Anomaly Detection in Electricity Consumption Data

Student: Soniya Gorbani

Supervisors: Hassan Nemati, Anita Sant’Anna, Sławomir Nowaczyk

Examiner: Prof. Antanas Verikas

ITE/ Halmstad University
Sweden

Halmstad University, September 16, 2017
"What you read, without any effort, it has been written many times with great effort."

(Enrique Jardiel Poncela)
ABSTRACT

Distribution grids play an important role in delivering electricity to end users. Electricity customers would like to have a continuous electricity supply without any disturbance. For customers such as airports and hospitals electricity interruption may have devastating consequences. Therefore, many electricity distribution companies are looking for ways to prevent power outages.

Sometimes the power outages are caused from the grid side such as failure in transformers or a break down in power cables because of wind. And sometimes the outages are caused by the customers such as overload. In fact, a very high peak in electricity consumption and irregular load profile may cause these kinds of failures.

In this thesis, we used an approach consisting of two main steps for detecting customers with irregular load profile. In the first step, we create a dictionary based on all common load profile shapes using daily electricity consumption for one-month period. In the second step, the load profile shapes of customers for a specific week are compared with the load patterns in the dictionary. If the electricity consumption for any customer during that week is not similar to any of the load patterns in the dictionary, it will be grouped as an anomaly. In this case, load profile data are transformed to symbols using Symbolic Aggregate approXimation (SAX) and then clustered using hierarchical clustering.

The approach is used to detect anomaly in weekly load profile of a data set provided by HEM Nät, a power distribution company located in the south of Sweden.

Keywords: electricity consumption, smart meter data, symbolic representation, anomaly detection
ACKNOWLEDGEMENTS

I would like to express my special thanks of gratitude to Prof. Antanas Verikas as well as my supervisors Hassan Nemati, Sławomir Nowaczy, and Anita Sant’Anna who gave me the opportunity to do this project. You all helped me in doing a lot of research and I came to know about so many new things.

I would also like to thank my parents (Soraya Shakoori - Mozafar Ghorbani), husband (Majid Ghanadi), all family members (Hossein Ghorbani) and friends (Philomena Igoe and Coraline Zhang) who helped me a lot in finalizing this project with their support.
**CONTENTS**

1 **INTRODUCTION** 1  
   1.1 Problem Statement 2  
   1.2 Background 3  
      1.2.1 Smart Grid 3  
      1.2.2 Smart Meters 3  
      1.2.3 Electricity Customers 4  
      1.2.4 Power Consumption 4  
      1.2.5 HEM 4  

2 **RELATED WORK** 7  
   2.1 Load Profile Clustering 7  
   2.2 Clustering Methods 10  
      2.2.1 K-means Clustering 10  
      2.2.2 Hierarchical Clustering 10  
      2.2.3 Fuzzy C-means (FCM) Clustering 11  
      2.2.4 Self-organization Mapping (SOM) 11  
   2.3 Anomaly Detection Methods 11  

3 **METHODS** 13  
   3.1 Load Pattern Dictionary 13  
      3.1.1 Input Data 13  
      3.1.2 Pre-Processing 14  
      3.1.3 Symbolic Representation 15  
      3.1.4 Clustering Methods 18  
   3.2 Anomaly Detection 20  
   3.3 Evaluation 21  

4 **INITIAL STEPS AND RESULTS** 23  
   4.1 Initial Steps And Results 23  
      4.1.1 Data Pre-processing 23  
      4.1.2 K-means 24  
      4.1.3 Hierarchical 24  
      4.1.4 Evaluation 24  
      4.1.5 Visualization 25  

5 **EXPERIMENT AND RESULTS** 29  
   5.1 Input Data 29  
   5.2 Pre-Processing 30  
   5.3 Symbolic Representation 32
5.4 Clustering And Creating Load Pattern dictionary 33
5.5 Anomaly Detection 36
5.6 Evaluation 38
  5.6.1 Creating Artificial Data 38
  5.6.2 Comparing Same Customer With Itself 38

6 CONCLUSIONS AND FUTURE WORK 43
  6.1 Conclusion And Future Work 43

BIBLIOGRAPHY 45
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>Process model of load classification [1].</td>
<td>8</td>
</tr>
<tr>
<td>Figure 2</td>
<td>System architecture of step 1: creating load pattern dictionary</td>
<td>13</td>
</tr>
<tr>
<td>Figure 3</td>
<td>Symbolic representation</td>
<td>16</td>
</tr>
<tr>
<td>Figure 4</td>
<td>the result of 500 means with different sample size of a data set [46]</td>
<td>17</td>
</tr>
<tr>
<td>Figure 5</td>
<td>An example of Gaussian fit for the electricity usage of all customers in each hour.</td>
<td>18</td>
</tr>
<tr>
<td>Figure 6</td>
<td>System architecture of step 2: anomaly detection</td>
<td>21</td>
</tr>
<tr>
<td>Figure 7</td>
<td>One example of artificial load profile patterns with noise</td>
<td>22</td>
</tr>
<tr>
<td>Figure 8</td>
<td>Initial system architecture by extracting features for clustering load profile data.</td>
<td>23</td>
</tr>
<tr>
<td>Figure 9</td>
<td>Result of K-means clustering using three principle components and $k = 20$.</td>
<td>25</td>
</tr>
<tr>
<td>Figure 10</td>
<td>Distance threshold of electricity consumption data and corresponding number of clusters.</td>
<td>26</td>
</tr>
<tr>
<td>Figure 11</td>
<td>Silhouette evaluation</td>
<td>26</td>
</tr>
<tr>
<td>Figure 12</td>
<td>Weekly electricity consumption in each cluster</td>
<td>27</td>
</tr>
<tr>
<td>Figure 13</td>
<td>Calendar for September 2013</td>
<td>29</td>
</tr>
<tr>
<td>Figure 14</td>
<td>(a) example of daily load consumption for customer 112367 during Sep 2nd 2013 (Monday), (b) example of weekly load consumption for customer 112367 during Sep 2nd 2013 to Sep 8th 2013 (Monday to Sunday)</td>
<td>31</td>
</tr>
<tr>
<td>Figure 15</td>
<td>Gaussian distribution of the data for each specific day during Sep 2nd 2013 to Sep 8th 2013 for all customers</td>
<td>32</td>
</tr>
<tr>
<td>Figure 16</td>
<td>Most common load patterns of clusters for daily electricity consumption calculated by considering 7 day of a week within Sep 2nd 2013 to Sep 8th 2013, 5551 daily electricity consumption samples.</td>
<td>34</td>
</tr>
</tbody>
</table>
List of Figures

