Analyzing Load Profiles of Electricity Consumption by a Time Series Data Mining Framework

I-Chin Wu[1](✉), Tzu-Li Chen[2], Yen-Ming Chen[3], Tzu-Chi Liu[3], and Yi-An Chen[2]

[1] Graduate Institute of Library and Information Studies,
National Taiwan Normal University, Taipei, Taiwan
icwu@ntnu.edu.tw

[2] Department of Information Management, Fu-Jen Catholic University, New Taipei City, Taiwan

[3] Industrial Technology Research Institute, Hsinchu, Taiwan

Abstract. 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. In turn, reducing energy consumption for sustainability and both tracking and managing energy efficiently have become critical challenges. In response, we analyzed electricity consumption from the perspective of load profiling, which charts variation in electrical load during a specified period in order to track energy consumption. As a result, we proposed a time series data mining and analytic framework for electricity consumption analysis and pattern extraction by streaming data mining and machine learning techniques. We identified key factors to predict the state of the annealing furnace and detect abnormal patterns of the load profile of their electricity consumption. Our experimental results show that the dimension reduction method known as piecewise aggregate approximation can help to detect the state of the annealing furnace.

Keywords: Energy consumption analysis · Load profiling · Piecewise aggregate approximation · Time-series data mining

1 Introduction

As a cornerstone of modern civilization and economic growth, electricity is critical for industrial and economic advancement, as well as a driving force for sustainable development. 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 malfunctions and thereby reduce the efficiency of both production and electricity. 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.

F.F.-H. Nah and C.-H. Tan (Eds.): HCIBGO 2017, PART II, LNCS 10294, pp. 443–454, 2017.
DOI: 10.1007/978-3-319-58484-3_35

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

Source: Taiwan Power Company (http://www.taipower.com.tw/)

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.

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 and 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, the kernel of which are 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 threefold:

1. 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.
2. To identify three states of the annealing process—heating-up, temperature retention, and cooling down—based on the temperature information of the operating machine and, following Keogh et al. [1], use piecewise aggregate approximation (PAA) to perform dimension reduction for data representation and, detect machine operational states according to energy load profiles that can inform real-time energy-optimization decisions; and
3. To propose and construct an electric energy monitor and analysis framework based on load profiles by data stream mining and machine learning techniques as a means to implement the proposed time series data-mining approach in co-operating iron and steel manufacture.

The overarching goal of the three objectives is to deploy a visualized decision-making support system and propose actionable energy-saving strategies for co-operating plants to solve real-world problems.

2 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 [2]. Time series data are large, as well as numerical and continuous in nature, which require continuous updating. Mörchen [3] 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) [4] classified time series analysis into four categories: description, modeling, prediction, and signal extraction. Sakurai et al. [5] 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 study, we focused on the first. Popeangă [6] 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.

Since time series are high-dimensional data, they are time consuming for computing and 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 [7], single value decomposition [8], discrete wavelet transformation [9], PAA [1], SAX [10], and indexable piecewise linear approximation [11]. We will adopt the intuitive method of PAA and discretized the PAA representation of a time series into a symbolic representation method SAX algorithm.

3 Time-Series Electricity Consumption Data Mining Framework

We collected energy consumption data and the corresponding product information of two annealing furnaces in 2014. Figure 1 shows the proposed 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 model to predict the operational states of machines (i.e., warm-up, heat retention, and cooling)?

