Journal of Library and Information Studies 15:2 (December 2017) pp.21-44
doi: 10.6182/jlis.2017.15(2).021

A Symbolic Time-series Data Mining Framework for Analyzing Load Profiles of Electricity Consumption

I-Chin Wu[1], Tzu-Li Chen[2], Guan-Qun Hong[3],
Yen-Ming Chen[4], Tzu-Chi Liu[5]

Abstract

Electricity is critical for industrial and economic advancement, as well as a driving force for sustainable development. In turn, reducing energy consumption for sustainability and both tracking and managing energy efficiently have become critical challenges. In this research, we analyzed electricity consumption from the perspective of load profiling, which charts variations in electrical load during a specified period in order to track energy consumption of an annealing furnace in a co-operating steel forging plant. We made a preliminary proposal to use a symbolic time-series data mining framework for electricity consumption analysis. First, we adopted a piecewise aggregate approximation (PAA) approach to perform dimension reduction. Then, we refined the distance measure of the symbolic aggregate approximation (SAX) algorithm. SAX is a symbolic representation of time-series for dimensionality reduction and indexing with a lower-bounding distance measure. Our experimental results showed that the dimension reduction method known as PAA can better detect the state of the annealing furnace compared to the fixed feature point (FFP) method. In addition, the refined lower-bounding distance measure proved to be better than the traditional measure for calculating the similarity between energy load profiles. The results can help the plant conduct further normal and abnormal electricity pattern detection.

Keywords: Electricity Load Profiling; Piecewise Aggregate Approximation; Symbolic Aggregate Approximation; Time-series Data Mining

1. Introduction

Given the problems of gradual oil depletion and global warming, energy consumption has become a critical factor for energy-intensive sectors, especially the semiconductor, manufacturing, iron and steel, and aluminum industries. Indeed, social development correlates positively with power consumption, which in Taiwan, especially the consumption of electricity, has risen rapidly due to economic, industrial, and commercial growth.

In relation to total exports, Taiwan's manufacturing-oriented economy exports a considerable share of manufactured goods. Currently, most industries in Taiwan have replaced manual operation with machine operation during fabrication, which requires a sufficient but not excessive supply of stable electricity. In fact, too much or too little electricity can cause mechanical

[1] Graduate Institute of Library and Information Studies, National Taiwan Normal University, Taipei, Taiwan
[2,3] Department of Information Management, Fu-Jen Catholic University, New Taipei, Taiwan
[4,5] Industrial Technology Research Institute, Hsinchu, Taiwan
* Corresponding Author: I-Chin Wu, E-mail: icwu@ntnu.edu.tw

malfunctions and thereby reduce the efficiency of both production. As Table 1 shows, Taiwan Power Company's statistics from 2015 reveal that the industrial sector consumes an exceptionally large proportion of electricity—even up to more than 50% of the total consumed in Taiwan. Specifically, according to the statistical data of the Bureau of Energy, the industrial sector accounts for the largest proportion of the total energy consumption by sectors in Taiwan for the past several years, i.e., 53% in 2016.

In response, manufacturers in Taiwan, is keen to identify the most cost-effective methods and techniques to increase electricity efficiency in their factories. In industries, many machines are highly energy intensive, and with machine data, we can analyze their tendencies regarding power and temperature, among other measures. We can also use anomaly detection to identify indicators of machine malfunction, which can then contribute to determining rules in order to explain the malfunctions. With such technologies, we can promptly correct abnormalities and thereby reduce the unnecessary waste of resources and improve the efficiency of electric consumption.

Table 1. Electricity Sales in Taiwan, 2015

Industry sector	GWh	(%)
Industrial	114,241.9	55.3
Residential	42,196.6	20.4
Commercial	32,511.0	15.7
Other	17,541.8	8.5
Total	206,491.3	100.0

Note. Adapted from "Electricity consumption," by Ministry of Economic Affairs, Bureau of Energy, 2017, Retrieved from http://web3.moeaboe.gov.tw/ecw/populace/web_book/WebReports.aspx?book=M_CH&menu_id=142

Without a doubt, energy is a vital resource for modern society, especially for long-term competitive sustainability. To reduce unnecessary energy consumption and improve energy efficiency, it is therefore critical to make informed decisions in real time. To that end, we collected data regarding energy consumption as well as information from the corresponding production and manufacturing domains from the plans of co-operating iron and steel manufacturers. Based on load profiles determined from data stream mining and machine learning techniques, we constructed an electric energy monitor and analysis framework, centered on a prediction model for identifying typical load profiles of each machine and a time-series data-mining engine for analyzing and extracting typical patterns based on the load profiles. The objectives of our research were fourfold:

1. To propose and construct an electric energy monitor and analysis framework, which is based on load profiles by data stream mining and machine learning techniques as a means to implement the proposed symbolic time-series data-mining approach in co-operating iron and steel manufacturers.
2. To observe and analyze relationships among various attributes (e.g., electric power, temperature, and product weight) in a data warehouse framework to allow researchers to select and confirm key attributes based on the results of analysis and consult with domain experts.
3. To identify three states of the annealing process—warm-up, heat retention, and cooling—based on the temperature information of the operating machine and, following Keogh,

Chakrabarti, Pazzani, and Mehrotra (2001), use piecewise aggregate approximation (PAA) to perform dimension reduction of time-series data for data representation.

4. To symbolize the time-series of electricity data by the symbolic aggregate approximation (SAX) algorithm (Lin, Keogh, Lonardi, & Chiu, 2003) and then refine the lower-bounding distance measure that calculates the similarity among symbolized energy load profiles. Finally, we adopted the tightness of lower bound (TLB) measure to evaluate the performance of our refined distance measure.

The rest of the paper is organized as follows: First, Section 2 briefly reviews the important research issues of time-series data mining. Section 3 illustrate the symbolic time-series data mining framework for electricity consumption analysis in this work and proposes the research problems. The electricity time-series data mining and analytics methods, i.e., the PAA approach and SAX algorithm, will be detailed in Section 4. Section 5 shows the experimental results for identifying the states of the annealing process and the performance of our refined distance measure for calculating the similarity among symbolized energy load profiles based on a real case study. Finally, Section 6 provides a summary and discusses some potential future research directions.

2. Related Works

2.1 Basic concepts of time-series data mining

Time-series data are easily obtainable from scientific, financial, and industrial applications, and given the deployment of numerous sensors and smart devices, the amount of accumulated time-series data continues to expand rapidly. By extension, the increased generation and use of time-series data have resulted in a great deal of research and developments in big data mining. Each time-series database consists of sequences of values or events obtained over repeated measurements of time (Han, Kamber, & Pei, 2011). Time-series data are large, as well as numerical and continuous in nature, which require continuous updating. Mörchen (2006) has identified two chief research-related goals of time-series analysis—to identify patterns represented by the sequence of observations and to forecast future values of time-series data—both of which require the identification of patterns of time-series data to enable the interpretation and integration of patterns with other data.

Kitagawa (2010) classified time-series analysis into four categories: description, modeling, prediction, and signal extraction. First, *description* refers to methods that effectively express or summarize the characteristics of time-series and can involve drawing figures of time-series or computing basic descriptive statistics (e.g., sample autocorrelation functions, sample autocovariance functions, and periodograms). Second, *modeling* requires the selection of a proper model class to estimate parameters in the model and generally depends upon the characteristics of the time-series and the objective of its analysis. Third, *prediction* enables the estimation of the future pattern of a time-series by extracting various current and past information in order to calculate correlations over time and among the variables. Fourth and lastly, *signal extraction* involves extracting essential signals or useful information from time-series according to the objective of analysis tasks. In a similar vein, Sakurai, Yamamuro, and Faloutsos (2015) have provided a comprehensive overview of key topics of time-series analysis: similarity search and pattern discovery, linear modeling and summary, nonlinear modeling and forecasting, and the extension of time-series mining and tensor analysis. In our research, we focused on the first, i.e., similarity search and pattern discovery.

