problems.

Power balance Power balance is the most basic-level problem that smart grids need to solve before anything else. The strongest requirement is that energy is available in the entire network. Hence, clustering the data and sources together to find hot-spots can detect specific points of danger in the network.

Multiple alternate sources In smart grids, supply and demand must be leveled across multiple alternate sources. Hence, combining clustering definitions for power demand and power supply should give indications on how to better level the sources.

Contingency analysis Contingency analysis tries to produce detection and reaction mechanisms to specific unexpected problems. Hence, monitoring the evolution of clusters of nodes, should help on detecting drifting sources of demand or supply.

Islanding Islanding is a concept that is directly connected with clustering, in the sense that it searches for subnetworks where demand and supply are leveled. Hence, local distributed clustering of sources and data should produce the expected definitions.

Self-healing Self-healing relates to the ability to rearrange and adapt the network on-the-fly to meet unexpected changes. Hence,

ad-hoc distributed clustering of sources, independently from a centralized server, should produce procedures for self-healing.

Online monitoring and asset management These features are strongly connected with incremental learning and adaptation of learned models. Hence, incremental models for sources and data clustering, and their evolution, should provide basic information.

Dynamic energy pricing Energy pricing largely depends on supply and demand balance. Hence, clustering power demand and supply together with buy and sell prices, should give insights on prospective energy pricing.

3 HOLISTIC DISTRIBUTED CLUSTERING

The smart grid produces different types of data, on each source (node or subnetwork), which must be taken into account: power demand, power supply, energy sell price, energy buy price. As previously stated, two clustering problems exist: clustering data and clustering data sources. This way, each node might be assigned to a cluster on (at least) eight different clustering definitions. For all problems, there is a common requirement: each node (meter) should process locally their own data. Only aggregated data should be shared between the different nodes in the grid.

From the previous section it became clear that a holistic approach to clustering in smart grids is needed and should produce benefits to energy sustainability. In this section we present such a proposal, based on two existing works on stream clustering (L2GClust and DGClust) and their prospective integration in a multi-dimensional clustering system. Next sections present the original clustering algorithms, their application to electricity demand sensor data streams, and how they could be merged into a holistic clustering system.

3.1 L2GClust: 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 [15]. 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 [15]. *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, while clustering is based on the furthest point algorithm 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.

Global evaluation of the L2GClust algorithm on synthetic data revealed high agreement with the centralized, yet streaming, counterpart, being especially robust in terms of cluster separability. Also, for stable concepts, empirical evidence of convergence was found. On the other hand, sensitivity analysis exposed the robusteness of the local algorithm approach. Figure 1 shows that agreement levels are robust to an increase on the number of clusters, being, however, a bit more sensitive with respect to network size and cluster overlapping. Nonetheless, the robusteness to network communication problems is exposed, as the proportion of agreement is harmed only for high levels of communication incompleteness.