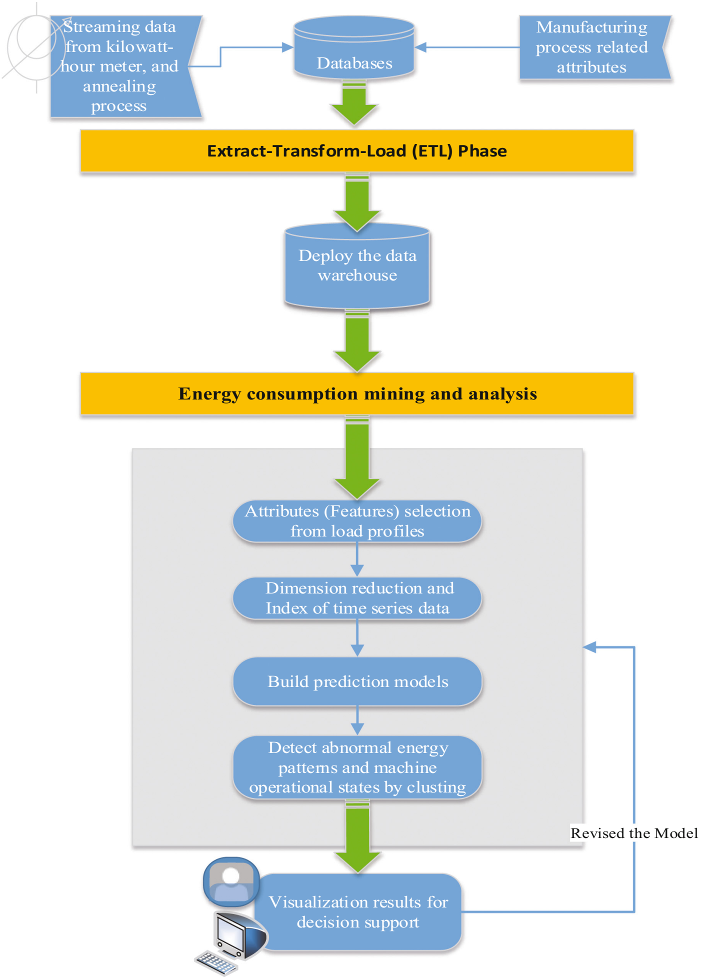

Streaming data from kilowatt-hour meter, and annealing process
Databases
Manufacturing process related attributes
Extract-Transform-Load (ETL) Phase
Deploy the data warehouse
Energy consumption mining and analysis
Attributes (Features) selection from load profiles
Dimension reduction and Index of time series data
Build prediction models
Detect abnormal energy patterns and machine operational states by clusting
Revised the Model
Visualization results for decision support

Fig. 1. Time-series data mining framework for electricity consumption analysis in industry

All tasks of analysis involved using the load profiles of the electricity consumption of the targeted machine. For the proposed framework, we preliminarily deploy 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 select and confirm key attributes to identify the state of the annealing furnace based on the results of analysis and consulted with domain experts. We confirm 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 min on average. We use the temperature information of the operating machine to identify three states: warm-up, heat retention, and cooling.

We apply the PAA method to discretize streaming data into n segments with timestamps in order to build the prediction model. We will refine the SAX algorithm, which is a symbolic representation of time series for dimensionality reduction and indexing with a lower-bounding distance measure to further extract subsequent patterns. It can help the system to detect abnormal energy patterns and machine operational states by symbolizing energy load profiles to make further energy-optimization decisions in real time. We will apply an agglomerative hierarchical clustering approach to discriminate normal and abnormal electric patterns—that is, to group the electric patterns for further analytical and prediction tasks. We plan to next conduct 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, and abnormal energy patterns. We also included associated experiments of parameter selection of the PAA method 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 co-operating iron and steel plants 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 3 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 (Fig. 2) and analyzed the relationships among various attributes (e.g., electric power, temperature, and product weight. Figure 3 presents the fact table of our research. The data warehousing platform had two chief goals: to analyze the load profiles of each annealing process and to define annealing states based on the selected attributes of load profiles.

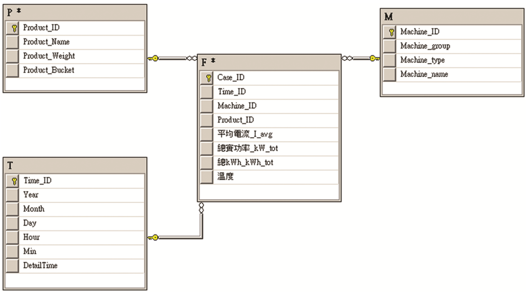


Fig. 2. Star schema of the data mart for EC Analysis

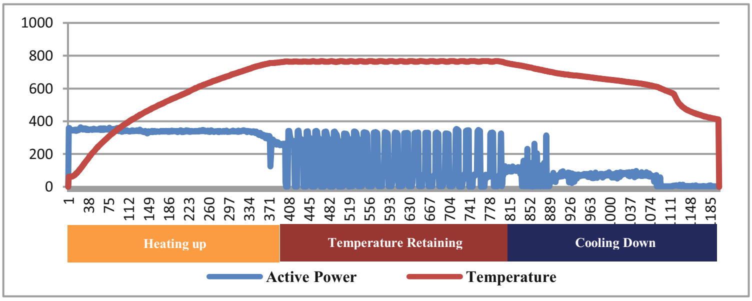


Fig. 3. Load profiles of active power and temperature

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. We used the data of annealing process from April 1, 2014, to December 31, 2014 to train and construct the prediction model to detect each machine's state and condition.

