

Early Pattern Classification Using Partial Pattern Matching

James Conigliaro, jconigliaro@ieee.org
Dr. Richard Povinelli
Marquette University

Abstract

The discovery of patterns within time-series energy consumption data is important when attempting to detect anomalous energy usage patterns early in the day, when changes to behavior can still affect the day's usage. This paper proposes a technique in which data clustering techniques are used to discover patterns in energy consumption. These patterns are then used to perform early classification of a day's usage for the purposes of detecting anomalous data. The study indicates that a basic partial pattern matching technique may be used to perform early classification with a high degree of success.

Introduction

Advances in data acquisition capabilities have resulted in large quantities of data being made available for practitioners in a number of fields. Much of this data is temporal in nature and thus constitute time series. The discovery of patterns within time-series data sets is a problem that receives much attention and may be applied to a wide range of fields. One field in which pattern discovery in time series data is particularly important is the energy industry. Unfortunately, knowledge discovery in energy time-series databases is not widely studied.

In the energy industry it is financially beneficial to be able to determine early in a any given day what electrical consumption pattern a facility is following for the purposes of the detection of anomalous behavior. Equally important is the ability to determine if an electrical consumption pattern will not match any previously known patterns early in the day.

This paper proposes a technique for early classification of periodic patterns in time-series energy consumption databases. Classification is defined as the process of assigning an object to a preexisting group of patterns. Clustering is the process of creating or discovering the groups through the use of data [2]. The technique first utilizes clustering to create a set of class under which data is expected to fall. Subsequently, as new data is acquired a partial pattern match is performed to determine which class or classes the remainder of the data may fall under.

Data Description¹

Data from a number of electrical meters or sensors are collected at 15-minute intervals over a period of several months. Energy consumption data typically have the following characteristics:

- Periodic in nature with a duty cycle of roughly 24 hours.
- Different patterns on weekends, weekdays, and holidays.
- Patterns fluctuate during different seasons.
- Subject to fluctuations that appear random.

Figure 1 illustrates seven days worth of energy demand data, recorded in 15-minute intervals. Four distinct patterns may be discerned through visual inspection. The data from Monday through Wednesday exhibit rapid rises in the electrical consumption rate at around 7:00 AM and 11:00 AM with a peak demand of 550 KW occurring shortly after noon. The data from Thursday and Friday exhibit similar rises in demand at 7:00 AM and 11:00 AM but have a much lower peak of 450 kW. Saturday's pattern shows a small rise in consumption at 7:00 AM and another at around 6:00 PM with a peak of only 300 kW. Sunday exhibits a relatively flat

¹ The data for this project are being provided, with permission, from Engage Networks, Inc.

demand curve. However after viewing a months worth of data, as shown in Figure 2, these patterns exhibit less visual distinction. Further, when one reviews a longer time period, seasonal patterns may be found, as shown in Figure 3 in which the range of electrical consumption increases with lower floors and higher peaks beginning in March.