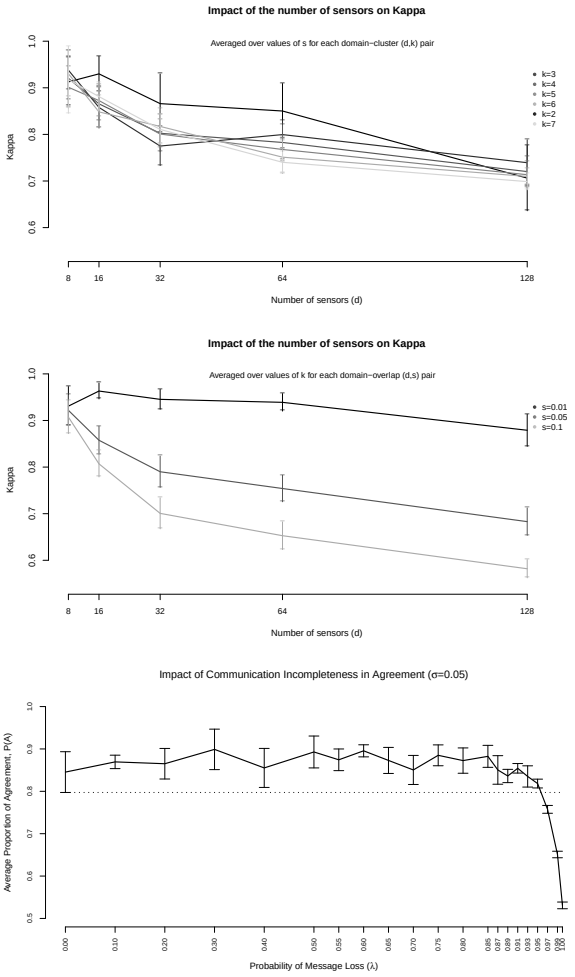

Figure 1. L2GClust: sensitivity of $\hat{\kappa}$ statistic to the number of sensors (d), for different number (k) and overlap (s) of clusters. Bottom plot presents the impact of communication incompleteness on average proportion of agreement for 5 clusters in a 128 sensor network.

One important task in electrical networks is to define profiles of consumers, to better predict their behavior in the near future. L2GClust was applied to a sample of an electrical network to try to find such profiles. From the raw data received at each sub-station, observations were aggregated on a hourly basis over more than two and a half years [14]. The log of electricity demand data from active power sensors was used to check whether consumer profiles would rise. The log has hourly data from a subsample (780 sensors) of the entire data set ($\sim$4000 sensors). Since no information existed on the actual electricity distribution network, the simulator used this dataset as input data to a random network and monitored the resulting clustering structures. Unfortunately, real data is never clean, and half of the sensors have more than 27% missing values, which naturally hindered the analysis. Given this, and the dynamic nature of the data, no convergence was possible in the clustering structures. However, we could stress that, as more data is being fed to the system, better agreement can be achieved with the centralized approach, as exposed in Figure 2. Hence, not only does the agreement tend to increase with more observations, but also changes on the clustering structure are apparently possible to detect. L2GClust presented good characteris-

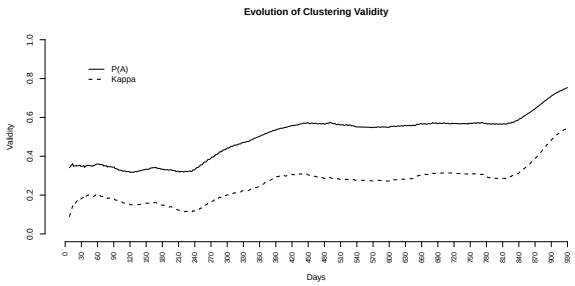

Figure 2. L2GClust evolution of clustering agreement (probability of agreement and $\hat{\kappa}$ statistic) for a real active power sensor data log.

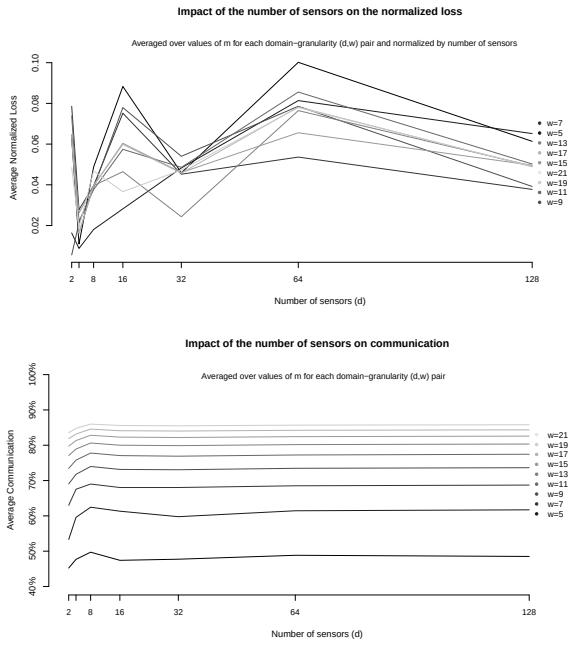

Figure 3. DGClust: impact of the number of sensors on loss to real centroids (top) and communication reduction (bottom) [5].

tics to find clusters of sensors in wide networks such as smart grids.

3.2 DGClust: Grid clustering of grid data 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 hotspots 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 [5]. 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 [5]. 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); and c) online clustering of frequent states (to keep high validity and adaptivity). 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.

To evaluate the sensitivity of the system to the number of sensors, synthetic data was used and the average result for a given value of granularity (w), averaged over all values of number of frequent states to monitor (m, as loss seemed to be only lightly dependent on this factor) was analyzed. In figure 3 we note no clear trend, strengthening the evidence of robusteness to wide sensor networks. Regarding communication reduction when compared with centralized clustering, figure 3 also shows that the amount of communication reduction does not depend on the number of sensors. This way, the benefits of reduced transmission rates are extensible to wide sensor networks.