Figure 17  Most common load patterns of clusters for weekly electricity consumption calculated by considering 4 weeks of a month within Sep 2nd 2013 to Sep 29th 2013, 3182 weekly electricity consumption samples. 35

Figure 18  Grouping customers based on the weekly electricity consumption dictionary load patterns for the week within Sep 2nd 2013 to Sep 8th 2013, 792 weekly electricity consumption samples. 37

Figure 19  Examples of customer’s patterns who are grouped as customers with abnormal electricity consumption for the week within Sep 2nd 2013 to Sep 8th 2013. 39

Figure 20  Patterns of electricity consumption for artificial data 40

Figure 21  Examples of abnormal electricity consumption patterns 41

Figure 22  Performance of the method while increasing the noise in data 41

Figure 23  Numbers of abnormal detected customers in each week of 4 months 42

Figure 24  Numbers of in common abnormal customers detected during 16 weeks 42
LIST OF TABLES

Table 1  A lookup table containing the breakpoints that divide a Gaussian distribution in an arbitrary number (from 5 to 9) of equiprobable regions  17
Table 2  An example of part of the input data for one customer  30
Table 3  The relation between the number of clusters and the number of customers who does not belong to any of the clusters for the weekly load profile based on different $\alpha$ values.  36
ACRONYMS

DM   Data Mining
AMI  Advanced metering infrastructure
DR   Demand response
EE   Energy efficiency
SAX  Symbolic Aggregate approXimation
PAA  Piecewise Aggregate Approximation
INTRODUCTION

Today distribution grids play an important role in human life. Distribution grids are the final stage in the delivery of electric power to electricity customers. These customers would like to have a continuous electricity supply without any disturbance. It is vital for customers such as airports and hospitals to avoid any electricity interruption that could have devastating consequences. Considering this, many electricity distribution companies are looking for ways and methods to understand patterns and behaviour of electricity customers and use this information to strengthen and modify their grids.

In order to modify the power grid architecture such as strengthening fuses, breakers, and cables in the delivery lines, information about the power consumption can be collected and analysed. Today this information is available by the usage of smart meter (SM). SMs are devices used to record electricity consumption, voltage and power factor in intervals of every hour or less. Therefore, monitoring and billing become easier by two-way daily communication between the SM and Distribution System Operators (DSOs).

SM information can be used for both DSOs and customers. Customers can benefit from SM data to adjust their electricity consumption strategies more economically and optimally. DSOs can use the SM data to characterize the type of electricity users, support the production planning, create competitive market policies, and provide more personalized electric power services [2, 1].

Clustering electricity customers based on their load profile, i.e., a graph of electricity load variation over time, is a useful method for DSOs toward achieving some of the mentioned objectives. Clustering is an important task in the field of data mining and in this application is considered as grouping load profiles into sets of clusters in which load profiles in one cluster are more similar compared to load profiles in other clusters. 
[3, 4].

In general, there is a considerable difference in load profiles of different type of customers, such as domestic, commercial, industrial, etc. For industrial users the amount of electricity usage during the weekdays is usually higher than the electricity usage during weekends. Other type of customers such as residential users usually have higher electricity consumption during the weekends compare to the rest of the week.

Hourly and daily consumption data can help identifying customers with different usage behaviour i.e. some customers have special routine in using electricity. Clustering load profiles provides a unique opportunity for DSOs to understand load variation of individuals as well as groups of consumers and accordingly modify the architecture of their power grid.

In the literature, several works have analyzed electricity consumption data for energy forecasting [5, 6, 7], and demand response programs [8, 9, 10, 11, 12] for tariff settings and developing personalized price offering.

1.1 PROBLEM STATEMENT

This thesis aimed to cluster electricity users based on recorded power consumption data. Hourly and daily electricity consumption is used to group different users by finding similarities in their energy consumption patterns known as load profile. The results can be used for electricity forecasting and prediction and moreover for detecting anomaly in the shape of load profile.

In this work, the main focus is on detecting anomalies in electricity consumption data by clustering hourly and daily load profiles. In this case anomaly is defined as load profile shapes which is dissimilar to others. We used a two step approach for anomaly detection: step 1) creating a load pattern “dictionary” consists of common load profile shapes, step 2) comparing a new entry of load data with the dictionary and specifying whether the load profile is similar to any load patterns in the dictionary or not. In step 1, load profile data are transformed to symbols using Symbolic Aggregate approXimation (SAX) and then clustered using hierarchical clustering. Clusters with considerably
high number of load profiles are added to the load pattern dictionary. This dictionary is then used in step 2 for detecting anomalies in electricity consumption data.

1.2 BACKGROUND

Some of the general terms used in the thesis are defined in the following.

1.2.1 Smart Grid

Power grids are an inter related network that delivers electricity from suppliers to consumers. Generating stations that produce electrical power, high-voltage transmission lines that carry power from distant sources to demand centres, and distribution lines that connect individual customers are all included in power grids [13].

The traditional power grids are used to carry power from a few central generating stations to a large number of users or customers. Nowadays and in 21th century there is a new generation of the power grids which is called smart grids (SG). SGs use two-way communication to create an automated and distributed advanced energy delivery network [14]. SGs include automated network, with real time, two-way and cyber-secured communication [15].

SGs are self monitoring and auto balancing power grids that transmit electricity from the source, such as sun, wind or coal, to the end user. They provide clean, safe, reliable, efficient and sustainable energy by using modern hardware and software technologies [15]. These grids allow society to optimize renewable energy source usage and minimize the collective human environmental footprint.

