A Clustering Model for Mining Consumption Patterns from Imprecise Electric Load Time Series Data

Qiudan Li, Stephen Shaoyi Liao, and Dandan Li

Laboratory of Complex Systems and Intelligence Science,
Institute of Automation, Chinese Academy of Sciences, Beijing
qiudan.li@ia.ac.cn, qiudanli@gmail.com
Department of Information System, City University of Hong Kong,
School of Economics and Management,
South West Jiao Tong University, China
issliao@cityu.edu.hk
Department of Automation and Computer-Aided Engineering,
The Chinese University of Hong Kong
ddli@acae.cuhk.edu.hk

Abstract. This paper presents a novel clustering model for mining patterns from imprecise electric load time series. The model consists of three components. First, it contains a process that deals with representation and preprocessing of imprecise load time series. Second, it adopts a similarity metric that uses interval semantic separation (Interval SS)-based measurement. Third, it applies the similarity metric together with the k-means clustering method to construct clusters. The model gives a unified way to solve imprecise time series clustering problem and it is applied in a real world application, to find similar consumption patterns in the electricity industry. Experimental results have demonstrated the applicability and correctness of the proposed model.

1 Introduction

With the rapid development of electricity information system, a big amount of electric load time series data are collected, which contain consumption pattern information of electricity consumers. Finding similar consumption pattern groups from these data can help the operators know more about regularities of consumption and make correct decisions. Time series clustering mining can serve for this purpose [1], [2]. However, due to collection or prediction reasons, the load time series data often involves imprecision. Therefore, the clustering mining model has to be designed to deal with this imperfection. Imprecision can be interval valued or fuzzy valued [3].

The purpose of this paper is thus to propose a load time series clustering mining model that can handle the imprecision of load time series efficiently. In [4], the authors discuss the sequence matching problem from a new angle focusing on the imprecise time series data represented by interval values. The method provides an efficient way for mining patterns from imprecise electric load time series.

L. Wang et al. (Eds.): FSKD 2006, LNAI 4223, pp. 1217 – 1220, 2006.

2 A Clustering Model

2.1 Representation and Preprocessing Process

The imprecise electric load time series is a finite sequence X of interval number: X = (x_1, x_2…x_n), where x_i is an interval number $[\underline{x}_i, \overline{x}_i]$ and n is the length of X. From the representation, we notice that X has two boundary sequences, one is lower sequence $\underline{X} = (\underline{x}_1, \underline{x}_2, ..., \underline{x}_n)$ and the other is upper sequence $\overline{X} = (\overline{x}_1, \overline{x}_2, ..., \overline{x}_n)$.

To guarantee the correctness when comparing two load time series, preprocessing procedure is used to handle the differences of baselines and scales [4]. Let $X^* = (x_1^*, x_2^*, \cdots, x_n^*)$ be any boundary sequence, where x_i^* $(1 \le i \le n)$ is any crisp number. In our study, normalization mappings including the Z-score normalization and the max-min linear normalization are used [5].

Definition1. Z-score normalization of X^* is defined as follows:

$$x_i^{*'} = \frac{x_i^* - \mu_{X^*}}{\delta_{X^*}} \tag{1}$$

where μ_{X^*} and δ_{X^*} are mean and standard deviation of the sequence X^*, respectively.

Definition2. Max-min linear normalization of X^* that transforms the values into scale [0,1] is defined as follows:

$$x_i^{*'} = \frac{x_i^* - \min X^*}{\max X^* - \min X^*} \tag{2}$$

where $\min X^*$ and $\max X^*$ are the minimum and maximum values of the sequence X^*, respectively.

2.2 Similarity Metric for Imprecise Electric Load Time Series

The similarity between two imprecise load time series is defined as follows [4]:
Definition3. Let $\delta > 0$ be an integer constant and φ is the normalization mapping. Two load time series *X and Y* have (γ,ε,φ,δ)-similarity if and only if, given $0 \le \varepsilon \le 1$, there exist $X_s = (x_{i_1}, ..., x_{i_l})$ and $Y_s = (y_{j_1}, ..., y_{j_l})$ that are subsequences in *X* and *Y*, respectively, and they satisfy the following conditions: 1)for any $1 \le k \le l$-1, $i_k < i_k$+1 and $j_k < j_k$+1; 2) for any $1 \le k \le l$, $|i_k - j_k| \le \delta$; 3) for any $1 \le k \le$ l, $SS([\varphi(\underline{x_{i_k}}), \varphi(\overline{x_{i_k}})], [\varphi(\underline{y_{i_k}}), \varphi(\overline{y_{i_k}})]) \le \varepsilon$. Interval SS is the similarity between two interval numbers.

Given ε and φ, the similarity degree between X and Y, Sim(X,Y), is represented as l/n, where l is the length of longest similar subsequences between X and Y.