3.3 HDClust: Holistic Distributed Clustering

The two algorithms previously exposed are designed for streaming data, and work with reduced computational costs in terms of memory and communications bandwidth. They present strong characteristics that could be even improved if used together. In L2GClust, each sensor node each node has an approximation of the global clustering. In DGClust, a centralized site maintains the global cluster structure of the entire network at reduced communication costs. The main idea of the *Holistic Distributed Clustering* (HDClust) is to integrate the local distributed approach of L2GClust, with the grid data clustering approach of DGClust, in order to achieve the holistic clustering of data and sources on sensor networks such as smart grids. Specifically, for each measured dimension:

- each local node (meter) keeps a sketch of its own data streams (as in L2GClust) and a local discretization grid (as in DGClust);
- communication is restricted to the neighborhood (as in L2GClust);
- at regular intervals, each local node receives from its neighbors the estimates of the clusters centroids (as in L2GClust) and the current data discretized grid cell (as in DGClust);
- each node keeps an estimate of the global clustering of nodes by clustering neighbors' centroids (as in L2GClust);
- each node keeps a frequent state list and maintains a clustering of the most frequent states (as in DGClust) from the neighbors;

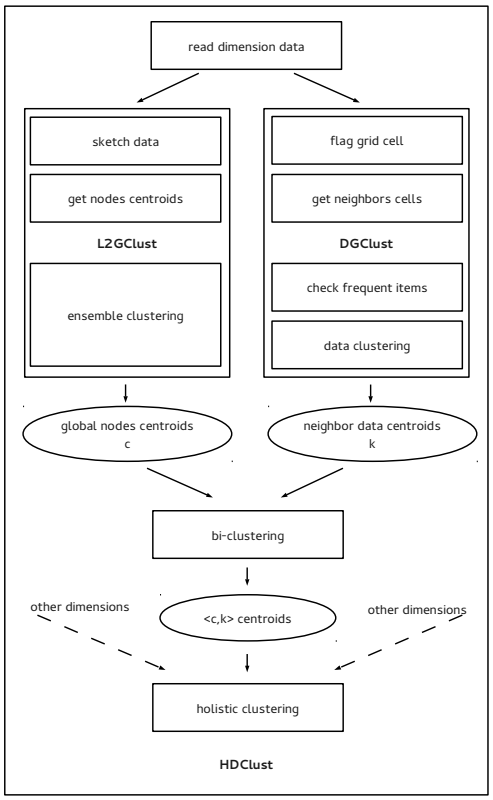

Figure 4. HDClust schema to be applied at each node, for each included dimension. Left branch applies L2GClust while right branch applies DGClust using data from the neighbors, each node acting also as central clustering agent. Both clustering definitions are then combined and integrated with other measured dimensions.

- to link clustering of sources with clustering of data, each node also receives from the neighbors their self assignment to a cluster.

In the resulting cluster structure, each sensor maintains **C** clusters of data sources, and **K** clusters of data points.

In a smart grid context, and taking advantage of the decomposable property of the grid network (microgrids), L2GClust and DGClust can work together. Assume a microgrid of **D** sensors, and 4 dimensions or quantities of interest: power demand, power supply, energy sell price and energy buy price. The resulting HDClust, the network is summarized by **C** clusters of data sources, and **K** clusters of data points, for each quantity of interest. In real-time and at each moment, each sensor is in a state $\langle c_i, k_i \rangle$ in each dimension. Figure 4 presents the global schema for a holistic approach to clustering, to be applied at each node of a smart grid. The combination of the characteristics both algorithms seems not only possible, but extremely relevant as complementary knowledge discovery in a holistic view of the grid.

4 REMARKS AND FUTURE PATHS

Smart grids are a paradigmatic example of ubiquitous streaming data sources. Data is produced at high speed, from a dynamic (time-changing) environment. Meters are geographically distributed, forming a network. On top of clustering algorithms, several tasks can be computed: profiling, anomaly and event detection, outliers detections, trends, deviations, etc. In this paper, we have discussed distributed clustering algorithm for data streams produced on wide sensor networks like smart grids. Furthermore, we have shown how smart grid problems can be addressed as clustering problems, and proposed a holistic approach to better extract knowledge from the grid. We believe that this holistic approach could be used to help solving some of the smart grid intelligent layer research problems. Current research focus on the integration of both algorithms into the schema and its evaluation on real-world electrical networks data.

