ON-LINE CHANGE POINT DETECTION IN HOUSEHOLD'S ELECTRICITY POWER CONSUMPTION DATA SERIES FOR SMART GRID APPLICATIONS

Istvan Pinter[1*], Lorant Kovacs[1], Andras Olah[1,2], Rajmund Drenyovszki[1,3], David Tisza[1,2] and Kalman Tornai[1,2]

[1]Kecskemet College, Faculty of Mechanical Engineering and Automation,
Izsaki ut 10, H-6000 Kecskemet, HUNGARY
[2]Pazmany Peter Catholic University, Faculty of Information Technology
Prater utca 50/a, H-1083 Budapest, HUNGARY
[3]University of Pannonia, Department of Computer Science and Systems Technology,
Egyetem u. 10, H-8200 Veszprem, HUNGARY

* *Corresponding author e-mail: pinter.istvan@gamf.kefo.hu*

Abstract

In the previous event of TEAM conference series we presented an off-line method for estimating CPTs (Change PoinTs) in household's electricity power consumption data. The basis of the method is the Jensen-Shannon divergence (JSD) contour, and the algorithm is recursive in its nature. However, for smart grids an on-line version would be more suitable, so we have elaborated a sliding-window based algorithm, and our recent results in this respect will be presented.

Keywords:
smart grid, change-point detection, sliding window, Jensen-Shanon divergence

1. Introduction

Smart grid is new vision of future evolution of electricity systems, this concept is very important. In contrast to the present grid, smart grid is adaptively managing the balance (supply and demand) of the system, and can handle the challenge of incorporation of renewable energy resources. Therefore, the role of a household should be revised considering the power consumption/production questions. For example, novel power consumption demand is charging PHEVs' batteries, and novel power generation sources at household-level are small wind turbines, photovoltaic panels, charged batteries of electric vehicles. These increase the fluctuation in electrical energy and makes balancing more difficult.

Based on recent ICT (Information and Communication Technology) the smart meters can be used for recording and evaluating the household's power consumption in a fine time-scale (e.g. in every minute or 15 minutes) which not only results in a large database, but calls for novel machine learning methods in managing household's electrical power demand. Recently an algorithm has been proposed in [1] for categorizing the households power consumption data into smaller number of typical consumption pattern and *daily* power consumption data series have been used. For categorizing daily data the first step was the normalization the power consumption in order to be considered as a probability distribution, secondly it was modeled as a mixture-of-Gaussian type distribution, and finally as a similarity measure the Jensen-Shannon divergence has been used.

However, using finer time-resolution the problem of automatic segmentation of the power consumption data series naturally appears. Our approach for this problem is the so-called change point detection, namely determining those time-instants where statistical properties of time-series data change [2], [3]. A lot of problem in smart grids, such as automatic detection of customer categories or anomaly detection are rooting in segmentation of non-stationary time series into stationary subsequences [4], [5]. Let's note that our method is for partitioning of the power consumption data series of the same household.

In algorithm development the source of real data was the UCI Machine Learning Repository [6]. In this database there are seven data series available from December 2006 to November 2010. The sampling interval is 1 minute, the recordings are global active power, global reactive power, voltage, intensity, sub-meterings in kitchen, laundry, and water heater with air-conditioner.

The structure of the paper is the following. After the introduction, in Section 2 both the mathematical background and the previously published results in developing the off-line (or recursive) version of the method are summarized, and augmented with some novel illustrations, too. In Section 3 the on-line (sliding window-based) version is presented, and its segmentation is compared with the off-line one. Section 4 presents our preliminary results in clustering the automatically segmented data. The article ends with conclusions and acknowledgement.

2. Summary of the recursive algorithm and its application

The algorithm is based on the concept of generalized Jensen-Shannon divergence. For probability distributions $P=\{p_1,p_2,\dots,p_i,\dots p_K\}$ and $Q=\{q_1,q_2,\dots q_i,\dots q_K\}$ and for weights $\omega_A,\omega_B>0$, so that $\omega_A+\omega_B=1$, the average distribution is $A=\omega_P P+\omega_Q Q$ and using these notations the definition of generalized JSD is the following:

$$JSD(P,Q)=H\big(\omega_P P+\omega_Q Q\big)-\big[\omega_P H(P)+\omega_Q H(Q)\big] \quad (1)$$

where $H(P)=-\sum_{k=1}^{K}p_k log p_k$.denotes the Shannon-entropy. Using the generalized JSD the so-called JSD-contour can be determined. A given *N*-symbol length sequence *S* can be divided into left and right sequences in positions $n=1,\dots,N-1$. At a given position *n* the number of symbols in left sequence is *n*, whilst the number of symbols in the right sequence is $N-n$. We can now estimate the *P* and *Q* distributions for the left and right sequences using the symbols relative frequencies, and by choosing the weights as $\omega_P=\frac{n}{N}$, $\omega_Q=\frac{N-n}{N}$ the generalized JSD can be computed. Repeating this procedure for all *n* positions, we get the function $JSD(n)$, which we call JSD-contour. The candidate for estimation of CPT is the index of maximum value of the JSD-contour. In order to accept or reject this candidate CPT an approximation of probability distribution function *F(x)* of JSD_{max} has been published, so the CPT-estimation can be checked at a given p_0 confidence level. The computed index is accepted as CPT-estimate, if $F(JSD_{max})>p_0$ is hold. Much more details can be found in the relevant article [3].