After selecting the attributes that were useful for periodical data analysis, we adopted the star schema to build the data mart (Fig. 2). 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 identify three states: warm-up, heat retention, and cooling. By extension, we could further identify the normal or abnormal states of each annealing process. We show one load profile of active power and temperature of one annealing furnace in Fig. 3 (Table 2).

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

4.2 Time Series Representation for Constructing the Prediction Model

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 [1, 12].

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, in a method we dub the fixed feature point (FFP) method. Figure 4 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 (Fig. 5). 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 [13]. Put differently, the transformation parameters are preset without consideration of the

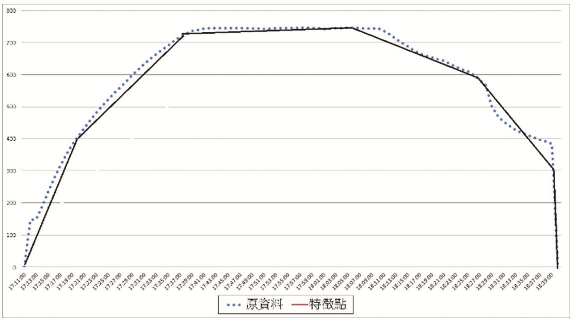


Fig. 4. FFP method for feature point extraction (temperature)

underlying data. We further adopted SAX after PAA to represent each feature point of the load profile symbolically. Due to the constraints of space, we report the results of the FFP method versus PAA for identifying states of the annealing process.

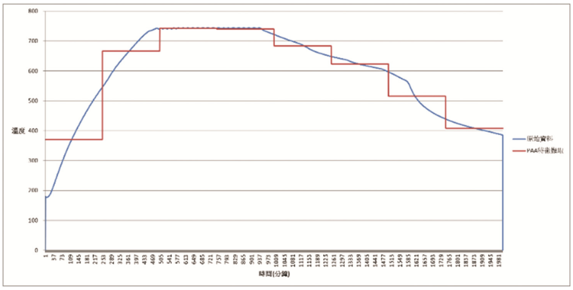


Fig. 5. PAA method for feature point extraction (temperature)

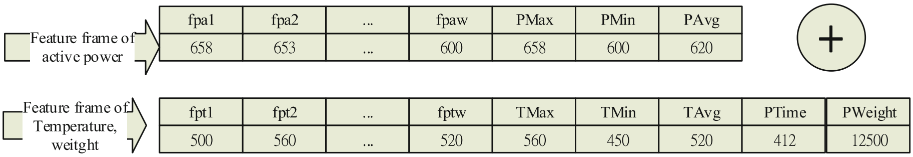
Feature frames of each annealing process. Based on the methods, we defined time series data and related notations (Table 3). We denoted time series data of an attribute i as S = (s1, s2,….sn), with the length of a time series in n and w as the dimensionality of the space to index the time series data. 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 $\overline{S} = (\overline{s_1}, \overline{s_2}, \ldots, \overline{s_w})$ – that is, w-dimensional space by vector $\overline{S}$. The ith feature point of $\overline{S}$ can be derived from Eq. (1).

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

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 = (fpa1, fpa2, …fpa$_w$)
FP_T (feature point of temperature)	A time series of temperature of length *w* after dimension reduction, FP_T = (fpt1, fpt2, …fpt$_w$)

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 (3) the weight of raw material information. The feature frame in Fig. 6 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 shown in Fig. 2.