Vikhorev, Greenough, and Brown (2013) has proposed a framework for energy monitoring and management in the factory. The framework incorporated standards for energy data exchange, on-line energy data analysis, performance measurement, and display of energy usage. Two methodological approaches were adopted in their research; that is, an action research (AR) framework and case study research. The AR can ensure a structured research process and its continual improvement. The research conducted a real case study by deploying the framework via a prototype information system in a machining line of a major European automotive manufacturer. The three operational states were identified by the manufacturing execution system (MES) and evaluated their energy usage patterns with an associated visualized interface. Le et al. (2012) has proposed a two-stage framework to identify six operational states of machines based on a real-time energy consumption pattern. The research adopted a finite-state machine (FSM) to model the industrial processes of six operational states: switch-off, warm-up, idle, pumping and heating, start-up, and production. The proposed framework is evaluated on two industrial injection molding machines by using a Savizky–Golay filter for advanced signal processing of energy measurement data and a neural network for classifying energy consumption patterns. The experimental results show that it can achieve 95.85% in identification of machine operational states and can detect abnormal energy patterns. Popeangă (2015) has proposed that energy production and consumption data recorded over a period at fixed intervals is a classic time (i.e., chronological) series data-mining problem. The entire process involves five steps: collecting data from various sources (e.g., the Internet, text, databases, data warehouses, sensors, and smart devices); conducting data filtering by eliminating errors or deploying a data warehouse to create an extraction, transformation, and loading (ETL) process in advance; selecting key attributes to be used in data mining for further analysis; detecting and analyzing new knowledge; and visualizing, validating, and evaluating results. The challenge of electricity consumption analysis is analyzing countless time-series to find similar or regular patterns and trends with a fast or even real-time response. Accordingly, time-series data mining techniques such as whole series clustering and classification, subsequent clustering and classification, time point clustering, anomaly detection, and motif discovery can be adopted for electricity consumption analysis and energy management. McLoughlin, Duffy, and Conlon (2015) adopts three clustering methods—k-means, k-medoid, and Self Original Map (SOM)—to analyze domestic electricity load Profile Classes (PC) for identifying various common patterns of home electricity usage. A Davies–Bouldin (DB) validity index was used to identify the most appropriate clustering method and the number of clusters. Finally, a multi-nominal logistic regression was applied on each PC to examine the influence of household characteristics on electricity use. The above studies provide good insights into building energy analytics and management frameworks by data stream analysis based on the energy load files. Then, the data mining related techniques can help build the model for detecting machines' operational states, analyzing energy consumption patterns, and executing energy management.

2.2 Data representation and similarity measures for time-series data

When dealing with time-series data efficiently, it is important to develop data representation techniques that can reduce the dimensionality of time-series, but still preserving the fundamental characteristics of the data (Ding, Trajcevski, Scheuermann, Wang, & Keogh, 2008; Han et al., 2011). Since time-series are high-dimensional data, they are time consuming for computing and have a high storage space cost. However, several techniques have been proposed that denote time-series data with reduced dimensionality. Well-known dimensionality reduction techniques include discrete Fourier transformation (Faloutsos, Ranganathan, & Manolopoulos, 1994), single value decomposition (Wall, Rechtsteiner, & Rocha, 2003), discrete wavelet transformation (Chan & Fu, 1999), PAA (Keogh et al., 2001), adaptive piecewise constant approximation (Keogh et al., 2001), SAX (Lin, Keogh, Wei, & Lonardi, 2007), and indexable piecewise linear approximation (Chen, Chen, Lian, Liu, & Yu, 2007). In this research, we adopted the intuitive method of PAA and discretized the PAA representation of a time-series into a symbolic representation method SAX algorithm. SAX transforms time-series into a symbolic string SAX, which is the first symbolic representation method for dimensionality reduction and indexing with a lower-bounding distance measure (Lin et al., 2003). The algorithm entails two primary steps: transforming the original time-series into the PAA representation and converting the PAA data into a string.

Another important issue of time-series data mining is determining the similarity or distance among the time-series data. Note that similarity and distance are two relative concepts; however, distance has to be non-negative. There are lots of distance measures for calculating the similarity of time-series data in the literature, e.g., Euclidean distance (ED) (Faloutsos et al., 1994), Dynamic Time Warping (DTW) (Berndt & Clifford, 1994), distance based on Longest Common Subsequence (LCSS) (Vlachos, Kollios, & Gunopulos, 2002), Edit Distance on Real sequence (EDR) (Chen, Özsu, & Oria, 2005), Edit Distance with Real Penalty (ERP) (Chen & Ng, 2004), Spatial Assembling Distance (SpADe) (Chen, Nascimento, Ooi, & Tung, 2007) and similarity search based on Threshold Queries (TQuEST) (Aßfalg et al., 2006). Ding et al. (2008) further categorized similarity measures as lock step, elastic, threshold-based, and pattern-based measures.

Euclidean distance, one of the lock-step measures, is the most common distance measure for time-series data and is surprisingly competitive with other more complex approaches, especially if the size of the training set/database is relatively large. However, since the mapping between the points of two time-series is fixed, these distance measures are very sensitive to noise and misalignments in time (Ding et al., 2008). Another well-known algorithm is Dynamic Time Warping (DTW), one of the elastic measures, can find an optimal alignment between two time-dependent sequences (Berndt & Clifford, 1994) and is a much more robust distance measure for time-series. In addition, DTW allows a time-series to be stretched or compressed to more precisely match with another one out of phase. Furthermore, several lower bounding measures have been introduced to speed up the similarity search using DTW (Keogh & Ratanamahatana, 2005; Nath & Baruah, 2014; Yi, Jagadish, & Faloutsos, 1998). Currently, tightness of lower bounding measure (TLB) is wildly used to compare the performance of data representation methods (Ding et al., 2008); that is, it measures the ratio of lower bound distance to the actual DTW or Euclidean distance. The ratio is in the range [0,1]. The higher the ratio, the tighter is the bound. Based on previous research (Ding et al., 2008; Lin et al., 2003), we adopted the TLB measure to evaluate the performance of our adjusted distance measure using SAX algorithm.

3. Symbolic Time-series Electricity Consumption Data Mining Framework

We collected energy consumption data and the corresponding product information of an annealing furnace in 2014. Figure 1 shows the proposed

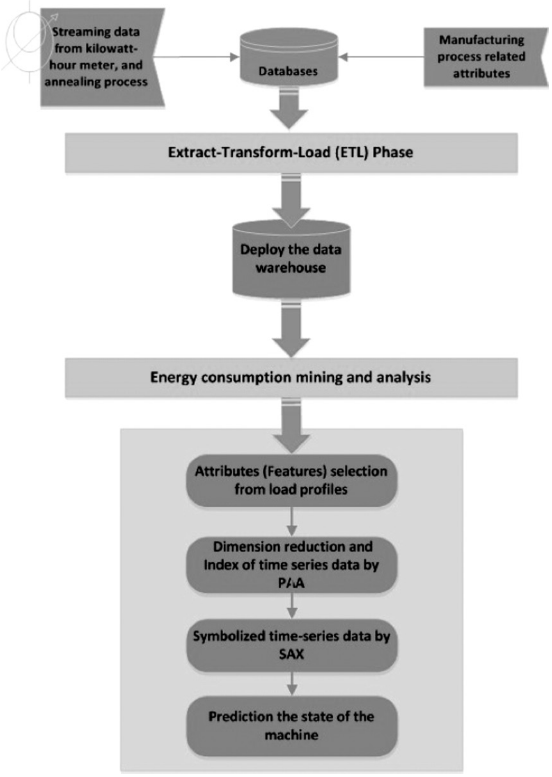
Streaming data from kilowatt-hour meter, and annealing process
Databases
Manufacturing process related attributes
Extract-Transform-Load (ETL) Phase
Deploy the data warehouse
Energy consumption mining and analysis
Attributes (Features) selection from load profiles
Dimension reduction and Index of time series data by PAA
Symbolized time-series data by SAX
Prediction the state of the machine

Figure 1. A Symbolic Time-series Data Mining Framework for Electricity Consumption Analysis

symbolic time-series data mining framework for electricity consumption analysis. The primary research questions were:

- What is a good attribute to identify the operational state of the machine?
- What is the best method of PAA approach to predict the operational states of machines (i.e., warm-up, heat retention, and cooling)?
- Evaluate the performance of the modified SAX algorithm compared to the typical one by tightness of lower bound (TLB) measure. Basically, we refined the distance measure adopted by SAX algorithm to calculate the similarity among energy load profiles.