Our first goal was realizing the recursive algorithm which appeared in [3]. For this the whole recording is necessary, e.g. one week long data. Then the data series is splitted into left and right sub-sequences using the index of the JSD-contour maxima (that is, using, the CPT), and this procedure is repeated until the desired level of recursion. In Figure 1 this procedure is illustrated.

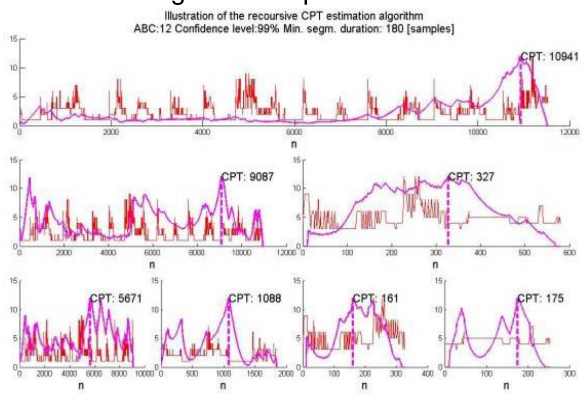

Figure 1. Illustration of the JSD-based recursive CPT detection algorithm

The statistical distribution of the left-sequence-data is different from the right-sequence one; a typical case can be seen in Figure 2.

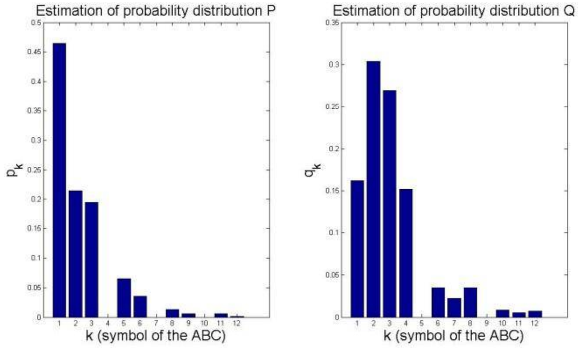

Figure 2. Estimated probability distributions of the left and right sequences (denoted by P, Q respectively)

This algorithm has been used for determining change in artificially concatenated data series using real data in case of seasonal global active power measurements. The change detection capability has also been tested similarly in case of appliances [4] (see Figures 3., 4). The automatic segmentation capability of daily measurements was presented in [5]. In order to be comparable with the recently proposed sliding window-based algorithm, we repeated the computations for different year/month/days combination (see Figure 5.).

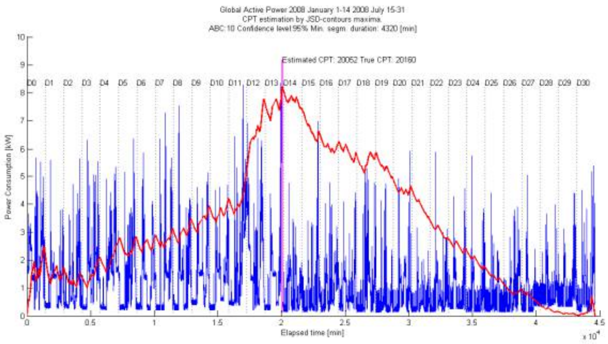

Figure 3. Seasonal change detection using the recursive algorithm.

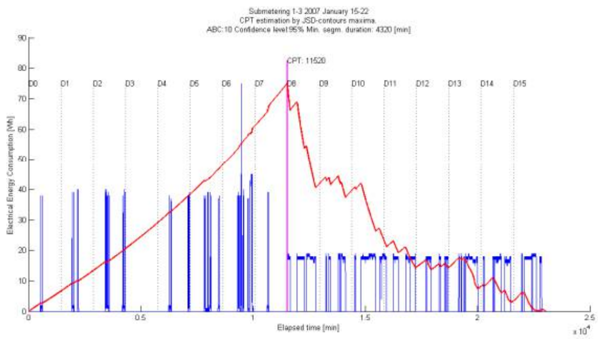

Figure 4. Appliance change detection using the off-line method.

