Abstract

Time series data grouping is essential for empirical data analysis and data summarization. Generally, grouping of time series data is based on similarity measure (distance function), thus time series in the same group are similar. The choice of similarity measure is very important during data grouping. In this paper we investigate the issue of smart meter data grouping according to daily consumption pattern using either Euclidian or correlation-based similarity. We find that the correlation-based measure is superior for data grouping with respect to consumption pattern.

Keywords: time series, clustering, classification, distance function, smart meter.

1. Introduction

Grouping time series data is of interest in many areas of data analytics. In this paper, we investigate the issue of grouping smart meter data according to consumption patterns and propose a statistical method for grouping. A smart meter is an electronic device that records energy consumption in intervals of an hour or less and communicates with an energy management system at regular time intervals to transmit meter readings. The readings for a day may be seen as a time series with for instance hourly readings. For analytic purposes, it is important to group smart meter time series correctly according to consumption patterns for different household types. Household types could be “working family with children”, “a youngster living alone”, “retired couple” etc. Different household types are expected to have different consumption patterns. Figure 1 illustrates different consumption patterns. In our previous work ([1]), we have suggested a novel method for generating synthetic smart meter data based on consumption patterns in real world data (seed). We want to enhance this method, so data for different household types can be generated. In order to make the generated data reflect daily consumption patterns for household types correctly; we need to be able to group our seed correctly according to daily consumption patterns for different household types. In data mining, there are two major approaches to group data: clustering and classification ([2]). Clustering is a process, where data is divided into a number of “natural groups” called clusters. Normally, the number of clusters is given in advance, but the features that characterise the clusters are not.
Classification, on the other hand, is done with outset in a number of pre-defined classes, and data is assigned to one or more of the classes according to closeness in features.

In Figure 1 above, \( P \) represents the pattern for a typical working family with children: Consumption peaks in the morning hours, is rather low during the day, but begins to increase in the afternoon, when the children come home from school and peaks again in the late afternoon and the evening. \( R1 \) represents the readings for a family with a consumption that is lower than \( P \), but close to the pattern. \( R5 \) represents a family with similar pattern, but much larger total consumption (more children, perhaps). \( R2 \) represents a household with a completely different pattern (for instance, a youngster living alone, gaming in the night and sleeping in the daytime).

Several methods for clustering and classification of time series data have been proposed and tested in many different domains, and there are many studies on similarity of time series, for instance [3], [4] and [2] give surveys, [5] considers kernel spectral clustering of smart meter data for forecasting purposes, and [6] and [7] report experimental tests of different approaches to clustering and classification of time series. Studies also (see for instance [8]) suggest that the choice of distance function or similarity measure is very important for obtaining good clusters. Most work on clustering and classification of smart meter time series has been done in order to do load forecasting of the consumption (for a single household or groups of households with some similarities), see for instance [9], [10], [11], [12], [5], [13] and [14]. In these works not only consumption pattern, but also absolute consumption is taking into account in the proposed grouping methods.

The purpose of this paper is to discuss different approaches to do this grouping according to consumption patterns and to choose an appropriate similarity
measure. Furthermore, we propose a simple method based on correlation that only considers consumption patterns. Firstly, we discuss how to prepare the raw meter readings for grouping (that is to create a profile ([15]) for the household), and secondly, we discuss different distance functions that may be used for grouping according to consumption patterns. Finally, we briefly discuss different approaches to grouping based on the chosen distance function.

2. Preparing the input (calculating profiles)

Household consumption profiles may be different for week days, when the family may be away at work or at school, and weekends and holidays. So it may be preferable to operate with two different profiles per household type, but in these preliminary investigations, we will disregard this. Also, we don’t take season and weather data into account at this stage. Firstly, because it is not relevant for investigating distance measures, and secondly, because our analytic platform first will be offered to Danish consumers and companies, and in Denmark electricity rarely is used for heating and cooling. For later versions, season and weather data should be taken into account, when profiles for households are calculated.

For calculating the consumption profiles for a set of households, we assume that we have at least one year of hourly meter readings for a set of meters. Given N meters with hourly readings for a year per meter, let RM be a set of raw meter readings:

\[ RM = \{RM_1, RM_2, \ldots, RM_N\} \]

where

\[ RM_i = \left\{ \left[ \sum_{d=1}^{365} r_{d,h} \right] \right\}_{d=0}^{23} \]

For every meter we want to generate a yearly average for that meter as basis for the clustering. We will call this the profile ([15]) for meter i.