All tasks of analysis involved using the load profiles of the electricity consumption of the targeted machine. For the proposed framework, we preliminarily deployed the data warehousing framework to observe and analyze the load profiles of electricity consumption and the relationships among various attributes (e.g., electric power, temperature, and product weight). Next, we selected and confirm key attributes to identify the state of the annealing furnace based on the results of analysis and consulted with domain experts. We confirmed that either the electric power or temperature information of the operating machine can help to identify the entire machine operational process, which is 1,440 minutes on average. We preliminary used the temperature information of the operating machine to identify three machine operational states: warm-up, heat retention, and cooling.

We applied the PAA method to discretize streaming data into *w* segments with timestamps in order to build the prediction model. We also refined the distance measure in SAX algorithm for dimensionality reduction and indexing with a lower-bounding distance measure to further extract subsequent patterns. We conducted a series of experiments to construct a prediction model in order to identify their operational states (i.e., warm-up, heat retention, and cooling), the target annealing furnace. We also included associated experiments of parameter selection of the PAA approach and SAX algorithm in our experiments. Ultimately, the goal of our series of studies is to deploy a visualized decision support system and propose actionable energy-saving strategies for the co-operating iron and steel plant to solve real-world problems. We present the entire framework for electricity consumption analysis and detail some of the modules in the following sections.

4. Data Preprocessing and Data Warehousing Deployment

4.1 Data preprocessing

Table 2 presents all of the attributes of the annealing furnaces related to electricity consumption analysis in our research. We adopted a data mart to visualize and observe the initial load profiles of electricity consumption. In general, data warehousing is fundamental to business intelligence, and data collection, data management, and data analysis techniques (e.g., data mart design with extraction, transformation, and loading tools) can help business analytics use data intelligently. Accordingly, we deployed the data warehousing framework to observe the load profiles of electricity consumption and analyzed the relationships among various attributes (e.g., electric power, temperature, and product weight.) The data warehousing platform had two

Table 2. Electricity Consumption Analysis Related Attributes

Attributes	Data type
Logtime	Date yyyy/mm/dd hh:mm:ss
Current (I_avg)	Numeric
Voltage (V_avg)	Numeric
Active power (kW_tot), total active power (kWh_tot)	Numeric
Reactive power (kvar_tot), total reactive power (kvarh_tot)	Numeric
Apparent power (kVA_tot), total apparent power (kVAh_tot)	Numeric
Power factor (PF_tot)	Numeric
Temperature	Numeric
Product weight	Numeric

chief goals: to analyze the load profiles of each annealing process and to define annealing states based on the selected attributes of load profiles.

Data warehousing helped us to confirm the load profiles of each annealing process in order to preliminarily identify the normal or abnormal state of the machines. We confirmed that either the electric active power or temperature information of the operating machine can help to identify the entire machine operational process, which is 1,440 min on average. After selecting the attributes that were useful for periodical data analysis, we adopted the star schema to build the data mart. The three dimension tables are the machine information table, the product information table with time information with different granularity table, and a fact table that shows the load profiles of current and temperature, among other things. Based on the analytical results of load profile, we used the temperature information of the operating machine to divide three operational states: warm-up, heat retention, and cooling. By extension, we could further identify the normal or abnormal states of each annealing process. We showed one load profile of active power and temperature of an entire annealing process in Figure 2.

4.2 Time-series representation for the load profile of electricity consumption

4.2.1 Time-series representation

To represent time-series data concisely and increase the index and processing times, we mainly adopted PAA in order to extract the primary features of time-series data (Keogh & Pazzani, 2000; Keogh et al., 2001).

We treated each annealing process as having streaming time-series data that are divisible based on the differing granularity of time units, each of which is a feature point of the data stream. Accordingly, an annealing process entails several feature points with timestamps. Herein, we introduce two methods to extract feature points: a fixed interval method as a baseline method and the PAA of a time-series. For the fixed interval method, if the length of the string was 1,000 and we aimed to extract 5 points, then we extracted the first, 250th, 500th, 750th, and 1000th points,

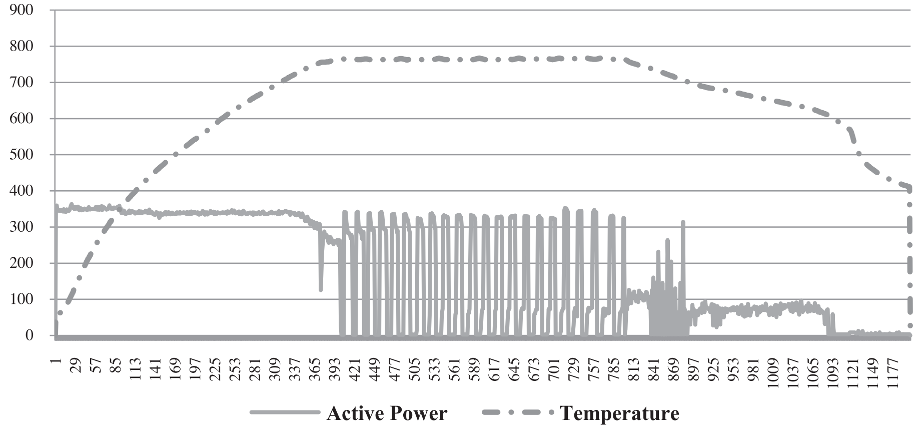

Figure 2. Load Profiles of Active Power and Temperature

in a method we called it as the fixed feature point (FFP) method. Figure 3 shows an example of the FFP representation curve. For PAA, we averaged the values of points in a fixed interval to represent a feature point (Figure 4). PAA is a non-data-adaptive representation model that transforms the time-series into a different space and has the same transformation parameters regardless of features of the data at hand (Kleist, 2015). Put differently, the transformation parameters are preset without consideration of the underlying data. We further adopted SAX after PAA to represent each feature point of the load profile symbolically.

4.2.2 Feature frames of each annealing process

Based on the methods, we defined time-series data and related notations (Table 3). We denoted time-series data of an attribute i as S = (s_1, s_2, ..., s_n), with the length of a time-series in n and w as the dimensionality of the space to index the time-series data. Put differently, a time-series of length n can be represented in w dimensional space and each feature point by a feature frame of fix length (i.e., n/w). For PAA, the result is $\overline{S} = (\overline{s}_1, \overline{s}_2, ..., \overline{s}_w)$—that is, w-dimensional space by vector $\overline{S}$. The ith feature point of $\overline{S}$ can be derived from Equation (1).

$$\overline{s_i} = \frac{w}{n} \sum_{j=\frac{w}{n}*(i-1)+1}^{\frac{w}{n}*i} s_j \qquad (1)$$

The attributes selected in a feature frame comprised all extracted points of active power (FP_A) and the minimum, maximum, and average values of active power; all extracted points of temperature (FP_T) and the minimum, maximum, and average values of temperature; and the weight of raw material information. The feature frame was the input of the training model. The specific notation, with a description of each attribute set of the feature frame, appears in Table 4. Attributes derive from the fact table in our designed data mart.

4.3 Symbolize the load profiles of electricity consumption data

Data size reduction techniques are helpful in the process of categorizing the electrical load consumption patterns on the basis of their shape.