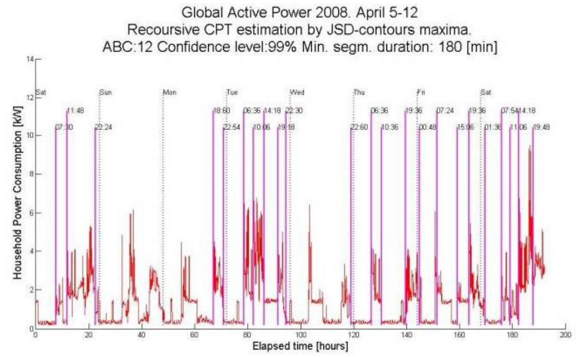

Figure 5. Automatic segmentation of global active power consumption data using the recursive CPT estimation algorithm (April 5-12, 2008)

3. Sliding window-based algorithm for on-line CPT-estimation and automatic segmentation

Because of its basically off-line nature the recursive method is not suitable for application in smart meters. Therefore, we elaborated a novel method, which is on-line and can be used in smart meters for real-time data analysis.

The basis of the method is the sliding window and a predefined statistical confidence level [3], [4], [5]. The JSD-contour is computed using the window's data and when the value of the probability distribution function at the JSD-contour's maxima lower, than the confidence level, the window is augmented with the new measurement value. This procedure is repeated until exceeding the above mentioned level. Moreover, the window-augmentation is also continued, when the length of either the left or the right sequence is lower, than a predefined minimal length. The resulted segmentation can be seen in Figure 6 using the same data as in Figure 5.

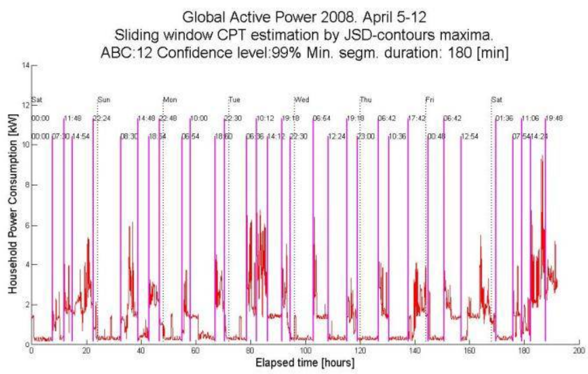

Figure 6. Automatic segmentation of the global active power data using sliding window in JSD-based CPT detection algorithm

It is important to note, that for the algorithm the symbol-sequence is necessary, not the original measured data. For this purpose a linear quantizer has been realized, which is illustrated in Figure 7.

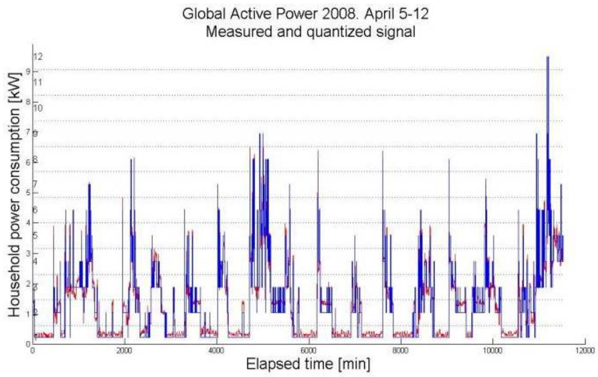

Figure 7. Generating symbol sequences from original measurements using linear quantization

When comparing the segmentation result of the recursive and the sliding window-based algorithm several differences can be observed. The cause of these, that the two algorithm deals with different symbol sub-sequences, or in other words, their 'scopes' are not the same. Moreover, the splitting process of the data into left and right subsequences in case of the sliding window based method is clear, because it deals with exactly two symbol-generating probability distributions (P and Q in Figure 2).

4. Preliminary results in clustering the segmented data

In households' electrical load balancing the existence of typical consumption patterns could be useful, for these can help in prediction of the consumption demand. Moreover, not only the consumption pattern is interesting, but its duration as well. The sliding window-based algorithm presented above automatically determines these data. Therefore, the next interesting problem is discovering similar patterns in the automatically segmented power consumption data series. This step is important because in case of existence of similar patterns a new direction of the research can be defined, namely developing nonlinear time-series prediction algorithms for predicting the next pattern and its duration.

In this paper our preliminary results in discovering the similar patterns in automatically segmented power consumption data series of a household are presented. Firstly we used distance-based clustering algorithms for this problem.

Although the results of segmentation is a doublet of <probability distribution of symbols in the segment><duration of the segment>, at first attempt we have neglected the probability distribution nature of the data. Instead, they have been considered as 12-dimensional vectors. The capabilities of two distance-based algorithms have been investigated: the hierarchical clustering which resulted in a dendrogram, and a version of k-means clustering. In both cases the L1 distance

(Manhattan-distance or 'cityblock' distance) has been used.
The time-series analyzed was "Global Active Power April 5-12 2008" data from [6]. The minimal segment duration was 180 minutes, and the sliding window based automatic segmentation algorithm has detected 32 segments. These 12 dimensional vectors were the input data set of clustering.
The resulted dendrogram can be seen in Figure 8.