Feature frame of active power
Feature frame of Temperature, weitght
+

fpa1	fpa2	...	fpaw	PMax	PMin	PAvg
658	653	...	600	658	600	620

fpt1	fpt2	...	fptw	TMax	TMin	TAvg	PTime	PWeight
500	560	...	520	560	450	520	412	12500

Fig. 6. An example of a feature frame as an input string for the prediction model

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

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. Notably, we discretized the streaming data into n segments with timestamps to construct the model. Based on our preliminary analytical results, we confirmed that either the electric power or temperature information of the operating machine can help to identify a machine's entire operational process, which is 1,440 min on average. We then used the temperature information of the operating machine to identify three states: warm-up, heat retention, and cooling. We collected energy consumption data and corresponding product information of two annealing furnaces from April 1, 2014, to December 31, 2014. Herein, we present the experimental results for one furnace. We explain the results of the two primary sets of experiments with the FFP method and PAA as feature extraction methods in what follows.

- Experiment 1 (FFP): The set of experiments included original stream data, data with the normalization process, data with the extreme value removal process, data with normalization, and the extreme value removal process—respectively, baseline_FFP, standardization _FFP, extreme_FFP, and hybrid_FFP.
- Experimental 2 (PAA): The set of experiments included original stream data, data with the normalization process, data with the extreme value removal process, data with normalization, and the extreme value removal process—respectively, baseline_PAA, standardization _PAA, extreme_PAA, and hybrid_PAA.

The purpose of data standardization with z-score standardization is to remove outlier data points and elucidate the relationship between a data point and the average value of all data points. The z-score converts all indicators to a common scale with an average of 0 and standard deviation of 1. The equation of the z-score method used appears in Eq. (2):

$$Normalized\left(e_i\right) = \frac{e_i - \overline{E}}{std(E)} \tag{2}$$

in which e_i represents the data points of the load profile, *std(E)* is the standard deviation of the data points of the load profile, and $\overline{E}$ is the mean value of the data points.

The purpose of removing outlier values is to avoid excessive noise in the time series data. We removed feature points outside twice the standard deviation of the average value, $\overline{E}$, of the target load profile. Ultimately, the hybrid method involved removing outlier data points and adopting the *z*-score.

We adopted sequential minimal optimization, in which a multilayer perceptron (MLP) is a feedforward artificial neural network model, and 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. The RMSE is the mean of the square of all errors, which is used to measure the differences between values.

5.2 Experimental Results for Identifying Operational States

Tables 5 and 6 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 and PAA approaches. Note that when we set *w* to 50, for example, we extracted 50 feature points to represent the entire load profile of the active power.

Observation 1 (FFP). For the FFP approach, the worst method on average is *standardization_FPP*. However, the *hybrid_FFP* can achieve the minimum RMSE in comparison to the other three methods under various *w* value settings. By contrast, *hybrid_FFP* and *extreme_FPP* have similar results under various *w* values, which indicates that we can help to remove the extreme value and then perform standardization. Overall, the best results on average occurred when *w* was 100. It seems that a larger *w* value (i.e., more feature points) with the FPP method does not generate better results in predicting states of machines.

Observation 2 (PAA). Like the FFP approach, the worst method for PAA is *standardization _PAA*. *extreme_PAA* and *hybrid_PAA* can achieve the minimum RMSE in comparison to the other two methods under various *w* values. Overall, the best results on average were with *w* at 150. The FFP method seems insensitive to *p*-values; however, more or fewer feature points does not yield better results in predicting states of machines.

Observation 3 (Comparison). Both dimension reduction approaches generated similar results between the methods. For example, after conducting data standardization without removing extreme values generated the worst results. When we compared the method between the approaches, we observed that the FFP method is worse than PAA, because the former is more sensitive to extract points in representing subsequent parts of the data stream. As such, we adopted PAA to further symbolize processing by the SAX algorithm and set *w* to 150 (i.e., 150 feature points to represent the entire data stream).