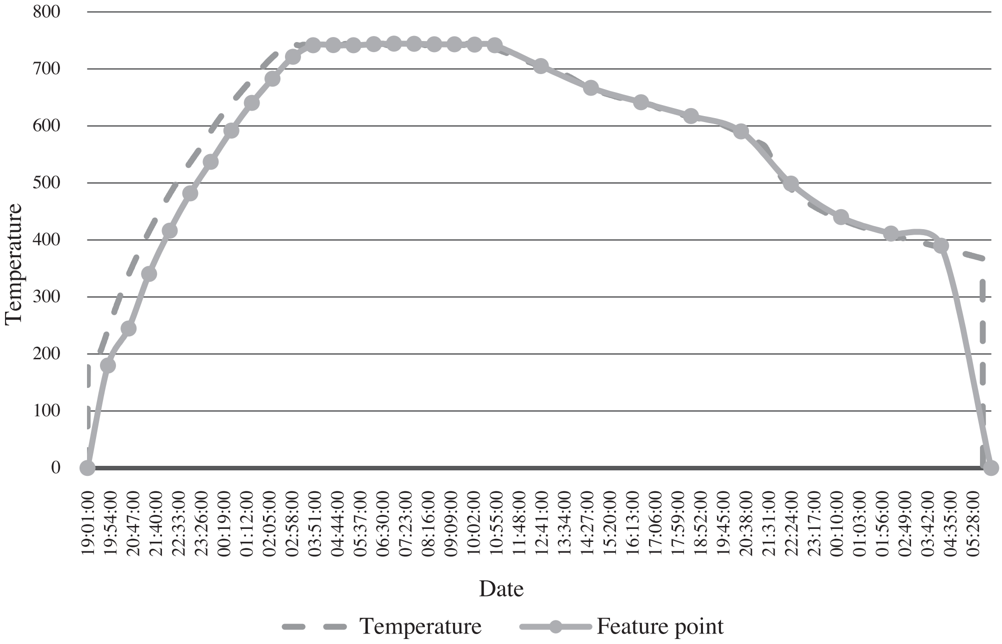

Figure 3. FFP Method for Feature Point Extraction (Temperature)

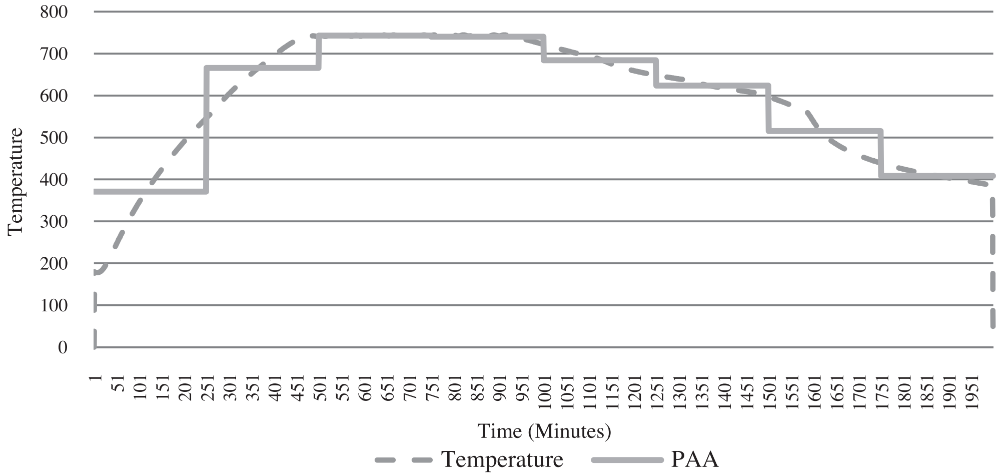

Figure 4. PAA Method for Feature Point Extraction (Temperature)

Table 3. Summary of Notation Used in PAA and SAX

Notations	Definitions
S_i	A time-series of length n, $S_i = (s_1, s_2, \ldots s_n)$
w	The dimensionality of the space, $1 \leq w \leq n$ That is, the *FFP* or *PAA* segments representing a time-series S
FF (feature frame)	A feature frame composed by set of attributes
$\overline{S}$	A piecewise aggregate approximation of a time-series
FP_A (feature point of active power)	A time-series of the active power of length w after dimension reduction, FP_A = $(fp_{a1}, fp_{a2}, \ldots fp_{aw})$
FP_T (feature point of temperature)	A time-series of temperature of length w after dimension reduction, FP_T = $(fp_{t1}, fp_{t2}, \ldots fp_{tw})$

Table 4. Summary of the Notation of the Feature Frame

Notations	Definitions
PMin	Minimum value of the active power of a state
PMax	Maximum value of the active power of a state
PAvg	Average value of the active power of a state
P_N	Number of extracted dimensions in a state of the active power
TMin	Minimum value of temperature of a state
TMax	Maximum value of temperature of a state
TAvg	Average value of temperature of a state
T_N	Number of extracted dimensions in a state of temperature
PWeight	Weight of materials for each operational process
PTime	Duration of each state of the entire operational process

The SAX algorithm is a symbolic representation for time-series and uses a synthetic set of symbols to reduce the dimensionality of the numerical series. The SAX follows a two-step process: (1) Piecewise Aggregate Approximation (PAA) and (2) conversion of a PAA sequence into a series of letters. PAA divides the data set of length n into w equally spaced segments or bins, and computes the average of each segment. This essentially means that we reduce the number of dimensions from n to w. For the specific example here, the solid line is before the dimensionality reduction and the dashed line is after the dimensionality reduction. The dimensionality has been reduced from 2000 to 8, as shown in the Figure 5 with discrete horizontal line segments, with each segment representing an aggregate 250 consecutive data points in the time-series. The SAX converts

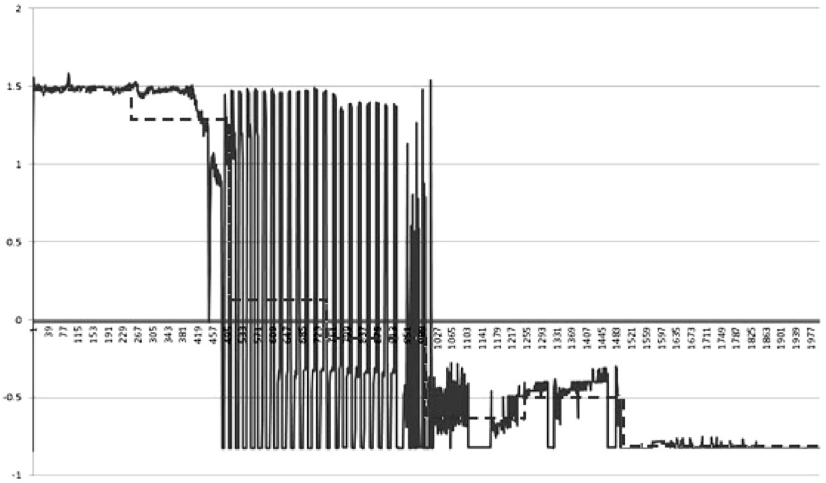

Figure 5. Dimensionality Reduction by PAA

different numerical data to different alphabet symbols through the breakpoint table proposed by Lin et al. (2007), as shown in Figure 6. The definition of the breakpoints references Lin et al. (2007). In Figures 5 and 6, the unit of x-axis is time in terms of minutes and y-axis is the value of *z*-score standardization for electricity consumption data.

Herein, we give an example to explain how to symbolize time-series data with the aid of the breakpoint table. If we decide to set α as equal to 3, there are three letters, A, B, and C to represent the time-series data in the given example. The SAX converts different numerical data to different symbols through the breakpoint table in Figure 6. Figure 7 shows a load profile of our electricity consumption data that was discretized by the PAA approach and then the PAA coefficients were mapped into SAX symbols using predetermined breakpoints. In this example, with n = 128, w = 8 and a = 3, the time-series was mapped into the word CCBBAAAA. Note that the solid line is before the dimensionality reduction and the dashed line is after the dimensionality reduction. The dimensionality has been reduced from n to 8, as shown in Figure 7. Noted that the solid line denotes raw data whereas the dashed line denotes PAA curve.