1.2.2 Smart Meters

SM is the next generation of energy meters that can record electricity consumption and power quality remotely [16]. SM technology delivers significant benefits to consumers as well
as a more sustainable energy management and an increased security of energy supply [17]. SM information can be used for both DSOs and customers. Customers can benefit from SM data to adjust their electricity consumption strategies more economically and optimally. DSOs can use the SM data to characterize the type of electricity users, support the production planning, create competitive market policies, and provide more personalized electric power services [2, 1]. Below listed some of the benefits of Smart Meters:

- Operational cost redundancy.
- Time saving to the consumers and utility companies for reporting the meter reading back to the energy providers.
- Possibility of online electricity bill payment.
- Power consumption can be significantly decreased during the high peaks with an intimation policy.
- Has a feature of automatically terminating the appliances off when they are not in use [18].

1.2.3 Electricity Customers

Customers are defined as power electricity users. The users are include individual households, offices, companies among others.

1.2.4 Power Consumption

The total amount of power consumed by an individual user is referred to power consumption. The consumption of power is an important aspect of electricity supply.

1.2.5 HEM

The data used in this project is provided by HEM Nät, Halmstad Energi och Miljö (HEM Nät). HEM Nät is an electricity distribution company which operates in and around Halmstad (a city located in the south of Sweden). HEM is owned by Halmstad municipality and supplies electricity, district heating and total energy services to more than 45000 customers. HEM deployed
38,000 advanced electricity meters in 2007 to monitor the electricity consumption and provide feedback to customers [15].
RELATED WORK

This chapter contains a review of some of the related works. Emphasis is placed on the techniques that used for clustering electricity consumption.

2.1 LOAD PROFILE CLUSTERING

Several studies have been done specifically in relation to clustering customers based on their energy usage. A survey done by Yang, et al. [1] reviews electric load clustering methods in smart grids. This paper proposed a process model of load classification (see Figure 1) containing five stages, namely, load data preparation, load data clustering preparation, load data clustering implementation, understanding and evaluation of load classification results, and applications of load classification results. In this work the authors refer to the K-means, fuzzy c-means, hierarchical clustering, and self-organization mapping methods as the well-known clustering algorithms used for load classification. The classification results are generally presented as a certain number of groups of load shapes and their corresponding representative load patterns. According to this paper, the goal of load classification process is to support the decision-making of power systems participants.

The paper entitled "Household Energy Consumption Segmentation" by Jungsuk et al. [19] focused on the shape of the patterns and used features derived from the encoded data. The encoding system based on a pre-processed dictionary is used to create the encoded data. Clustering used in this study deals with multivariate data. The data they analysed was household data samples for a large utility. A combination of multiple clustering criteria to segment the households was employed.

Our work has some dissimilarities with work presented in [19]. In their work, they focused on the shape of the patterns and clustering electricity load profiles. One of the dissimilarities that we can name here is modelling. In their work they modelled daily electricity consumption by mixture of two log normal
distribution. This summarizes the load shape into maximum four values.

Lin et al. in [20] introduced a new symbolic representation of time series to cluster the electricity customers. The proposed new symbolic representation method allows dimensionality/numerosity reduction while for many symbolic representations of time series, dimensional of the symbolic representation is the same as the original data. This new symbolization method also allows distance measures to be defined on the symbolic approach that lower bound corresponding distance measures defined on the original series. This symbolic representation allows the real valued data to be converted in a streaming fashion, with only an infinitesimal time and space overhead. They demonstrated the utility of representation on the classic data mining tasks of clustering and anomaly detection which are the main focus of this thesis as well. This survey studied numbers of time series representations such as Discrete Fourier Trans-
form (DFT) [21], the Discrete Wavelet Transform (DWT) [22], Piecewise Linear, and Piecewise Constant models (PAA) [23], (APCA) [24, 23], and Singular Value Decomposition (SVD) [23]. One important feature of all these representations is that they are real valued. This limits the algorithms, data structures and definitions available for them. As an example, in anomaly detection the probability of observing any particular set of wavelet coefficients cannot be defined because the probability of observing any real number is zero [23]. This limitation is one of the reasons that researchers used a symbolic representation of time series. Discrete nature of this symbolic approach make it possible to tackle emerging tasks such as anomaly detection and motif discovery.

Wijaya et al. in symbolic representation of smart meter data[26] propose a symbolic representation approach to analyse smart meters large amount of data. This approach reduces smart meters data and each symbol represents a range of data. The proposed method has the ability to run on real time system data. It is possible to apply the method where the alphabet size for the symbolising is not fixed. In our case we can have the fixed alphabet size and our current data is not the real time data. The data set used in this work follows a log-normal distribution. However in our work we assume that the data is normally distributed and we do not focus on the power forecasting and instead we want to do the anomaly detection. As described there is similarities between this work and our work in using the symbolic representation approach but our work’s focus area is to find our the anomalies and this paper work has power forecasting as one of it’s focused goals.

Kaushik et al. in [23] focused on similarity search on time series data bases also by Adaptive Piecewise Constant Approximation (APCA). APCA approximates each time series by a set of constant value segments of varying lengths such that their individual reconstruction errors are minimal therefore this work was part of our interest.

The idea of using symbolic representation in this thesis originate from the work by Lin et al. [20]. In this study Lin et al. proposed a symbolization approach for anomaly detection. Thus this work is of interest. This thesis has several similarities to these studies [20, 26]. Symbolic representation approach has
been used in these works as well as this thesis.

2.2 CLUSTERING METHODS

In this section we go through some of the common clustering algorithms and describe them.

2.2.1 K-means Clustering

K-means is one simple unsupervised learning algorithm that is used to solve clustering problems. The main idea is to define one center for each cluster (k centres) [29]. To achieve this in the first step there is a need to choose the k random centres. Then each data point will be assigned to the cluster which has the closest center. At this step there is a need to do a re centring by calculating the mean of all the points in each cluster and data points assigning action will continue till the centres of the clusters do not change.

One of the effective factors on k-means is how the data is spread. Different distributions causes different results therefore placing the centres should be done in a cunning way. In this case the better choice will be to place them as far away as possible from each other [30].