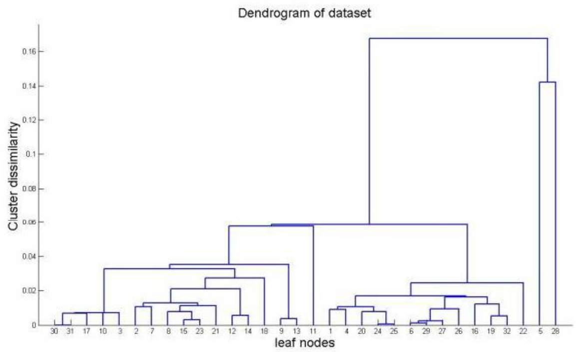

Figure 8. Dendrogram, computed from automatically segmented data using sliding window-based CPT estimation

It can be seen in the figure that there exists a similarity level where the data are acceptably well clusterable, and the number of clusters is 3 or 4. This interpretation is strengthened by the results of *k*-means clustering, which is illustrated in Figure 9 for different cluster numbers (*k* = 2..9). The silhouette diagram shows that 3 or 4 clusters are acceptable in this case too. For all computations we used the MATLAB software.

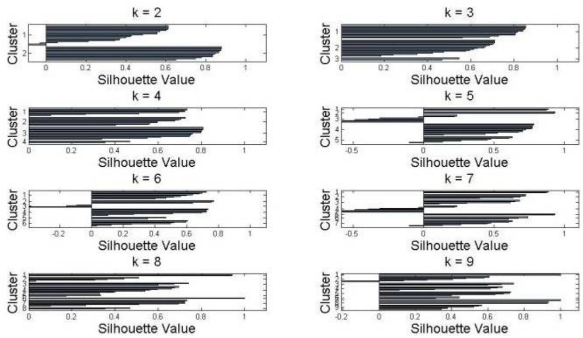

Figure 9. Silouette diagram, computed from automatically segmented data using sliding window-based CPT estimation

5. Conclusion

In this paper we proposed a novel algorithm for JSD-based CPT estimation, which is sliding window-based. By comparing it with the previously elaborated recursive algorithm we concluded, that the latter is suitable for global analysis of the data series, while the novel approach can be considered as a candidate for real-time applications. Moreover, we presented our first results in discovering similar patterns in a household's power consumption by applying hierarchical and k-means clustering methods.

Acknowledgement

This research and publication have been supported by the European Union and Hungary and co-financed by the European Social Fund through the project TÁMOP-4.2.2.C-11/1/KONV-2012-0004: National Research Center for the Development and Market Introduction of Advanced Information and Communication Technologies. This source of support is gratefully acknowledged.

References

[1] H. Hino, H. Shen, N. Murata, S. Wakao, Y. Hayashi: *A Versatile Clustering Method for Electricity Consumption Pattern Analysis in Households.* IEEE Transactions on Smart Grid, Vol. 4. No. 2., June 2013, pp. 1048 – 1057.

[2] Y. Kawahara, M. Sugiyama: *Change-Point Detection in Time-Series Data by Direct Density-Ratio Estimation.* Proceedings of the SIAM International Conference on Data Mining, SDM 2009, April 30 - May 2, 2009, Sparks, Nevada, USA (2009), pp. 389-400

[3] I. Grosse, P. Bernaola-Galván, P. Carpena, R. Román-Holdán, J. Oliver and H. E. Stanley: „Analysis of symbolic sequences using the Jensen-Shannon divergence", Physical Review E, Volume 65, 2002, 041905 pp. 65-81.

[4] I. Pintér, L. Kovács, A. Oláh, R. Drenyovszky, D. Tisza, K. Tornai: *Application of Jensen-Shannon Divergence in Smart Grids* 5th International Scientific and Expert Conference TEAM 2013 Technique, Education, Agriculture & Management Prešov, 4th to 6th November 2013

[5] I. Pintér, L. Kovács, A. Oláh, R. Drenyovszky, D. Tisza, K. Tornai: *Automatic Segmentation of Electricity Consumption Data Series with Jensen-Shannon Divergence* Proceedings of the Energycon 2014, IEEE International Energy Conference, 13-16 May, 2014, Dubrovnik, Croatia pp. 873-877.

[6] Bache, K., Lichman, M.: „*UCI Machine Learning Repository*", Irvine, University of California, School of Information and Computer Science, [http:/archive.ics.uci.edu/ml], 2013.