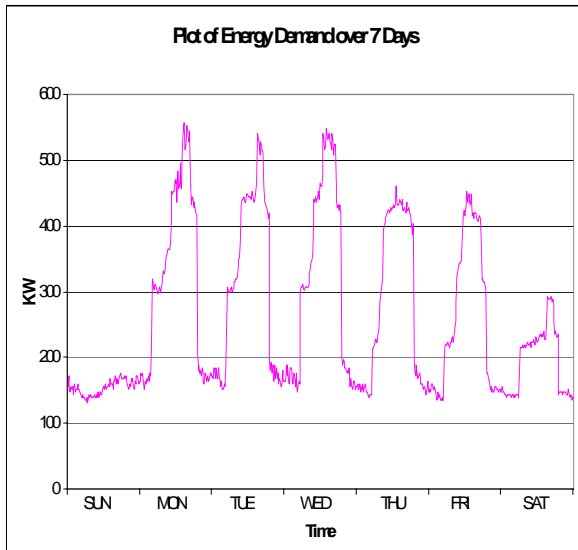


Figure 1: Seven Data KW Plot

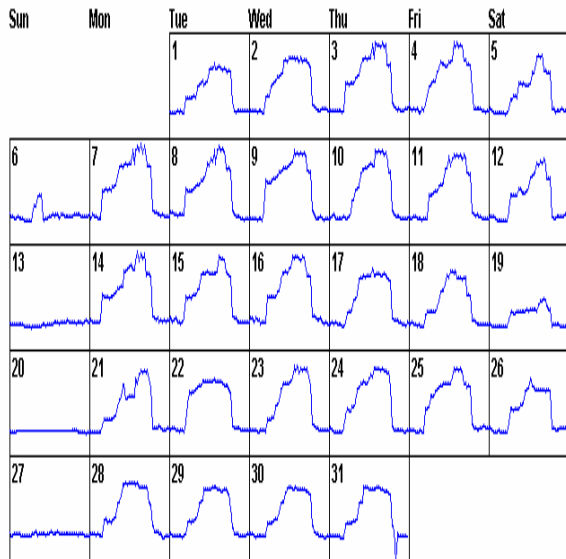


Figure 2: A monthly plot

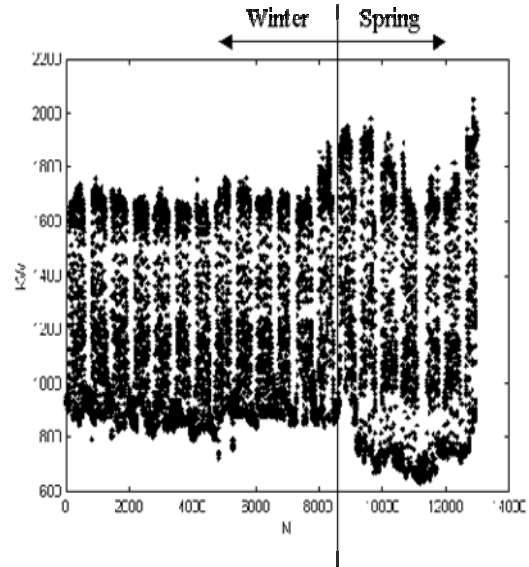


Figure 3: Plot of energy demand over a 20-week period.

Model Description

Given the characteristics of electrical demand time-series data, the system under discussion is designed to classify daily consumption patterns into distinct classes. These classes are then utilized to identify deviations from known patterns. Furthermore, the system should be efficient enough to run in an on-line manner.

The first step in the sequence is to cluster the time series into distinct daily patterns consisting of 96 data points per day. Incoming data is then matched to one or more clusters using a partial pattern-searching algorithm with the intent of determining what pattern the time-series is likely to follow throughout the remainder of the day.

Clustering Algorithm

There are many techniques that may be used to cluster time-series data. For this experiment, a simple bottom-up algorithm is employed [1, 3]. This algorithm begins with a set of N daily patterns of dimension R . The distance between each of these sets is computed using some predefined distance measure, $f_d(p_1, p_2)$. The two sets with the smallest distance measure are then merged, creating a set of $N-1$ patterns [1, 3]. This process is then repeated until a desired

number of patterns is reached or the minimum distance between two patterns exceeds some threshold [1, 7, 10

```

Do While Termination Condition Not Met
  For each combination of two patterns
    Compute distance
  Next
  Merge closest two patterns
  Replace closes two patterns with the merged
pattern
End Loop

```

Figure 4: Pseudo-code for clustering algorithm

There are a number of methods that may be used to compute the distance between two patterns. Many of which are described in [1] and [3]. A modified version of a distance measure presented in [1] was utilized for this experiment. This measure as

presented in [1] is defined as.

$$f_d = \sqrt{\sum_{i=0}^R (p_{1i} - p_{2i} - \Delta)^2} / R \quad 1.$$

$$\text{Where } \Delta = \sum_{i=0}^R (p_{1i} - p_{2i}) / R \cdot \quad 2.$$

This measure provides a level of normalization to the patterns to place more emphasis on shape and less on magnitude. A slight modification to the algorithm was made to allow the distance measure to take the respective peaks of each pattern into account. The new distance measure is defined as:

$$f_d = \sqrt{\left((m_1 - m_2 - \Delta)^2 + \sum_{i=0}^R (p_{1i} - p_{2i} - \Delta)^2 \right) / (R+1)} \quad 3.$$

Where m is the maximum value of the pattern.

While a number of more advanced techniques may be used to cluster patterns, many of them were found to require more data than is available or are too inefficient.