The positive points of k-means it that it is simple and effective, on the other hand initial cluster center selection is effected on algorithm and it is difficult to determine suitable numbers of clusters. Noise sensitivity and usage limitation to find groups in spherical data set are disadvantages of this method [1].

2.2.2 Hierarchical Clustering

A method of clustering which builds hierarchy of clusters is called hierarchical clustering. There are two types of hierarchical clustering agglomerative (bottom-up) and divisive (top-down). Agglomerative: In this type each data point will be a cluster itself. Pairs of clusters will merge together when a cluster moves up the hierarchy [31]. Divisive: In this type all data points will be one cluster. A cluster will split recursively when a cluster
moves down the hierarchy. It is easy to implement but it is difficult to select merges and splits points [32].

2.2.3 Fuzzy C-means (FCM) Clustering

FCM [33] is a local search fuzzy clustering method. In the k-means and hierarchical clustering, a data object in a data set belongs and only belongs to one group. However, in the fuzzy clustering, each data object belongs into more than one groups with a certain degree of membership. FCM algorithm starts with determining the number of clusters followed by randomly chosen the initial cluster centers. Then every data object is assigned a membership degree to each cluster. Each cluster center point and corresponding membership degree are updated iteratively by minimizing the objective functions until the positions of the cluster centers does not change or the difference of objective function values between two iterations ranges in a permitted extent [33].

2.2.4 Self-organization Mapping (SOM)

SOM [34] is a kind of unsupervised neural networks method which contains input layer and competitive layer. There are $N$ input neurons in the input layer and $M$ output neurons in the output layer. Evaluation function is not needed in SOM, and it can identify the most significant characteristics with self-stability. SOM also has a strong ability of anti-noise.

2.3 ANOMALY DETECTION METHODS

In general there are two ways to detect anomalies: prediction-based anomaly detection, and clustering-based anomaly detection [35]. The prediction-based, is based on an observed pattern and the assumption that it is reoccurring in the future. Therefore, if this assumption does not hold true, the predicted values may be far off the measured values. In this work we focus on clustering-based anomaly detection which is based on similarity between data. In clustering-based, we assume often-observed patterns to be the usual behavior and rarely occurring patterns
to be abnormal. Following this idea, we first have to define and compute the similarity of patterns in order to detect whether a pattern occurs more than once.

In this thesis, we perform clustering-based anomaly detection. Observe that, the main focus in this work is on the shape of the patterns and clustering electricity based on load profile shapes. Therefore the anomaly is mainly related to the visual analytic of power consumption data. To cluster the data we use symbolic representation originate from [20]. k-means, hierarchical, and modified version of hierarchical clustering is used. Performing anomaly detection based on other type of clustering methods are left for future work.
In this chapter the method that we used for customer clustering is presented. In the first step, a load pattern dictionary which consist of common load profiles is created. In the second step, a new entry of load profile is compared with the dictionary and if the load profile is dissimilar with all the load patterns in the dictionary, the new entry will be classified as anomaly consumer. Each of these steps are described in a more details in the following.

3.1 LOAD PATTERN DICTIONARY

Creating load pattern dictionary from load profile data consists of several tasks (see Figure 2). First the raw input data is sent to the pre-processing step where different procedures such as data cleaning, normalization, and outlier detection can be applied on the input data. The pre-processed data is then transformed to symbols using Symbolic Aggregate approXimation (SAX). Finally hierarchical clustering is performed on symbolized data.

![Figure 2: System architecture of step 1: creating load pattern dictionary](image)

3.1.1 Input Data

The input data is the electricity consumption of several customers within a period of time. This electricity consumption is hourly measurement data recorded by smart meters. Depending on the type of the load profile shape that one would like to cluster e.g. daily, weekly, monthly, the data will be modified. For
example, if clustering of the weekly load shape is investigated, the electricity consumption during every week will be used. Accordingly every weekly data contains 7 values or 7 features.

Furthermore, the period of time when the data is recorded can be one week, one month, or more. Usually the electricity consumption load shapes are highly correlated to the change in weather (outdoor temperature). Therefore we do not recommend using electricity usage of different seasons for creating dictionary. Moreover, a load shape to be considered as a load pattern when we observe the load shape frequently. For this purpose we consider more than one day or one week of electricity usage as input data. In this case, if a customer does not have a similar load shape compare to other customers but still using electricity in a similar way (for every day or every week), we expect the method to cluster the customer as normal.

3.1.2 Pre-Processing

The first step in pre-processin is about smoothing the noisy data and filling in missing values. Missing data (missing values) occurs when no observation was made in a specific time stamp and may happen when a SM fails to record, store, or send the data to the DSO. There are several methods introduced in literature [36] that can be used to deal with missing data. A very trivial method is to estimate the missing values according to the previous or next correctly received values.

The next step after data cleaning is normalization. In general, normalization is defined as adjusting measured values to a notionally common scale [37]. There are various normalization methods which depend on the type of data and application such as standard score, student's t-statistic, and studentized residual. In this work we used the following normalization formula known as standard score to have load profile data with mean of zero and standard deviation of one. Using standard score normalization keeps the shape of the load profile.

\[ S_{\text{customer}_c} = \frac{e_i - \mu_i}{\sigma_i} \] (1)
3.1 Load Pattern Dictionary

Where $\mu_i$ and $\sigma_i$ are the mean and standard deviation of the electricity consumption for customer $c_i$ within a week, and $e_i$ is daily electricity consumption for customer $c_i$ in day $i$.

In a data set an outlier is a value that has a considerable distance from other measured values. If there is a value with a much larger distance compare to other values it is said the value "lies outside". An outlier may be caused by a variability in measurements or it may indicate experimental error [38]. Outlier detection helps to control the range of the values after the normalization task. For outlier detection we used the simple classical standard deviation approach. In this case any value which is not within the range of $\mu \pm 3\sigma$ is considered as outlier. The variables $\mu$ and $\sigma$ are the mean and the standard deviation of the all data set within each day.