4.4 Modify the distance measure of SAX algorithm

Herein, we discussed the issues related to measuring the similarities between the time-series data after conducting the SAX algorithm. The most common distance measure for time-series data is the Euclidean distance. The weakness of the Euclidean distance is its sensitivity to distortion in the time axis (Keogh & Ratanamahatana, 2005), that is, when there are two time-series sequences which have an overall similar shape but are not aligned in the time axis. Given two time-series *T1* and *T2,* with the same length, *n,* we then conducted dimension

β_i \ *a*	3	4	5	6	7	8	9	10
β_1	-0.43	-0.67	-0.84	-0.97	-1.07	-1.15	-1.22	-1.28
β_2	0.43	0	-0.25	-0.43	-0.57	-0.67	-0.76	-0.84
β_3		0.67	0.25	0	-0.18	-0.32	-0.43	-0.52
β_4			0.84	0.43	0.18	0	-0.14	-0.25
β_5				0.97	0.57	0.32	0.14	0
β_6					1.07	0.67	0.43	0.25
β_7						1.15	0.76	0.52
β_8							1.22	0.84
β_9								1.28

Figure 6. A Lookup Table of Breakpoints for the Corresponding Cutoff Lines from Dimensions 3 to 10 (Lin et al., 2007)

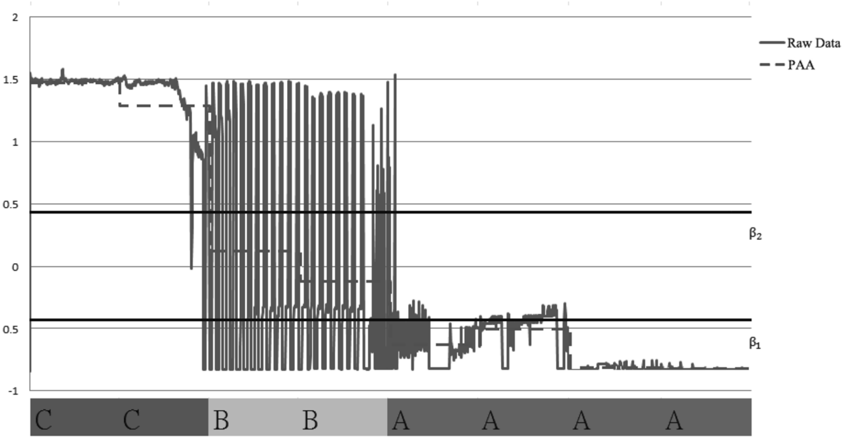

Figure 7. An Example of a Load Profile of Symbolize Electricity Consumption

reduction using the PAA approach to transform the original *T1* and *T2* into *T1'* and *T2'*, respectively, as shown in Figures 8 (A) and 8 (B). Based on Chakrabarti, Keogh, Mehrotra, and Pazzani (2002), we obtained a lower bounding Euclidean distance approximation between the original time-series data by Equation (2) as illustrated in Figure 8 (B). The lower bounding Euclidean distance measure can be applied using the reduced-dimension time-series representation method, which ensures the reduced dimension can be less than or equal to the true distance on the raw time-series data (Ding et al., 2008).

$$D_{LB}(T_1',T_2') = \sqrt{\frac{n}{w}}\sqrt{\sum_{i=1}^{w}(t'_{1i}-t'_{2i})^2} \qquad (2)$$

We further transformed the data into the symbolic representation, i.e., SAX, with a lower-

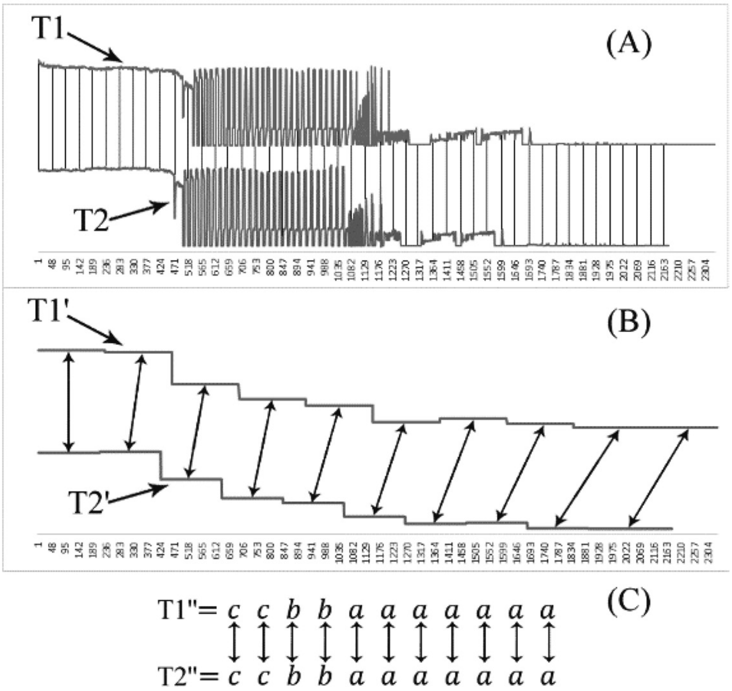

Figure 8. The Example Extracted from Our Data. (A) The Euclidean Distance between Two Operational Annealing Processes. (B) The Distance Measure after PAA. (C) The Distance of Two Sequences of Symbols after Conducting SAX Algorithm.

bounding distance measure. In the SAX algorithm Lin et al. (2007) defines a D_{min_dist} function to calculate the minimum distance between two sequences of symbols shown as shown in Equation (3).

$$D_{\min\_dist}(T_1",T_2")=\sqrt{\frac{n}{w}}\sqrt{\sum_{i=1}^{w}dist((t"_{1i}*i-t"_{2i}*i))^2} \qquad (3)$$

Table 5 shows a look up table using 4-letters of the alphabet for each cell in Table 5 for calculating distance between two sequences of symbols. That is, the distance between two symbols can be read off by checking the corresponding row and column. The equation is shown in Equation (4) as below.

$$Dist(R,C)=\begin{cases} 0,if\left|R-C\right|\le 1 \\ \beta_{\max(R,C)-1}-\beta_{\min(R,C),otherwise} \end{cases} \qquad (4)$$

Where βi is the element of the breakpoint list $\mathbf{B}=(\beta_1, \beta_2\ , \ldots \beta_{W-1})$ and $\beta_{i-1}<\beta_{i..}$ Four alphabet letters be defined using 3 breakpoints. In addition, R denotes row and C denotes column.

In this study, we improved the D_{min_dist} function by considering the variance of time-series data to get the new distance of two continuous strings. In Equation (4), the distance between two consecutive symbols is set to zero. Herein, we

Table 5. A Lookup Table Used by the D_{min_dist} Function

	a	b	c	d
a	0	0	0.67	1.34
b	0	0	0	0.67
c	1.67	0	0	0
d	1.34	0.67	0	0

refined this part by considering the variance of each time-series and incorporating the information into the equation, as shown in Equations (5) and (6). That is, if the two symbols are continuous letters, their distance is essentially the average variance of two corresponding time-series data, as shown in Equation (6). Then, we used the maximum variance of each time-series data, i.e., max (var_all), divided by the average variance of T1 and T2 due to the limited value range of [0,1]. Moreover, we multiplied the *Var*(*T*1, *T*2) with *p*. The parameter, *p,* is a parameter used to control whether the value will be equal to or less than the distance value of the value of two consecutive or same symbols in the breakpoint table.

$$Dist'(R,C)=\begin{cases} Var(T1,T2)\times p, if\,|R-C|\leq 1 \\ Var(T1,T2)\times(\beta_{\max(R,C)-1}-\beta_{\min(R,C)}), otherwise \end{cases} \tag{5}$$

Where

$$Var(T1,T2)=\frac{\max(\mathrm{var}\_all)}{(\mathrm{var}(T1)+\mathrm{var}(T2))/2} \tag{6}$$

Where 'β' is the breakpoint value, 'max(*var*)' is the maximum variance of all time-series, *var*(*T*1) is the variance of time-series T1, and *var*(*T*2) is the variance of time-series T2. Noted that the data range of *p* will depend on the α value of SAX algorithm, as shown in Figure 7. For example, we set α to be 10 in our electricity consumption data. Thus, the smallest distance value between 9 breakpoint values will be 0.25. Accordingly, the value range of *p* will be [0, 0.25] based on Figure 6 and Equation (4). We conducted the experiment to test if the *p* value will influence the experimental result.

