Krzysztof Gajowniczek[1,2], Tomasz Ząbkowski[1]

Department of Informatics, Warsaw University of Life Sciences (1), Systems Research Institute, Polish Academy of Sciences (2)

Clustering Approach for Short Term Electricity Forecasting

***Abstract**. Smart meters open new possibilities for short-term electricity load forecasting at different scales. In this paper, our contribution is twofold: (1) we deal with short term electricity load forecasting for 24 hours ahead on the individual household level, (2) we investigated different time series similarity measures for the purpose of clustering and further increase the electricity load forecasting.*

Keywords: artificial neural networks, hierarchical clustering, smart metering system, time series clustering, time series similarity measures.

Introduction

Installation of smart meters opens new possibilities for advanced analytics of electricity consumption at the individual household level. One of the most important tasks in various Smart Grid applications is the short-term electricity load forecasting at different scales, from an individual customer to a whole group of customers.

In this paper, based on various similarity measures and clustering algorithms, we will study an approach to forecast the hourly electricity loads for individual consumers for 24 hours ahead, taking into account the one neural network model for all the households in each group and comparing the results with the base model built for all the consumers.

Time series similarity

The time series distance measures are usually divided into four categories [1, 2]: (1) shape based – Euclidean, Short Time Series, Dynamic Time Warping, (2) edit based - Real Sequences, Longest Common Subsequence, (3) features based – Autocorrelation, Crosscorrelation, Fourier coefficients, TQuest, Integrated Periodogram, Wavelet Feature Extraction and (4) structure based. This article focuses on the first three categories because they are applicable to all cases.

Clustering method

Ward's minimum variance method [3] can be defined and implemented recursively by a Lance–Williams algorithm. The Lance–Williams algorithms are the infinite family of agglomerative hierarchical clustering algorithms which are represented by a recursive formula for updating cluster distances in terms of squared similarities at each step (each time a pair of clusters is merged).

To choose optimal number of clusters the Calinski and Harabasz index [4] was used.

Numerical experiment

The data have been obtained from Pecan Street Inc. via the WikiEnergy project [5]. The data set contains data from 61 homes, in which the household aggregate power demand are monitored at 1 hour intervals over 6 months from December 2012 until June 2013.

To assess the model performance for forecasting by 10-fold cross validation, we used Mean Absolute Percentage Error and resistant Mean Absolute Percentage Error [6].

For training neural networks we used the BFGS (Broyden – Fletcher – Goldfarb – Shanno) algorithm, which belongs to the broad family of quasi-Newton optimization methods. Input layer consisted of 52 perceptrons, hidden layer consisted of 20 perceptrons and finally, the output layer consisted one perceptron. All the perceptrons were activated by logistic function and as loss function we chose the least squares estimator.

In all cases the optimum number of clusters based on Calinski-Harabasz was set to 2. Number of households in each clusters was not constant and is ranged from 30 and 31 for Real Sequences, up to 53 and 8 for Crosscorrelation. The greatest increase in accuracy was obtained for Dynamic Time Warping. Please, see Tab. 1 for details.

Table 1. Obtained results

	MAPE		r-MAPE	
Cluster number	1	2	1	2
Base model	1.670		0.462	
Euclidean	0.714	1.360	0.427	0.480
Short Time Series	2.220	0.593	0.500	0.420
Dynamic Time Warping	0.695	0.701	0.417	0.452
Real Sequences	0.588	1.870	0.404	0.524
Longest Common Subsequence	1.870	4.820	0.524	0.468
Autocorrelation	2.230	0.835	0.427	0.531
Crosscorrelation	0.831	1.360	0.471	0.480
Fourier coefficients	0.714	1.360	0.427	0.480
TQuest	0.677	1.810	0.440	0.487
Integrated Periodogram	3.300	0.701	0.517	0.452
Wavelet Feature Extraction	0.714	1.360	0.427	0.480

Conclusion

In this paper, we presented an approach to forecast electricity load on individual level data for a group of customers. The result of MLP network model used for 24 hours ahead short term load forecast shows that models' performance differs substantially and it is dependent of the time series similarity measures.

The study is cofounded by the European Union from resources of the European Social Fund. Project PO KL „Information technologies: Research and their interdisciplinary applications", Agreement UDA-POKL.04.01.01-00-051/10-00.

REFERENCES

[1] Mori U., Mendiburu A., Lozano J.A. TSdist: Distance Measures for Time Series data, (2014), R package version 1.2, http://CRAN.R-project.org/package=TSdist

[2] Montero P., Vilar J.A. TSclust: An R Package for Time Series Clustering, Journal of Statistical Software, 62(1), (2014), 1-43

[3] Ward J.H. Hierarchical Grouping to Optimize an Objective Function, Journal of the American Statistical Association, 58(301), (1963), 236–244.

[4] Calinski T., Harabasz J. A Dendrite Method for Cluster Analysis, Communications in Statistics – Theory and Methods, 3(1), (1974), 1–27.

[5] https://dataport.pecanstreet.org/

[6] Moreno J.J.M., Pol A. P., Abad A.S., Blasco B.C. Using the R-MAPE index as a resistant measure of forecast accuracy, Psicothema 25(4), (2013), 500-506.

***Authors**: Krzysztof Gajowniczek, Warsaw University of Life Sciences, Nowoursynowska 159, 02-787 Warsaw, Poland, krzysztof_gajowniczek@sggw.pl and Systems Research Institute, Polish Academy of Sciences, Newelska 6, 01-447 Warsaw, Poland **Tomasz Ząbkowski**, Warsaw University of Life Sciences, Nowoursynowska 159, 02-787 Warsaw, Poland, tomasz_zabkowski@sggw.pl*