3.1.3 Symbolic Representation

There are many representations of time series data proposed in literature [39, 22, 24, 40], to facilitate similarity search. Symbolic interval time-series has been shown an appropriate data format for discovering temporal knowledge [41, 42]. An important practical advantage of working with symbols is that the efficiency of numerical computations is greatly improved through compression. Furthermore, symbolic data is often less sensitive to noise in measurements [43].

If we assume that time series sub-sequences tend to have a Gaussian distribution we can symbolize the data using Symbolic Aggregate approXimation (SAX) method [41]. Let’s assume $C : C_1, C_2, ..., C_n$ are the daily measurement of the electricity consumption for a customer after the normalization. Then the representing symbol $\hat{C} : \hat{C}_1, \hat{C}_2, ..., \hat{C}_w$ is a string of arbitrary length $w$. The number of alphabet can be defined by $\alpha$ e.g. for $\alpha = 5$ then alphabet = \{a, b, c, d, e\}.

In case of reducing the arbitrary length $n$ to $w$, known as dimensionality reduction in SAX, we first transfer the data into Piecewise Aggregate Approximation (PAA) representation [23, 44], $\tilde{C} : \tilde{C}_1, \tilde{C}_2, ..., \tilde{C}_w$, using the following formula and then symbolize the PAA representation into a discrete string.
Figure 3: Symbolic representation

In the matter of having one measurement per hour, we have 24 values per day $n = 24$ (24 dimensional time series). We can reduce this 24 dimension into a time series with smaller dimensions, e.g., 3 dimensions by using equation (2) and $w = 8$.

$$\tilde{c}_i = \frac{w}{n} \sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}i} c_j$$  \hspace{1cm} (2)

Accordingly, to reduce the time series from $n$ dimensions to $w$ dimensions, the data is divided into $w$ equal sized “frame”. Then the data-reduced representation is the mean value of the data falling within a frame. An example for symbolic representation using SAX is shown in Figure 3.

The central limit theorem [45] states that given a distribution, the sampling distribution of the mean approaches a normal distribution as the sample size increases. In fact, no matter what the shape of the original distribution is, the sampling distribution of the mean approaches a normal distribution. Note that, the sample size for each mean is not equal to the number of samples. In the sampling distribution the number of samples is assumed to be infinite. For example, Figure 4 shows the result of 500 means with different sample size of a data set.

We can assume that our data points are samples of means of the orginal independent and identically distributed data, therefore are normally distributed. Thus even though we might not know the shape of the distribution where our data comes from, the central limit theorem says that we can treat the sampling distribution as if it were normal. Then we can determine
the breakpoints that will produce \( \alpha \) equal-sized areas under Gaussian curves. This breakpoints are a sorted list of numbers \( B = \beta_1, \ldots, \beta_{n-1} \) such that the area under a \( N(0,1) \) Gaussian curve from \( \beta_i \) to \( \beta_{i+1} \) is equal to \( 1/\alpha \) [41]. Table 1 shows an example of the breakpoints for values from 5 to 9 [25].

Table 1: A lookup table containing the breakpoints that divide a Gaussian distribution in an arbitrary number (from 5 to 9) of equiprobable regions

<table>
<thead>
<tr>
<th></th>
<th>( \alpha = 5 )</th>
<th>( \alpha = 6 )</th>
<th>( \alpha = 7 )</th>
<th>( \alpha = 8 )</th>
<th>( \alpha = 9 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_1 )</td>
<td>-0.84</td>
<td>-0.97</td>
<td>-1.07</td>
<td>-1.15</td>
<td>-1.22</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>-0.25</td>
<td>-0.43</td>
<td>-0.57</td>
<td>-0.67</td>
<td>-0.76</td>
</tr>
<tr>
<td>( \beta_3 )</td>
<td>0.25</td>
<td>0</td>
<td>-0.18</td>
<td>-0.32</td>
<td>-0.43</td>
</tr>
<tr>
<td>( \beta_4 )</td>
<td>0.84</td>
<td>0.43</td>
<td>0.18</td>
<td>0</td>
<td>-0.14</td>
</tr>
<tr>
<td>( \beta_5 )</td>
<td>-</td>
<td>0.97</td>
<td>0.57</td>
<td>0.32</td>
<td>0.14</td>
</tr>
<tr>
<td>( \beta_6 )</td>
<td>-</td>
<td>-</td>
<td>1.07</td>
<td>0.67</td>
<td>0.43</td>
</tr>
<tr>
<td>( \beta_7 )</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.15</td>
<td>0.76</td>
</tr>
<tr>
<td>( \beta_8 )</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.22</td>
</tr>
</tbody>
</table>

We have chosen values between 5 and 9 by performing some experiments (clustering) and comparing the results, suggested by Lin et al. [20].

To fit the Gaussian distribution we consider all the electricity consumption for each time stamp (every hour or every day depending on the daily or weekly load shape) to determine the
Figure 5: An example of Gaussian fit for the electricity usage of all customers in each hour.

parameters. For example, if hourly load shape is considered, all the consumption data corresponding to time 00:00 will be used for finding the Gaussian distribution parameters for this time. Then, for the rest of the 23 hours of the day this procedure of fitting Gaussian with the data will be repeated. In fact, every hour correspond to a different Gaussian fit and different parameters. Figure 5 illustrates an example of this procedure for the electricity usage of all customers in each hour. After this step, electricity consumption of each customer, individually, will be symbolized using this global fit at each hour.

3.1.4 Clustering Methods

In this work we used hierarchical clustering with some modification. In general, hierarchical clustering can be bottom-up or top-down. In bottom-up clustering, initially, every data point is assigned to one cluster. By considering a connection criteria such as distance the method iteratively merges the clusters until all the data objects are in single cluster or fulfilled the termination condition (see algorithm 1).

The top-down hierarchical clustering starts by assigning all items to one cluster and then dividing the clusters based on a split criteria such as distance. The method iteratively split the clusters until the centres of the clusters do not change (see
Algorithm 1 Hierarchical clustering bottom-up

1. Start by assigning each item to its own as a cluster. So for N items, there will be N clusters.

2. Merge those two that have the closest (most similar) pair of clusters into a single cluster.

3. Calculate the distances (similarities) between the new cluster and each of the old clusters.

4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N.