5. Experimental Design and Results

5.1 Experimental setup

We next conducted a series of experiments to construct a prediction model in order to identify operational states for the target annealing furnace. We collected electricity consumption data and corresponding product information of an annealing furnace during April 1–December 31, 2014. The industry of our co-operating plan was steel forging, located in the middle of Taiwan. The company had more than 25 years of experience in designing and developing custom die and manufacturing original equipment manufacturer (OEM) forged products. Herein, we focused on the process of an annealing furnaces, which consumes a large among of electricity, making it important to analyze its operational efficiency and electricity consumption conditions. Annealing is a heat treatment process used to relieve internal stresses, induce ductility, and improve the mechanical properties of cold forged products. In this research, there were 70 records of the annealing processes of the operating machine collected in total. We mainly analyzed

Table 6. Prediction the Operational State by FFP Method in Terms of RMSE

Method/w	150	200	250	300	350	Average
baseline_FFP	0.108	0.085	0.086	**0.083**	0.100	0.092
normalization _FFP	0.497	0.337	0.344	0.365	0.422	0.393
hybrid_FFP	0.062	0.059	0.065	0.060	0.064	**0.062**
Average	0.222	**0.160**	0.165	0.169	0.195	

Table 7. Prediction the Operational State by PAA Method in Terms of RMSE

Method/w	150	200	250	300	350	Average
baseline_PAA	0.126	0.150	0.211	0.115	0.129	0.146
normalization _PAA	0.090	0.109	0.164	0.104	0.126	**0.119**
hybrid_PAA	0.127	0.119	0.117	0.107	0.172	0.128
Average	0.114	0.126	0.164	**0.109**	0.142	

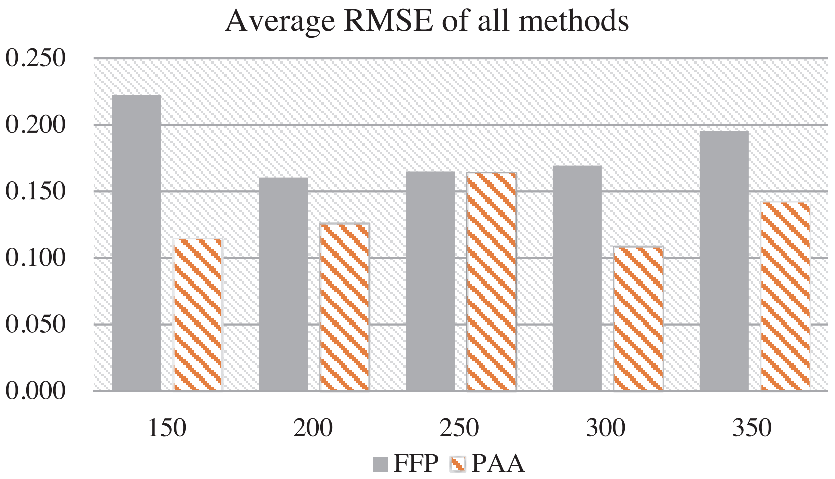

Figure 9. Comparison of Average Results of Each Approach with the FFP Method and PAA Method

the electric power and temperature information of the operating machine during the entire machine operational process, which was 1,440 minutes on average. Among 70 machine operational processes, there were 36 normal and 34 abnormal processes tagged based on load profiles. We used the data to train and construct the prediction model to detect each machine's three operational states during the annealing process—warm-up, heat retention, and cooling. We adopted five-fold cross-validation to evaluate the root mean squared error (RMSE) of the prediction results.

In what follows, we present the experimental results for one furnace, which we explain in terms of two primary sets of experiments with the FFP method and PAA as feature extraction methods.

Experiment 1 (FFP): The first set of experiments included original stream data, data regarding the normalization process, and data regarding both normalization process, and the extreme value removal process—respectively, baseline_FFP, normalization_FFP, and hybrid_FFP.

Experiment 2 (PAA): The second set of experiments included original stream data, data regarding the normalization process, data regarding both normalization process, and the extreme value removal process—respectively, baseline_PAA, normalization_PAA, and hybrid_PAA.

The purpose of data normalization with *z*-score standardization was to remove outlier data points and elucidate the relationship between a data point and the average value of all data points. The *z*-score converted all indicators to a common scale with an average of 0 and standard deviation of 1. The equation for calculating the *z*-score appears in Equation (7).

$$Normalized\left(e_i\right)=\frac{e_i-\bar{E}}{std(E)} \tag{7}$$

In which e_i is the data points of the load profile, *std(E)* is the standard deviation of the data points of the load profile, and is the mean value of the data points. The purpose of removing outlier values was to avoid excessive noise in time-series data. We removed feature points outside twice the standard deviation of the average value, $\bar{E}$, of the target load profile. Ultimately, the hybrid FFP and PAA methods involved adopting the *z*-score and removing outlier data points.

We adopted sequential minimal optimization, in which a multilayer perceptron (MLP) is a feedforward artificial neural network model, as well as a radial basis function (RBF). We tuned different learning rates to train the best MLP model and adopted five-fold cross-validation to evaluate the root mean squared error (RMSE) of the prediction results. We used the RMSE, as the mean of the square of all errors, to measure differences between values.

5.2 Experiment 1: Identifying operational states of the machine

Tables 6 and 7 show the average results of the three data mining approaches (i.e., MLP, radial basis function, and sequential minimal optimization) for the FFP method and PPA. We discretized the time-series data into *w* points and listed the results of each variation method based on the FFP method and PAA. When we set *w* to 150, for example, we extracted 150 feature points to represent the entire load profile of the active power.

Observation 1 (FFP method): For the FFP method, the worst FFP-based method on average was *normalization_FFP*. By contrast, the *hybrid_*

FFP achieved the minimum RMSE better than the other two methods under various *w* value settings. Overall, the best results on average occurred when *w* was 200, and it seems that a larger *w* value (i.e., more feature points) with the FFP method does not generate better results in predicting the machine states.

Observation 2 (PAA): The worst method for PAA was *baseline_PAA*. Both *normalization_PAA* and *hybrid_PAA* achieved similar RMSE results which are better than the baseline under various *w* values. Overall, the best results on average were with *w* at 300 on average. The FFP method seems insensitive to *w*-values; however, more or fewer feature points did not yield better results in predicting the machine states.

Observation 3 (Comparison): Figure 9 shows a comparison of the average RMSE between the FFP method and PAA. When we compared the methods in terms of the approaches, we observed that the FFP method is worse than PAA, because the former is more sensitive than the latter to extract points in representing subsequent parts of the data stream. Apparently, PAA can accommodate a smaller RMSE than the FFP method with most *w*-values. As such, we adopted PAA to further symbolize processing with the SAX algorithm and set *w* to 150 or 300 (i.e., 150 or 300 feature points to represent the entire data stream).

Based on experimental results, we will adopt the *normalization_PAA* to execute the SAX algorithm in order to achieve better effectiveness and efficiency with the experimental results.

5.3 Experiment 2: Effectiveness of modifying the distance measure of the SAX algorithm

We preliminary examined the degree of distortion of two time-series data with the SAX algorithm and modified SAX algorithm proposed in Sections 4.3 and 4.4. To do that, we first introduced the evaluation metric (i.e., tightness of the lower bound, or TLB).

5.3.1 Distance evaluation metrics: Measuring the tightness of lower bound

Lin et al. (2007) have proposed empirically determining the best values by simply measuring the TLB, defined as the ratio in the range [0,1] of the lower bound distance to the actual true Euclidean distance. The higher the ratio, the tighter is the bound. Since we aimed to achieve the tightest possible lower bounds, we can simply estimate the lower bounds over all possible parameters and select the best settings. To identify the TLB, we used Equation (8):

$$\mathrm{TLB} = \frac{\textit{lower bounds distance}}{\textit{true euclidean distance}} \tag{8}$$

The lower bound distance represents the distance after symbolization, whereas the true Euclidean distance represents the true distance of two time-series data. The value range of the TLB is always between 0 and 1; the higher the TLB value, the closer it is to the true Euclidean distance, which indicates better results. To test the required parameter and effectiveness of our revised distance measure, we have presented the evaluation results in Section 5.3.2.