ACKNOWLEDGEMENTS This work is funded by the ERDF through Programme COMPETE and by the Portuguese Government through FCT, projects PEst-C/SAU/UI0753/2011 and PTDC/EIA/098355/2008. The authors acknowledge the help of Luís Lopes and João Araújo.

REFERENCES

[1] H. Allcott, 'Rethinking real-time electricity pricing', *Resource and Energy Economics*, **33**(4), 820–842, (2011).
[2] S.M. Amin, 'Smart grid: Overview, issues and opportunities. advances and challenges in sensing, modeling, simulation, optimization and control', *European Journal of Control*, **17**(5-6), 547–567, (2011).
[3] F. Benzi, N. Anglani, E. Bassi, and L. Frosini, 'Electricity smart meters interfacing the households', *IEEE Transactions on Industrial Electronics*, **58**(10), 4487–4494, (2011).
[4] Xi Fang, Satyajayant Misra, Guoliang Xue, and Dejun Yang, 'Smart grid – the new and improved power grid: a survey', *IEEE Communications Surveys & Tutorials*, (2012). (to appear).
[5] João Gama, Pedro Pereira Rodrigues, and Luís Lopes, 'Clustering distributed sensor data streams using local processing and reduced communication', *Intelligent Data Analysis*, **15**(1), 3–28, (January 2011).
[6] Sudipto Guha, Adam Meyerson, Nina Mishra, Rajeev Motwani, and Liadan O'Callaghan, 'Clustering data streams: Theory and practice', *IEEE Transactions on Knowledge and Data Engineering*, **15**(3), 515–528, (2003).
[7] A. Iwayemi, P. Yi, X. Dong, and C. Zhou, 'Knowing when to act: An optimal stopping method for smart grid demand response', *IEEE Network*, **25**(5), 44–49, (2011).
[8] J.A. Kavicky, 'Impacts of smart grid data on parallel path and contingency analysis efforts', in *IEEE PES General Meeting*, (2010).
[9] R.H. Lasseter, 'Smart distribution: Coupled microgrids', *Proceedings of the IEEE*, **99**(6), 1074–1082, (2011).
[10] S. Muthukrishnan, *Data Streams: Algorithms and Applications*, Now Publishers Inc, New York, NY, 2005.
[11] A. Pasdar and H.H. Mehne, 'Intelligent three-phase current balancing technique for single-phase load based on smart metering', *International Journal of Electrical Power and Energy Systems*, **33**(3), 693–698, (2011).
[12] B. Ramachandran, S.K. Srivastava, C.S. Edrington, and D.A. Cartes, 'An intelligent auction scheme for smart grid market using a hybrid immune algorithm', *IEEE Transactions on Industrial Electronics*, **58**(10), 4603–4612, (2011).
[13] Pedro Pereira Rodrigues, Zoran Bosnić, João Gama, and Igor Kononenko, 'Estimating reliability for assessing and correcting individual streaming predictions', in *Reliable Knowledge Discovery*, 267–287, Springer Verlag, (2012).
[14] Pedro Pereira Rodrigues and João Gama, 'A system for analysis and prediction of electricity load streams', *Intelligent Data Analysis*, **13**(3), 477–496, (June 2009).
[15] Pedro Pereira Rodrigues, João Gama, João Araújo, and Luís Lopes, 'L2GClust: Local-to-global clustering of stream sources', in *Proceedings of ACM SAC 2011*, pp. 1011–1016, (March 2011).
[16] Pedro Pereira Rodrigues, João Gama, and João Pedro Pedroso, 'Hierarchical clustering of time-series data streams', *IEEE Transactions on Knowledge and Data Engineering*, **20**(5), 615–627, (May 2008).
[17] K. Tanaka, A. Yoza, K. Ogimi, A. Yona, T. Senjyu, T. Funabashi, and C.-H. Kim, 'Optimal operation of dc smart house system by controllable loads based on smart grid topology', *Renewable Energy*, **39**(1), 132–139, (2012).