Algorithm 2 Hierarchical clustering top-down

1. Start by assigning all items to one cluster. So for N items, there will be 1 cluster.

2. Divide to two cluster that have the furthest (not similar) items in each cluster.

3. Calculate the distances (dissimilarities) between the new cluster and each of the old clusters.

4. Repeat steps 2 and 3 until all items each are clustered into a single cluster. For N item it will be N clusters.

The distance measure used in our hierarchical clustering is defined by the following:

$$\text{cell}_{r,c} = |r - c|$$ (3)

The $|r - c|$ corresponds to a value within $[2, a]$ depending on the ordering of the alphabet. For example the $|r - c|$ between alphabet $a$ and $b$ is 1, the $|r - c|$ for $a$ and $c$ is 2, the $|r - c|$ between $a$ and $c$ is 3, and so on.

Normally for multidimensional features, hierarchical clustering uses the sum of the distances for all the features as the
outcome of the similarity measure. For example in case of daily consumption data, the sum of the all distances between every hours (here each hour is one feature) will be considered as the overall distance. However, using this sum of the distances could not cluster the customers based on their load profiles correctly.

To avoid the problem related to using the distance measure in hierarchical clustering (this will be explained in the next chapter), we considered the distances of every hour (feature) separately. In fact, the outcome of the distance measure will be a 1-dimensional matrix corresponding to the distances of every individual feature. Therefore we implemented a clustering method similar to the hierarchical clustering bottom-up. The difference is in comparing the distances and grouping. In the first step each customer assigns to one cluster. Then merge those two that have the closest distances between their 1-dimensional matrix. Repeat this until there is no change in the clusters.

We consider the mean and standard deviation of load profile data in each cluster to be representative of patterns and the corresponding confidence bounds in the load pattern dictionary. To create the dictionary we take the mean value of the load profiles with in a cluster. Therefore, the dictionary contains load patterns of (weekly or daily) consumption data.

3.2 Anomaly Detection

In the next step, every new load profile entry first will be symbolized using SAX and then compared to the patterns in the dictionary. If the new entry is not similar to any of the load patterns in the dictionary it will be grouped as anomaly consumption. This process is illustrated in Figure 6. Note that the process of symbolic representation for the new entry is same as section A.2. Symbolic representation.

3.3 Evaluation

We performed two tests to evaluate the accuracy of the method in detecting anomalies:

1. Creating artificial abnormal data as weekly load profile, then increasing random noise
2. Comparing same customer with itself during different intervals (different weeks of different months)

The procedure of creating artificial data is explained in the following:

1. Creating artificial load profiles patterns for a certain number of customers within a month to be used in the dictionary. To do this, we create some artificial load profiles, for example four. Then, we regenerate the same load shapes with additional small noise. Figure 7 illustrates an example of a created load profile patterns with additional noise. According to our method, customers with similar load shape will be clustered together and their load shape will be considered as normal.

2. Creating number of abnormal load profile shapes i.e creating shapes which are completely different from the patterns in the dictionary.

3. Adding Gaussian noise to the data abnormal load profile shapes.

4. Running the program and counting how many of the artificial abnormal load shapes are actually detected as anomaly.

5. Repeat step 4 and 5 several times and take the average of detected number of anomaly load shapes.

6. Repeat step 3 and 5 several times and increasing the randomness in the artificial data.

The procedure of comparing same customer with itself is explained in the following:
Figure 7: One example of artificial load profile patterns with noise

1. Running the method for real data and for specific period of the time e.g. only September 2013, creating dictionary, and specifying anomaly load shape customers.

2. Running the method for another period of time e.g March, June, and December.

3. Performing pairwise comparison between the list of anomalies.

4. Specifying which customers were always detected as abnormal, and which of them appear occasionally.
4.1 INITIAL STEPS AND RESULTS

In this chapter we describe the initial steps that has been done in this thesis and before using symbolic representation. Figure 8 shows these steps and the results of using k-means and hierarchical clustering are shown in the following sections.

![Diagram](image)

Figure 8: Initial system architecture by extracting features for clustering load profile data.

4.1.1 Data Pre-processing

Different steps are needed to organize data in preparation for application of clustering methods. The data cleaning and normalization steps are the same as what we have explained in chapter 3.

- Data cleaning
- Normalization
- Feature extraction
- Feature selection
In feature extraction we have chosen seven features for customer clustering and anomaly detection. These features are listed in the following:

1. \( f_1 = \min(x) \)
2. \( f_2 = \max(x) \)
3. \( f_3 = \text{mean}(x) \)
4. \( f_4 = \text{sum}(x) \)
5. \( f_5 = \text{var}(x) \)
6. \( f_6 = \frac{\text{sum(weekdays)}}{\text{sum(weekends)}} \)
7. \( f_7 = \frac{\text{mean(weekdays)}}{\text{mean(weekends)}} \)

where \( x = [\text{Total}_{\text{elec}}\text{Monday}, \ldots, \text{Total}_{\text{elec}}\text{Sunday}] \), one week of consumption data.

4.1.2 \( K \)-means

First we tried k-means clustering with different number of clusters. Figure 9 shows the result of clustering using three principle components when \( k = 20 \). In k-means clustering, the problem is related to defining the number of clusters \( k \). To avoid this manual selection we used hierarchical clustering.

4.1.3 Hierachical

A dendogram plot of using hierarchical clustering is shown in Figure 10. According to this dendogram we chose the maximum number of clusters as 30. Then we grouped customers into 30 different clusters.

4.1.4 Evaluation

On of the most commonly used evaluation methods is called silhouette coefficient[47]. The silhouette value is a measure of average distance to elements in the same cluster with the average distance to elements in other clusters. The silhouette is between -1 to 1, where a high value are considered well clustered. If most objects have a high value, this indicates that the clustering configuration is appropriate. On the other hand if points have
Figure 9: Result of K-means clustering using three principle components and $k = 20$.

If the silhouette calculation for one point is described by this formula:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where $b(i)$ is the lowest average distance of point $i$ to any other cluster, and $a(i)$ is the average dissimilarity of $i$ with all other data within the same cluster [48].