First Stage Clustering

The purpose of the first stage of clustering is to create the groups into which subsequent data sets are to be categorized. The first step is to cluster sub-divide the time-series into distinct segments consisting of 96 data points per segment. (A single day divided into 15 minute divisions has 96 points.) This process will divide the time-series into N clusters, where N is some number less than the total number of days in the time-series. Clustering of this sort using statistical models was presented in [1] for the purposes of visualization.

Figure 5 illustrates data collected from an office building for the month of July. Upon initial inspection once can readily see distinct patterns for weekdays, Saturdays, and Sundays. Once can also note that July 4th, a U.S. holiday, looks much like a Sunday. When this data was run through the clustering algorithm, ten distinct patterns were found in the data set. These ten patterns create the set of categories that subsequent data will be classified in.

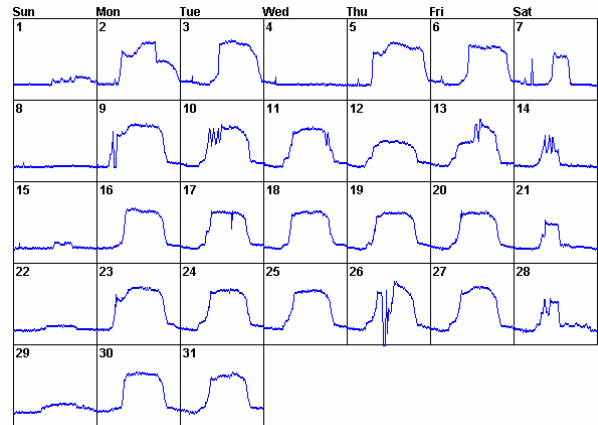


Figure 5: Data for July

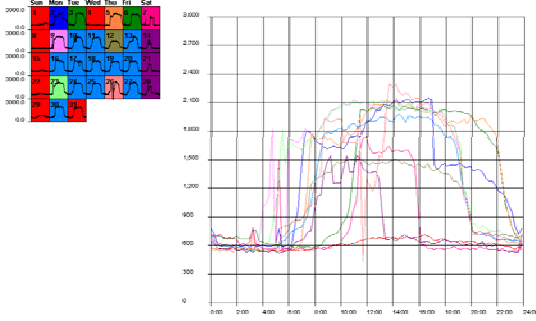


Figure 6: Clustering of July Data

Partial Pattern Matching

Once a set of classifications has been produced from the training data set, new data is fed into the system one point at a time. As each data point for a given day is acquired a partial pattern search is performed on the first M points of each known pattern where M is the number of points acquired for a given day. The partial pattern distance measure is defined as:

$$f_d = \sqrt{\left((\Delta)^2 + \sum_{i=0}^M (p_{1i} - p_{2i} - \Delta)^2 \right) / (M + 1)} \quad 4.$$

Equation 4 is derived from Equation 3 by making the assumption that if the patterns are assumed to be matched then the peak values are likely to be near equal. As such the two m terms will cancel out. The winning pattern is the closes pattern within the termination limit.

Experimental Results

The experiment involved two data sets. For each data set one month's data was used to train the system (generate the classifications). These models were then used to classify the subsequent 13 days of data. The system was configured to return two types of results. The first result was a winner take all partial pattern search in which the closest partial pattern match that was within the limiting threshold was returned as the matched pattern. The second output was a list of all patterns that fell within the limiting threshold. Table 1 shows the time that the winner take all method stabilized and the time

that the nearest neighbors search converged to a single pattern. If the search did not converge, an N/A is listed.

Conclusion

A method for utilizing pattern clustering techniques to perform early pattern classification was proposed in this paper. This method involved the use of a training data set and clustering algorithms to generate a set of base patterns. These base patterns were then used within an on-line partial pattern-matching algorithm to pre-classify a day into one of the base patterns. The technique was applied in one of two ways: winner take all and nearest neighbors. The winter take all technique consistently illustrated the earliest classification.

This technique has a number of applications in industry. It may be used to create alarm thresholds to notify individuals when the data being acquired exhibits abnormal behavior. The technique may be utilized to assist in demand forecasting. The technique is also useful in medicating risk in real-time energy trading.