Let M be a set of yearly average meter readings (profiles):

\[ M = \{M_1, M_2, \ldots, M_N\} \]

Each reading \( M_i \) \((1 \leq i \leq N)\) is a time series of 24 hourly readings: \( M_i = \{r_{i,0}, r_{i,1}, \ldots, r_{i,23}\}\). This algorithm outlines the computation of M:

Input: \( RM \)
Output: \( M \)

\[
\begin{align*}
\text{foreach}(\text{Meter } M_i, 1 \leq i \leq N) \\
\quad \text{foreach}(\text{hour } h, 0 \leq h < 24) \\
\quad \quad \text{- compute } s = \sum_{d=1}^{365} r_{d,h} \\
\quad \quad \text{- set } r_{ih} = s / 365 \\
\quad \quad \text{- add } r_{ih} \text{ to } M_i
\end{align*}
\]
3. Distance functions

Let $P_k$ be a time series of readings representing the consumption pattern for a day (hourly based readings) for a household of type $k$ ($1 \leq k \leq K$, $K$ being the number of wanted clusters (household types)):

$$P_k = \{p_{k,0}, p_{k,1}, \ldots, p_{k,23}\}$$

We want to group meters according to their similarities with some given pattern. Most clustering and classification techniques require a similarity measure, metric or distance function. Hence, we need to decide on a distance function that computes how close a meter profile is to a pattern (that is, a function: $\text{dist}(P_k, M)$) that returns a measure of how close the average readings $M_i$ (profile) is to the pattern $P_k$.

Further, we need to decide how to create the patterns and how to measure the distance between a series of readings and a pattern – that is, define $\text{dist}(P_k, M_i)$. Firstly, we will look at how to define the distance function.

Several different approaches to similarity measures are proposed. In general Euclidian distance seems to be most popular. For clustering and classifying time series data (and especially smart meter time series for forecasting purposes) Dynamic Time Warping (DTW), Dynamic Wavelet Transformation (DWT) and other approaches has been proposed and evaluated in many works ([3], [4], [6], [2], [7], [14], [16] and [17]). DTW and DWT measure similarity in pattern and absolute consumption with differences in time scaling. This is not needed for our purpose, so we will consider a simpler approach. A few papers propose to use Pearson correlation ([18]) as a good measure for similarity in patterns ([9] and [19]) and in [20] correlation is used in voice recognition (looking for patterns independent of loudness). So in the following we will consider Euclidian distance and distance based on correlation.

Euclidean distance is defined as:

$$\text{euclDist}(P_k, M_i) = \sqrt{\sum_{i=0}^{23} (p_{k,i} - r_i)^2}$$

It measures average absolute distance between the two time series.

Pearson correlation is defined as:

$$\text{cor}(P_k, M_i) = \frac{\text{Cov}(P_k, M_i)}{\sigma(P_k)\sigma(M_i)}$$

where $\sigma$ denotes standard deviations and the covariance, $\text{Cov}(P_k, M_i)$ is defined as:

$$\text{Cov}(P_k, M_i) = E[(p_{k,h} - \mu_{P_k})(r_h - \mu_{M_i})]$$
where $\mu$ denotes the averages (means). The correlation is expressed as follows:

$$
cor(P_k, M_i) = \frac{\sum_{h=0}^{23} (p_{kh} - \mu_{P_k})(r_{ih} - \mu_{M_i})}{\sqrt{\sum_{h=0}^{23} (p_{kh} - \mu_{P_k})^2} \cdot \sqrt{\sum_{h=0}^{23} (r_{ih} - \mu_{M_i})^2}}
$$

From a theoretical viewpoint, correlation is preferable, since it measures relative variation over time, while Euclidean distance measures absolute distance. We want to group according to daily patterns in consumption, not absolute size. For instance, two households of working families with children may have similar patterns, but if one household has more children, then their absolute consumption may be larger. This leads us to define a distance function based on correlation as:

$$
corDist(P_k, M_i) = 1 - cor(P_k, M_i)
$$

This distance function will give a distance of zero for perfectly correlated (correlation = 1) time series and a distance of 2 (correlation = -1) will indicate a complete opposite pattern.

These distance functions are illustrated by the following experiments: In Figure 2a, $R2$ is a household with a completely different pattern, $R3$ and $R4$ are permutations of $R1$ (from figure 1) and $P$ respectively and are included in order to have readings with the same total, mean and variance as $P$ and $R1$ respectively. In Figure 2b, we have the households from Figure 1 representing households with similarity in consumption pattern, but different absolute consumption.