We evaluate the result of hierarchical clustering with 30 clusters. Figure 11 illustrates the result for silhouette coefficient evaluation method. According to [47] the values close to 1 correspond to a "good" clustering and close to -1 correspond to a "bad" clustering. In this figure, the size of each cluster represented by the thickness of the size for each cluster. For example, cluster number 6 contains many customers and cluster number 3 contains a few customers.

### 4.1.5 Visualization

According to the result of the evaluation, the clustering in the majority of the groups, and in particular, the groups which have
Figure 10: Distance threshold of electricity consumption data and corresponding number of clusters.

Figure 11: Silhouette evaluation

A large number of customers is quite good (silhouette coefficient close to 1). However, if we look at the weekly consumption data at each group (see Figure 12), the customers show very different behaviour. In fact, the load profile shapes in each group are not the same. This is because of using the distance measure in the hierarchical (or k-means) clustering. In these clustering methods, the distance is the sum of all the distances for all the features. The overall distances at each group might be the same compare to another group, but the load profile shapes are not. Therefore, we need another way to measure the distances for our clustering method if we want to capture the load profile shapes within
each cluster.

Figure 12: Examples of weekly electricity consumption in each cluster.
EXPERIMENT AND RESULTS

The proposed symbolic representation method and investigated clustering methods were examined with a real data set which was provided by the local distribution company (HEM Nät). In this chapter the input data and the experimental results are explained.

5.1 INPUT DATA

In this work we used electricity consumption data from HEM Nät distribution company. The data contains hourly electricity consumption for customers with fuse limit greater than 63 A. According to the Swedish law [49] all the distribution companies should collect the hourly consumption for customers with fuse limit greater than 63 A and daily consumption for customers below 63 A.

The hourly data contains the electricity consumption of 939 customers for the period between January 2013 to December 2013. From this data set we have chosen one month, Sep 2013, for further analysis. As you can see from Figure 13 this month of 2013 contains 4 working weeks and does not have any holidays during the weekdays (Monday to Friday).

![Calendar for September 2013](image_url)

Figure 13: Calendar for September 2013
An example of part of the input data for one customer is shown in Table 2. Every customer has a unique ID. For each customer $30 \times 24 = 720$ values for every hour of the month is recorded. These values are corresponding to the active energy consumption of that specific customer during each hour of 30 days of the month.

Table 2: An example of part of the input data for one customer

<table>
<thead>
<tr>
<th>Customer ID</th>
<th>Date and Time</th>
<th>Electricity Consumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customer 1</td>
<td>01.09.13 01:00</td>
<td>2.2</td>
</tr>
<tr>
<td>Customer 1</td>
<td>01.09.13 02:00</td>
<td>2.4</td>
</tr>
<tr>
<td>Customer 1</td>
<td>01.09.13 03:00</td>
<td>2.5</td>
</tr>
<tr>
<td>Customer 1</td>
<td>01.09.13 04:00</td>
<td>2.4</td>
</tr>
<tr>
<td>Customer 1</td>
<td>01.09.13 05:00</td>
<td>2.4</td>
</tr>
<tr>
<td>Customer 1</td>
<td>01.09.13 06:00</td>
<td>2.6</td>
</tr>
<tr>
<td>Customer 1</td>
<td>01.09.13 07:00</td>
<td>2.4</td>
</tr>
<tr>
<td>Customer 1</td>
<td>01.09.13 08:00</td>
<td>2.3</td>
</tr>
<tr>
<td>Customer 1</td>
<td>01.09.13 10:00</td>
<td>2.4</td>
</tr>
</tbody>
</table>

5.2 PRE-PROCESSING

In the pre-processing step we start with dealing with missing values. For this purpose if a customer does not have any recorded electricity consumption for more than one day, it will be excluded from further analysis. Accordingly, the number of customers with non-zero electricity consumption during the four weeks of Sep 2013 are 792, 793, 797, and 800 respectively. Furthermore, if a dataset for a customer contains some missing values corresponding to hourly consumption, the missing values will get the next or previous closest recorded hourly consumption. Other possibility is to replace the missing value with the corresponding value of the day before or after.

In this work the sum of hourly consumption during each day (daily consumption) and for one week is considered as load profile data. Since most of the customers are industries and commercial buildings, usually the electricity consumption during the weekend is lower than the rest of the week. By considering the weekly load profile we are able to capture these
characteristics. These weekly load profiles are then sent to the normalization step for re-scaling while keeping the shape of the load profile.

An example of this hourly and weekly load profile for a customer is shown in Figure 14. Consider that using this normalization we actually do not change the shape of the load profile of the original values.

Figure 14: (a) example of daily load consumption for customer 112367 during Sep 2nd 2013 (Monday), (b) example of weekly load consumption for customer 112367 during Sep 2nd 2013 to Sep 8th 2013 (Monday to Sunday)

The normalized values for hourly consumption of a day and daily consumption of a week are considered as the features to be used for clustering.
Figure 15: Gaussian distribution of the data for each specific day during Sep 2nd 2013 to Sep 8th 2013 for all customers

5.3 SYMBOLIC REPRESENTATION

We assume that all the normalized load profile data have Gaussian distribution at each time stamp. Considering this assumption we used SAX to symbolize the data. In this case, if the daily load profile is used for clustering then the symbolic representation string contains 24 characters corresponding to each hourly consumption; and, if the weekly load profile is used for clustering then the symbolic representation string contains 7 characters corresponding to the sum of daily consumption for 7 days of a week, starting from Monday.

An example of a weekly load profile for the period during Sep 2nd 2013 to Sep 8th 2013 for all customers is shown in Figure 15. We assume that the sum of the daily consumption for different days comes from different distribution. Therefore, for each day we calculate the mean and standard deviation of the all data point in that day separately. Then we fit the Gaussian distribution to the data for each specific day. After that we symbolized the data. Similarly in daily load profile we have 24 hours and correspondingly 24 mean and standard deviation for each hour of the day (from hour 01:00 to 24:00).