The work done thus far is preliminary and may serve as a baseline upon which to judge the success of future endeavors. One area of research that may be investigated is the use of neural networks and fuzzy systems to perform the partial pattern matching. ART2 Networks may be applicable toward the partial pattern search. Additional techniques for performing the pattern clustering should also be investigated to improve speed and stability.

Data Set	Day	Time of Stable WTA	Time of Convergence
1	1	3:30 AM	3:30 AM
1	2	7:30 AM	10:00 AM
1	3	10:00 AM	12:30 PM
1	4	8:00 AM	10:15 AM
1	5	10:15 AM	1:00 PM
1	6	8:15 AM	N/A
1	7	7:30 AM	N/A
1	8	7:45 AM	9:15 AM
1	9	1:00 PM	1:00 PM
1	10	8:00 AM	12:15 PM
1	11	10:15 AM	11:30 AM

Data Set	Day	Time of Stable WTA	Time of Convergence
1	12	8:30 AM	11:49 AM
1	13	8:15 AM	1:00 PM
2	1	10:30 AM	10:30 AM
2	2	1:00 AM	12:15 PM
2	3	1:00 AM	8:15 AM
2	4	10:15 AM	11:45 AM
2	5	6:15 AM	11:45 AM
2	6	9:30 AM	12:30 PM
2	7	10:45 AM	11:30 AM
2	8	6:00 AM	11:45 AM
2	9	12:30 PM	12:30 PM
2	10	5:00 AM	8:30 AM
2	11	12:00 PM	12:00 PM
2	12	10:30 AM	10:30 AM
2	13	5:15 AM	3:15 PM

Table 1: Experimental Results Without On-Line Learning

References

[1] Van Wijk, Jark J., Edward R. van Selow., “Cluster and Calendar based Visualization of Time Series Data”, Proceedings of the 1999 IEEE Symposium on Information Visualization, pp. 4-9, 140, 1999.

[2] Keller, P.R. and M.M Keller, Visual Cues, IEEE Press, 1993 p. 188.

[3] Kaufman L. and P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley, 1990.

[4] Ware, Colin, Information Visualization: Perception for Design, Morgan Kaufman, 2000 p. 349.

[5] Adamopoulos, A.V., S.D Likothanassis, and E.G. Georgopoulos, “Extracting Structural Characteristics of a Nonlinear Timeseries Using Genetic Algorithms” *Proceedings of the IASTED International Conference on Intelligent Information Systems*, pp. 179 –183, 1997.

[6] Guimarães, G. “Temporal Knowledge Discovery in Multivariate Time Series with Enhanced Self-Organizing Maps,” *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks*, pp 165-170, 2000.

[7] Guoqing, Chen, Qiang Wei, and Hong Zhang, “Discovering Similar Time-Series

Patterns with Fuzzy Clustering and DTW Methods”, *Proceedings of the Joint 9th IFSAA World Congress and 20 NAFIPS International Congress*, pp 2160-2164, 2001;

[8] Keim, Daniel A., Han-Peter Kriegel, and Mihael Ankerst “Recursive Pattern: A Technique for Visualizing Very Large Amounts of Data” *Proceedings of the 6th IEEE Visualization Conference*, pp. 279 -286, 463, 1995.

[9] Li, Bin, Zhang Tan, Jinsong Lixiang, and Zhuang Zhenquan “Using Fuzzy Neural Network Clustering Algorithm in The Symbolization of Time Series”, *Proceedings of The 2000 IEEE Asia-Pacific Conference on Circuits and Systems*, pp. 379 –382, 2000

[10] Lin, Chin-Teng and C. S. George Lee, *Neural Fuzzy Systems*, Upper Saddle River, NJ, Prentice Hall, 1996

[11] Tachibana, Yuko and Mikihiro Ohnari, “Prediction Model of Hourly Water Consumption in Water Purification Plant through Categorical Approach,” *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, pp. 569-574. 1999.

[12] Volchan, Sergio B., Ashok N. Srivastava, Renjeng Su, and Andreas S. Weigend, “Data Mining for Features Using Scale-Sensitive Gated Experts,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 21, No. 12, pp. 1268-1279, Dec. 1999

[13] Giordano, Frank R. Maurice D. Weir, William P. Fox, “*A First Course In Mathematical Modeling* Brooks-Cole, 1997 pp 420 – 425.