The results of computing the Euclidean distance, the correlation and the correlation based distance to the profile are shown in Table 1. From the results in Table 1, it is seen that $R1$, $R3$ and $R4$ have Euclidean distances to $P$, which are rather close to each other, and $R5$ has a much larger Euclidean distance, although $R1$ and $R5$
have the same pattern as $P$ (Figure 2b) while $R3$ and $R4$ have completely different patterns (Figure 2a). On the other hand, the correlation-based distances to $P$ are small for $R1$ and $R5$ and relatively large for $R2$, $R3$ and $R4$, which is what we want.

<table>
<thead>
<tr>
<th></th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
<th>R5</th>
</tr>
</thead>
<tbody>
<tr>
<td>euclDist($P, R_i$)</td>
<td>4,3275</td>
<td>7,8338</td>
<td>5,7640</td>
<td>5,5209</td>
<td>12,7697</td>
</tr>
<tr>
<td>corDist($P, R_i$)</td>
<td>0,1012</td>
<td>1,3764</td>
<td>0,6082</td>
<td>0,5768</td>
<td>0,1012</td>
</tr>
</tbody>
</table>

Table 1. Comparison of distance functions.

From this it seems that the correlation-based distance is most suitable for our purpose, but more systematic tests and experiments should be conducted. Criteria for evaluating metrics can be found in [3].

4. Clustering methods

After choosing an appropriate distance function, a clustering or classification algorithm must be chosen. Many approaches to clustering and classification are proposed ([3] and [2] provides comprehensive overviews). For the purpose of grouping time series according to typical consumption patterns for household types, we will consider three approaches. The first approach is based on clustering, and the second and third approaches are based on classification. In all three approaches, we will predefine a set of typical consumption patterns representing household interest types.

The first approach uses standard $K$-Means clustering ([21]). In this approach the patterns for the chosen household types may be used as initial centroids for the clustering process. As the clustering process moves on, the centroids will change as they are adjusted each time a new series of readings is added to the cluster. In the second approach, classification may be done using the $k$-Nearest Neighbours technique ([21]) with $k = 1$ and using the chosen patterns for household types as centroids. And finally, in the third approach, the chosen patterns for household types may be used to define groups. A series of readings will be added to a group; if it is sufficient close to the consumption pattern that represents the household type. For each meter, the distances between the meter's profile and the defined patterns are to be computed; and the meter is assigned to the closest group. If the distance is greater than some threshold, we will not assign that meter to any group. This may leave us with readings that are not in any group. In the last two approaches, the centroids of the groups do not change.

5. Discussion and future work

In this paper we have discussed the problem of grouping smart meter readings according to daily consumption patterns for household types (“working family with children”, “younger living alone”, “working family with no children”, “retired couple”
Firstly, we propose to compute consumption profiles for each household in the raw data as the annual average consumption per hour. These profiles may have to be computed as one profile for weekdays and one for weekends and holidays. Also in other geographical areas than Denmark, weather and climate data may be taken into account. Secondly, we investigated different similarity measures for smart meter time series, and concluded that for our needs the correlation based distance (10) seems most feasible. Often Euclidian distance (6) is used, but as it is seen from table 1, it may be large for time series with similar patterns, but different absolute consumption, while the correlation based distance is small. Hence, we will investigate the use of the correlation based distance function further. Finally, we briefly discussed different approaches to grouping the time series data, either using K-Means clustering or classification based on pre-defined consumption patterns for household types. At the moment it is not clear which approach is best. K-Means clustering will try to group the data according to “natural groups”, which is appealing, but it may lead to clusters that do not correspond with household types. The other two approaches will classify according to the pre-defined patterns. The k-Nearest Neighbours will place all households in some group, but the groups may turn out very inhomogeneous in consumption patterns. The third approach will only assign households to a group, if it is close to the defining pattern, but may leave household without a group. Also, for this approach, more experiments are needed to determine the value of the threshold.

More work on implementing and testing the three approaches discussed in this paper is needed in order to determine, which approach is the better for our purpose. This may also strengthen or weaken our hypotheses about the superiority of the correlation-based distance function compared to Euclidian distance functions.

6. References

Residential Customers: Exploiting Aggregation and Correlation between Households’

Sustainable Internet and ICT for Sustainability, SustainIT 2013, pp. 1–6, 2013.