2.3 Construction of the Cluster

K-means clustering algorithm is widely used in many real world applications [5]. We introduce the proposed similarity metric to k-means to handle imprecise time series clustering. The construction process consists of the following steps:

Step1: Randomly select k objects as initial cluster centers;
Step2: Assign each object to its most similar cluster center, the process can be regarded as a similarity matching process. The similarity degree between each object and cluster center is the length of longest similar subsequences between them, dynamic programming approach can server this purpose;
Step3: Recalculate the cluster center for each cluster;
Step4: The process will continue until the changing value of E in the last two iterations is less than a predefined threshold. E is total sum of dissimilarity degree for object to its cluster center, which can be defined as follows:

$$E = \sum_{i=1}^{k} \sum_{b \in C_i} (1 - Sim(b, m_i)) \tag{3}$$

where k is the number of clusters, b is an object, m_i is the center of cluster C_i, and $1 - Sim(b, m_i)$ describes the dissimilarity between b and m_i.

3 Use of the Proposed Model in an Electricity Industry

In this section, we apply the proposed model in an electricity industry to demonstrate its correctness and practicability.

We use about one year's electric load time series data of a real electricity system in this study. Each time series has 96 values, which are collected per 15 minutes interval in a day. Cluster patterns are constructed by the proposed model, during the cluster construction process, the total sum of dissimilarity degree decreases at each iteration as objects searching new similar cluster centers. Fig.1 shows the found three representative consumption patterns. Similarity parameters ε, δ in this study are 0.1 and 2, respectively.

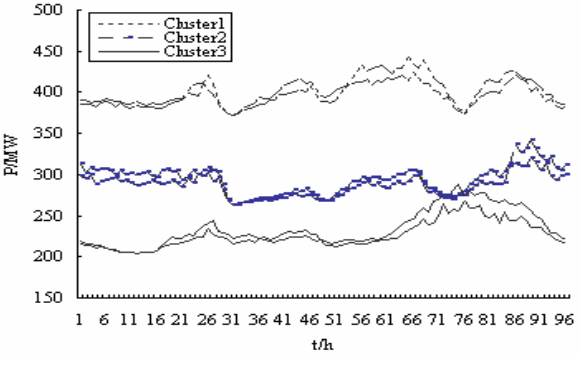

Fig. 1. Three representative cluster patterns

It can be seen from Fig.1 that: 1) Cluster3's load values are relatively low from time points 1 to 20, 85 to 96, while values are relatively high from time points 25 to 44, 65 to 84. By taking time series' date type into account, we can find that cluster3 is a typical holiday's consumption pattern, since some big factories rest and do not

consume electricity from time points 1 to 20, 85 to 96. 2) Cluster1 and cluster2's load values' trends at the above time points are opposite to that of cluster3, the reason is that the big factories work on those periods to lessen costs. This trend of cluster2 is more obvious than that of cluster1. Generally speaking, cluster1 and cluster2 can be regarded as normal days' consumption patterns. Since more use of air-condition will leads to higher electricity consumption amount in summer, and some factories will also rest on very hot days, therefore, load values from time points 1 to 20, 85 to 96 are slightly lower, however, the total amount of electricity consumption is still higher than that of other seasons. So cluster1 shows more characteristics of summer while cluster2 shows characteristics of other seasons.

Based on the above analysis, we can find that the proposed model still mines the reasonable consumption patterns from imprecise electric load time series. The found patterns can not only distinguish consumption patterns of holidays from those of normal days, but also show the consumption characteristics of different seasons. Based on the consumption clusters, operators can generate corresponding strategy for different consumer to make the enterprise be competitive.

4 Conclusions and Future Work

We address time series clustering problem from a new angle with a focus on clustering of imprecise time series that are represented by interval numbers. The model is also applicable in other applications to find clustering patterns from imprecise time series data.

Future work will involve the following aspects: First, to further validate the performance of the proposed model. Second, to mine other kinds useful consumption patterns from imprecise electric load time series. Finally, to integrate the model in an electric load time series decision support system.

References

1. Fátima Rodrigues, Jorge Duarte, Vera Figueiredo, Zita A. Vale, Manuel Cordeiro:A comparative analysis of clustering algorithms applied to load profiling. MLDM (2003)73-85
2. Keogh E. J., Kasetty S.: On the need for time series data mining benchmarks: A survey and empirical demonstration. Proc.8th ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining. (2002) 102-111
3. Parsons, S.: Current approaches to handling imperfect information in data and knowledge bases. IEEE Trans. Knowledge Data Eng. (1996)353-372
4. Liao, S. S., Tang, T. H., Liu W.Y.: Finding relevant sequences in time series containing crisp, interval, and fuzzy interval data. IEEE Tran. Syst. Man, Cybern. B (2004)2071-2079
5. Han J., Kamber, M.: Data Mining: Concepts and Techniques. San Mateo, CA: Morgan Kaufmann(2001)