Table 5. Prediction the operational state by FFP method in terms of RMSE

Method/w	50	100	150	200	250	300	350	Average
baseline_FFP	0.070	0.086	0.108	0.085	0.086	**0.083**	0.100	0.088
standardization _FFP	0.340	**0.314**	0.497	0.337	0.344	0.365	0.422	0.374
extreme_FFP	0.082	0.081	0.072	**0.069**	0.0703	0.078	0.072	0.075
hybrid_FFP	0.065	**0.058**	0.062	0.059	0.065	0.060	0.064	0.062
Average	0.139	***0.135***	0.185	0.137	0.142	0.146	0.165	0.150

Table 6. Prediction the operational state by PAA method in terms of RMSE

Method/w	50	100	150	200	250	300	350	Average
baseline_PAA	0.121	0.136	0.126	0.150	0.211	**0.115**	0.129	0.141
standardization _PAA	0.397	0.218	**0.090**	0.109	0.164	0.104	0.126	0.173
extreme_PAA	0.115	0.097	**0.091**	0.099	0.091	0.125	0.100	0.103
hybrid_PAA	**0.083**	0.123	0.127	0.119	0.117	0.107	0.172	0.121
Average	0.179	0.143	***0.109***	**0.119**	0.146	**0.113**	**0.132**	0.134

6 Conclusions and Future Works

We proposed a time series data mining and analytic framework for electricity consumption analysis in energy-intensive industries. 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. We then compared the results of two dimension reduction approaches with various data preprocessing methods to predict the state of the annealing process of target furnaces. We preliminarily confirmed that PAA with data outlier removal and data standardization processing can achieve slightly better results than the FFP approach. In the future, we will finalize all modules mentioned in the framework and conduct a series of experiments to confirm the effectiveness of the proposed framework and approaches to identify electricity patterns and machine operational states in real time.

Acknowledgments. This research was supported by Ministry of Science and Technology of Taiwan under Grant MOST 105-2410-H-003-153-MY3 & Bureau of Energy of Taiwan is gratefully acknowledged.

References

1. Keogh, E., Chakrabarti, K., Pazzani, M., Mehrotra, S.: Dimensionality reduction for fast similarity search in large time series databases. Knowl. Inf. Syst. **3**(3), 263–286 (2001)
2. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques, 3rd edn. Morgan Kaufmann, San Francisco (2011)
3. Mörchen, F.: Time series knowledge mining. Ph.D. thesis, Philipps-University Marburg, Germany (2006)
4. Kitagawa, G.: Introduction to Time Series Modeling. Monographs on Statistics & Applied Probability. Chapman & Hall/CRC, Boca Raton (2010)
5. Sakurai, Y., Matsubara, Y., Faloutsos, C.: Mining and forecasting of big time-series data. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Tutorial, pp. 919–922, Melbourne, Victoria, Australia (2015)
6. Popeanga, J.: Data mining smart energy time series. Database Syst. J. **6**(1), 14–22 (2015)
7. Faloutsos, C., Ranganathan, M., Manolopoulos, Y.: Fast subsequence matching in time-Series databases. In: Proceedings of the 1994 ACM International Conference on Management of Data (SIGMOD), pp. 419–429 (1994)
8. Wall, M.E., Rechtsteiner, A., Rocha, L.M.: Singular value decomposition and principal component analysis. In: Berrar, D.P., Dubitzky, W., Granzow, M. (eds.) A Practical Approach to Microarray Data Analysis, pp. 91–109. Kluwer, Norwell (2003)
9. Chan, K.A., Fu, W.C.: Efficient time series matching by wavelets. In: Proceeding of the 15th International Conference on Data Engineering (ICDE), pp. 126–133 (1999)
10. Lin, J., Keogh, E.J., Wei, L., Lonardi, S.: Experiencing SAX: a novel symbolic representation of time series. Data Mining Knowl. Discov. **15**(2), 107–144 (2007)
11. Chen, Q., Chen, L, Lian, X., Liu, Y., Yu, J.X.: Indexable PLA for efficient similarity search. In: Proceeding of the 33rd International Conference on Very Large Data BaseS (VLDB), pp. 435–446 (2007)
12. Keogh, E.J., Pazzani, M.J.: A simple dimensionality reduction technique for fast similarity search in large time series databases. In: Terano, T., Liu, H., Chen, A.L.P. (eds.) PAKDD 2000. LNCS (LNAI), vol. 1805, pp. 122–133. Springer, Heidelberg (2000). doi:10.1007/3-540-45571-X_14
13. Kleist, C.: Time series data mining methods: a review, (Master's thesis, Humboldt-Universit¨at zu Berlin, Germany). http://edoc.hu-berlin.de/master/kleist-caroline-2015-03-25/PDF/kleist.pdf