5.3.2 TLB of the SAX algorithm

We evaluated the degree of distortion after dimension reduction by SAX algorithm to evaluate the accuracy of the method. We first evaluated the effect of *w*-dimensional space on

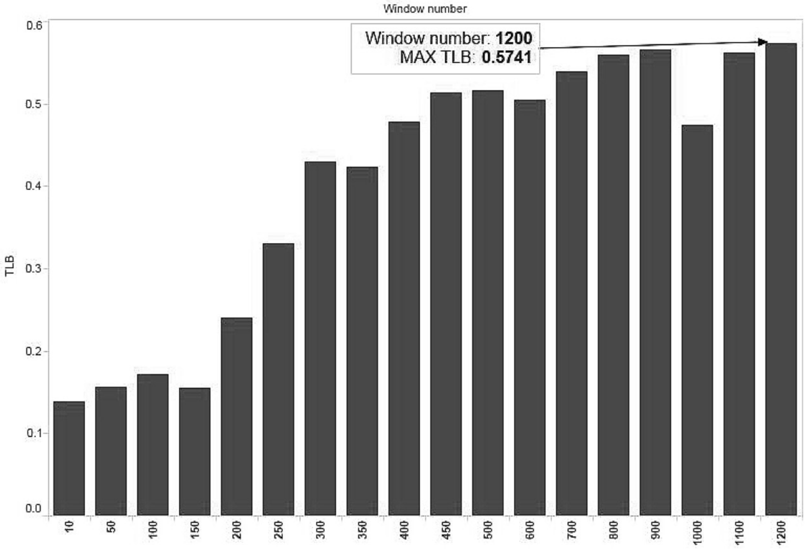

Figure 10. TLB Value after Dimension Reduction with the SAX Algorithm

TLB values. Figure 10 shows that the larger the *w* value, the higher the TLB value. When *w* was set to 300, it clearly improved the TLB value, and the slope of curve was gentler once we increased *w* from 300 to 1,200, meaning that the value of *w* could be greater than 250. Although a *w* with a larger value can yield a better TLB, it will also increase computation time. Thus, a tradeoff exists between the degree of distortion and computation time cost, which merits further investigation.

5.3.3 TLB of the modified SAX algorithm

We tested the *p*-value in Equation 5 to determine the best value of *p*. As mentioned, we set α to be 10 in our electricity consumption data in order to have alphabetical labels. As such, the least distance value between nine breakpoint values was 0.25. Accordingly, the value range of *p* was [0, 0.25). We adjusted the *p*-values to be 0, 0.05, 0.1, 0.15, and 0.2 to evaluate changes in TLB values (Figure 11). Interestingly, TLB values decreased when time window number was set to be 1,000 under each *p*-value in the case study. Based on the results of our tests, the *p*-value can be set to 0.20 for the abnormal and normal electricity patterns for future study.

6. Conclusions and Future Works

A symbolic time-series data mining and analytic framework for electricity consumption analysis in energy-intensive industries was proposed in this work. We deployed a data warehouse framework to analyze the load profiles of each attribute in order to select key attributes for further data mining tasks. Subsequently, we compared the results of two dimension reduction strategies with various data preprocessing methods to predict the state of the annealing process of target furnaces. We preliminarily

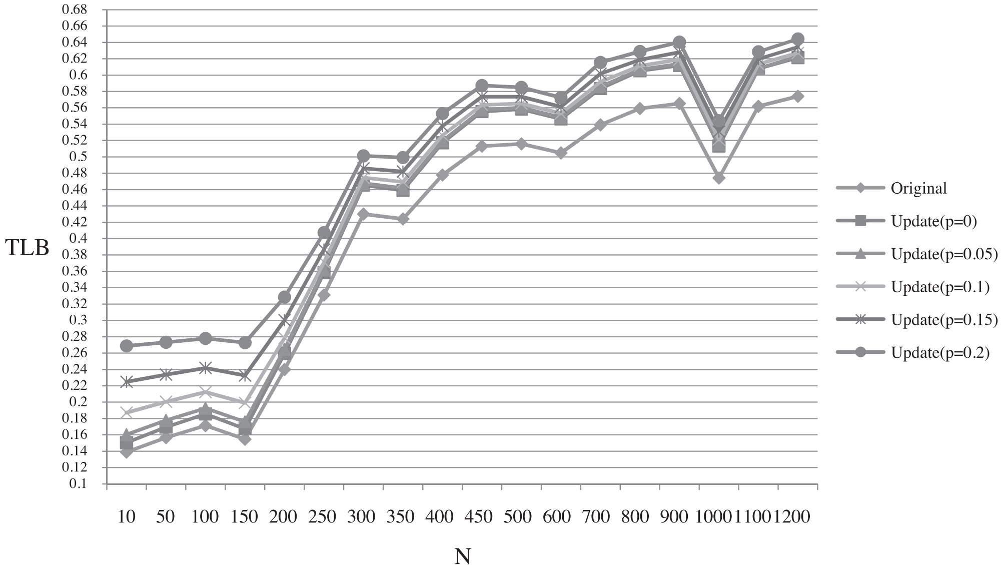

Figure 11. A Comparison Curves of TLB Values under Different *p* Values

confirmed that PAA with data outlier removal and data normalization processing (i.e., hybrid one) can achieve slightly better results than the FFP method. Based on results with PAA, we further experimented with the SAX algorithm to symbolize the electricity load profiling and to evaluate our adjusted distance measure by TLB values. We used a real-life case to demonstrate the application of the related methods. In the future, we will seek to decrease the RMSE of the prediction results by refining the method adopted in this work and apply the clustering approach to discriminate normal and abnormal electric patterns—that is, to group electric patterns for further analytical and prediction tasks. In addition, we will finalize all modules mentioned in the framework and conduct a series of experiments to comprehensively confirm the effectiveness of the proposed framework and approaches. The overarching goal of our research is to help the co-operating plant to make energy-optimization decisions in real time.

Acknowledgments

The support provided by Bureau of Energy and Industrial Technology Research Institute is gratefully acknowledged.

References

Aßfalg, J., Kriegel, H.-P., Kröger, P., Kunath, P., Pryakhin, A., & Renz, M. (2006). Similarity search on time series based on threshold queries. In Y. Ioannidis, M. H. Scholl, J. W. Schmidt, F. Matthes, M. Hatzopoulos, K. Boehm, ... C. Boehm (Eds.), *Lecture Notes in Computer Science: Vol. 3896. Advances in Database Technology-EDBT 2006* (pp. 276-294). Berlin, Germany: Springer-Verlag. doi: 10.1007/11687238_19

Berndt, D. J., & Clifford, J. (1994). Using dynamic time warping to find patterns in time series. *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, USA, 1994*, 359-370.

Chakrabarti, K., Keogh, E., Mehrotra, S., & Pazzani, M. (2002). Locally adaptive dimensionality reduction for indexing large time series databases. *ACM Transactions on Database Systems, 27*(2), 188-228. doi: 10.1145/568518.568520

Chan, K. A., & Fu, W. C. (1999). Efficient time series matching by wavelets. In M. Kitsuregawa, L. Maciaszek, M. Papazoglou, & C. Pu (Eds.), *Proceedings of the 15th International Conference on Data Engineering* (pp. 126-133). Los Alamitos, CA: IEEE Computer Society. doi: 10.1109/ICDE.1999.754915

Chen, L., & Ng, R. (2004). On the marriage of Lp-norms and edit distance. In M. A. Nascimento, M. T. Özsu, R. J. Miller, J. A. Blakeley, & K. B. Schiefer (Eds.), *Proceedings of the Thirtieth International Conference on Very Large Data Bases* (pp. 792-803). San Francisco, CA: Morgan Kaufmann. doi: 10.1016/B978-012088469-8/50070-X

Chen, L., Özsu, M. T., & Oria, V. (2005). Robust and efficient similarity search for moving object trajectories. In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data* (pp. 491-502). New York, NY: ACM.

Chen, Q., Chen, L., Lian, X., Liu, Y., & Yu, J. X. (2007). Indexable PLA for efficient similarity search. In C. Koch, J. Gehrke, M. N. Garofalakis, D. Srivastava, K. Aberer, A. Deshpande, ... E. J. Neuhold (Eds.), *Proceedings of the 33rd International Conference on Very Large Data Bases* (pp. 435-446). New York, NY: ACM.

Chen, Y., Nascimento, M. A., Ooi, B. C., & Tung, A. K. H. (2007). SpADe: On shape-based pattern detection in streaming time series. *Proceedings of the 23th International Conference on Data Engineering, Turkey, 2007*, 786-795. doi: 10.1109/ICDE.2007.367924

Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., & Keogh, E. (2008). Querying and mining of time series data: Experimental comparison of representations and distance measures. *Proceedings of the VLDB Endowment, 1*(2), 1542-1552. doi: 10.14778/1454159.1454226

Faloutsos, C., Ranganathan, M., & Manolopoulos, Y. (1994). Fast subsequence matching in time-series databases. In R. T. Snodgrass & M. Winslett (Eds.), *Proceedings of the 1994 ACM International Conference on Management of Data* (pp. 419-429). New York, NY: ACM. doi: 10.1145/191839.191925

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Waltham, MA: Morgan Kaufmann.

Keogh, E., & Pazzani, M. J. (2000). A simple dimensionality reduction technique for fast similarity search in large time series databases. In T. Terano, H. Liu, & A. L. P. Chen (Eds.), *Lecture Notes in Computer Science: Vol. 1805. Knowledge Discovery and Data Mining. Current Issues and New Applications* (pp. 122-133). Berlin, Germany: Springer-Verlag. doi: 10.1007/3-540-45571-X_14

Keogh, E., & Ratanamahatana, C. A. (2005). Exact indexing of dynamic time warping. *Knowledge and Information Systems, 7*(3), 358-386. doi: 10.1007/s10115-004-0154-9

Keogh, E., Chakrabarti, K., Pazzani, M., & Mehrotra, S. (2001). Dimensionality reduction for fast similarity search in large time series databases. *Knowledge and Information Systems, 3*(3), 263-286. doi: 10.1007/PL00011669

Kitagawa, G. (2010). *Introduction to time series modeling*. Boca Raton, FL: CRC Press. doi: 10.1201/9781584889229

Kleist, C. (2015). *Time series data mining methods: A review* (Unpublished master's thesis). Humboldt-Universität zu Berlin, Germany.

Le, C. V., Pang, C. K., Gan, O. P., Chee, X. M., Zhang, D. H., Luo, M., ... Lewis, F. L. (2012). Classification of energy consumption patterns for energy audit and machine scheduling in industrial manufacturing systems. *Transactions of the Institute of Measurement and Control, 35*(5), 583-592. doi: 10.1177/0142331212460883

Lin, J., Keogh, E., Lonardi, S., & Chiu, B. (2003). A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery* (pp. 2-11). New York, NY: ACM. doi: 10.1145/882085.882086

Lin, J., Keogh, E., Wei, L., & Lonardi, S. (2007). Experiencing SAX: A novel symbolic representation of time series. *Data Mining and Knowledge Discovery, 15*(2), 107-144. doi: 10.1007/s10618-007-0064-z

McLoughlin, F., Duffy, A., & Conlon, M. (2015). A clustering approach to domestic electricity load profile characterization using smart metering data. *Applied Energy, 141*, 190-199. doi: 10.1016/j.apenergy.2014.12.039

Ministry of Economic Affairs, Bureau of Energy. (2017). *Electricity consumption*. Retrieved from http://web3.moeaboe.gov.tw/ecw/populace/web_book/WebReports.aspx?book=M_CH&menu_id=142

Mörchen, F. (2006). *Time series knowledge mining* (Unpublished doctoral dissertation). Philipps-Universität Marburg, Germany.

Nath, H., & Baruah, U. (2014). Evaluation of lower bounding methods of dynamic time warping (DTW). *International Journal of Computer Applications, 94*(20), 12-17. doi: 10.5120/16550-6168

Popeangă, J. (2015). Data mining smart energy time series. *Database Systems Journal, 6*(1), 14-22.

Sakurai, Y., Yamamuro, M., & Faloutsos, C. (2015). Mining and forecasting of big time-series data. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data* (pp. 919-922). New York, NY: ACM. doi: 10.1145/2723372.2731081

Vikhorev, K., Greenough, R., & Brown, N. (2013). An advanced energy management framework to promote energy awareness. *Journal of Cleaner Production, 43*, 103-112. doi: 10.1016/j.jclepro.2012.12.012

Vlachos, M., Kollios, G., & Gunopulos, D. (2002). Discovering similar multidimensional trajectories. In R. Agrawal, K. Dittrich, & A. H. H. Ngu (Eds.), *Proceedings of IEEE the 18th International Conference on Data Engineering* (pp. 673-684). Los Alamitos, CA: IEEE Computer Society. doi: 10.1109/ICDE.2002.994784

Wall, M. E., Rechtsteiner, A., & Rocha, L. M. (2003). Singular value decomposition and principal component analysis. In D. P. Berrar, W. Dubitzky, & M. Granzow (Eds.), *A practical approach to microarray data analysis* (pp. 91-109). Norwell, MA: Kluwer. doi: 10.1007/0-306-47815-3_5

Yi, B. K., Jagadish, H. V., & Faloutsos, C. (1998). Efficient retrieval of similar time sequences under time warping. *Proceedings of the 14th International Conference on Data Engineering, USA, 1998*, 201-208. doi: 10.1109/ICDE.1998.655778

(Received: 2017/8/14; Accepted: 2017/10/10)

圖書資訊學刊　第15卷第2期　（2017.12）　頁21-44
doi: 10.6182/jlis.2017.15(2).021

基於符號化時間序列資料探勘架構之電力消耗負載分析

A Symbolic Time-series Data Mining Framework for Analyzing Load Profiles of Electricity Consumption

吳怡瑾[1]　陳子立[2]　洪冠群[3]　陳彥銘[4]　劉子吉[5]

I-Chin Wu[1], Tzu-Li Chen[2], Guan-Qun Hong[3],
Yen-Ming Chen[4], Tzu-Chi Liu[5]

摘　要

能源在永續發展工業已被視為重要的管理資產，因此，如何減少能源消耗並有效率地追蹤及管理能源為重要的挑戰。本研究基於電力負載追蹤電力消耗狀況提出符號化時間序列電力資料探勘架構，首先，研究應用分段聚合近似法（piecewise aggregate approximation, PAA）進行時間序列降維處理，接著採用符號聚合近似演算法（symbolic aggregate approximation, SAX）將降維後序列進行符號化，並改良SAX演算法的時間序列下限制（lower-bounding）距離衡量計算公式。研究以鋼鐵鍛造公司的大型退火爐為例進行方法驗證，實驗結果顯示採用PAA法較傳統的固定端點取法較能預測機器狀況；另一實驗結果顯示改良SAX之下限制距離公式能更準確地計算負載曲線之間的相似度。本研究所提出之架構與方法將有助於工廠進行後續正異常電力樣式預測。

關鍵字：電力負載曲線、分段聚合近似法、聚合近似演算法、時間序列資料探勘

[1] 國立臺灣師範大學圖書資訊學研究所
Graduate Institute of Library and Information Studies, National Taiwan Normal University, Taipei, Taiwan

[2,3] 輔仁大學資訊管理學系
Department of Information Management, Fu-Jen Catholic University, New Taipei, Taiwan

[4,5] 工業技術研究院綠能與環境研究所
Industrial Technology Research Institute, Hsinchu, Taiwan

* 通訊作者Corresponding Author: 吳怡瑾I-Chin Wu, E-mail: icwu@ntnu.edu.tw

註：本中文摘要由作者提供。

以APA格式引用本文：Wu, I.-C., Chen, T.-L., Hong, G.-Q., Chen, Y.-M., & Liu, T.-C. (2017). A symbolic time-series data mining framework for analyzing load profiles of electricity consumption. *Journal of Library and Information Studies, 15*(2), 21-44. doi: 10.6182/jlis.2017.15(2).021

以Chicago格式引用本文：I-Chin Wu, Tzu-Li Chen, Guan-Qun Hong, Yen-Ming Chen, and Tzu-Chi Liu. “A symbolic time-series data mining framework for analyzing load profiles of electricity consumption.” *Journal of Library and Information Studies* 15, no. 2 (2017): 21-44. doi: 10.6182/jlis.2017.15(2).021