To symbolize the data we can use different breakpoint size of alphabet $\alpha = 5, 6, 7, 8, 9$ and after clustering we select the best number of $\alpha$. Since the Gaussian for each daily or weekly profile does not have the form $N(0,1)$, the breakpoints are calculated
by the following \( N(\mu_i, \sigma_i) \). The same break point is used for data at each day.

\[
\hat{\beta}_i = \mu_i + \beta_i \cdot \sigma_i
\]

where \( \hat{\beta}_i \) is the breakpoint values defined in Table 1.

5.4 CLUSTERING AND CREATING LOAD PATTERN DICTIONARY

The clustering method we have used is bottom-up hierarchical clustering. Therefore at the beginning every customer belongs to a different cluster. If clusters are close enough (the sum of the distance is below 24 and maximum distance is one), they will be joined together. We performed the clustering to create a load pattern dictionary. This clustering is not only based on one day or one week of electricity consumption but it is based on several days and weeks. Meaning that, for example, for the daily load profile dictionary we have considered 7 days of a week of all the customers. Accordingly for the weekly load profile we have considered 4 weeks. The advantage of this is to avoid the sensitivity of the load pattern to a specific day or a specific week. After this, the load profiles which at least more than 1% of the customers are consuming electricity similar to them will be considered as common load patterns and will be added to the dictionary. It is also possible to improve the sensitivity of the load patterns by adding more weeks and more days from other months of a year, but in this work we only worked with the collected data from Sep 2013.

We performed the symbolization by different number of \( \alpha \) from 5 to 9. The results of the number of clusters and the number of data which does not belong to any of the clusters for the weekly load profile are presented in Table 3. In this case we have chosen the \( \alpha = 7 \) since the number of un-clustered data are below thirty percent of the total number of data and also the standard deviation of the clustered data is less than \( \alpha = 5 \) and 6. The results for weekly load profile clustering using 7 alphabet shows 61 common load patterns and for daily load profile shows 71 common load patterns. The six first most common daily and weekly load patterns with the corresponding standard deviation are shown in Figure 16 and Figure 17 respectively.
Figure 16: Most common load patterns of clusters for daily electricity consumption calculated by considering 7 day of a week within Sep 2nd 2013 to Sep 8th 2013, 5551 daily electricity consumption samples.
Figure 17: Most common load patterns of clusters for weekly electricity consumption calculated by considering 4 weeks of a month within Sep 2nd 2013 to Sep 29th 2013, 3182 weekly electricity consumption samples.
Table 3: The relation between the number of clusters and the number of customers who does not belong to any of the clusters for the weekly load profile based on different $\alpha$ values.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of clusters</td>
<td>59</td>
<td>76</td>
<td>61</td>
<td>59</td>
<td>63</td>
</tr>
<tr>
<td>Number of un-clustered data</td>
<td>205</td>
<td>499</td>
<td>930</td>
<td>1179</td>
<td>1332</td>
</tr>
</tbody>
</table>

In Figure 16 the plots in the first row, the right plot in the second row, and the left plot in the third row look very similar. However, there are some differences. For example, the start hour of the peak in consumption is not hour 5 in all of the plots. The end hour of the peak in consumption is not hour 18 in all the plots. The duration from low consumption and the peak in consumption and vice-versa is not the same. The quantitative measure for detecting these differences is the corresponding symbolic representation of the load profile data.

5.5 ANOMALY DETECTION

This is the second step in our proposed method which is related to anomaly detection. In this case the load profile data for customers in one day or in one week (depending on type of the load profile) are compared to the load patterns. If they are similar (the sum of the distance is below 7 and maximum distance is one) to any of the clusters in the load pattern dictionary it will be grouped as that cluster. If the customer load profile shape is not similar to any of the load patterns in the dictionary it is grouped as a customer for abnormal electricity consumption for that specific day or week. September 2013 month’s data has been used to build the dictionary and the first week of the same month has been used for anomaly detection.

Figure 18 shows the customers whose weekly load profile belongs to each of the six clusters presented in Figure 17. Accordingly, some of the patterns for customers who are grouped as customers with abnormal electricity consumption are shown in Figure 19.

Accordingly, some of the patterns for customers who are grouped as customers with abnormal electricity consumption
Figure 18: Grouping customers based on the weekly electricity consumption dictionary load patterns for the week within Sep 2nd 2013 to Sep 8th 2013, 792 weekly electricity consumption samples.
for the week within Sep 2nd 2013 to Sep 8th 2013 is shown in Figure 19.

5.6 Evaluation

5.6.1 Creating Artificial Data

In this test, first we created 4 groups of artificial patterns for 80 customers during a month according to chapter 3, Evaluation. These patterns are shown in Figure 20. Then we added 20 artificial data which we are sure they are abnormal shapes (completely different from the 4 groups of created patterns). Figure 21 included some examples of these 20 abnormal load profiles. At this step if we run the program it can detect all 20 as abnormal. Then, we added Gaussian noise to 2 of the abnormal load profiles, then 4, and so on up to all 20 customers. The Gaussian noise is added in the creation step of generating these data. For example if the initial abnormal weekly data is \([1, 0, 1, 0, 1, 0, 1]\) then the Gaussian noise would change the fixed values to a value between \([0, 1]\). Therefore for this example, the data after adding noise may look like \([0.8, 0.6, 0.2, 0, 0.6, 1, 0.1]\).

We run the program at each step for 100 times and generate new artificial data. This gives us the mean value of the number of detected abnormal customers with a standard deviation. The result of this test is shown in Figure 22.

\[
\text{Accuracy} = \frac{\text{number of detected abnormal}}{\text{total number of abnormal}} \times 100\%
\]

5.6.2 Comparing Same Customer With Itself

To compare the detected anomaly customers with the load pattern of themselves in other periods we took the steps described in Part 3. Figure 23, illustrates numbers of customers detected as abnormal in each week of the four months March, June, September, and December.

The comparison between the customers in each list shows that only 9 customers were always detected as anomaly during these
Figure 19: Examples of customer’s patterns who are grouped as customers with abnormal electricity consumption for the week within Sep 2nd 2013 to Sep 8th 2013.
Figure 20: Patterns of electricity consumption for artificial data

16 weeks. Furthermore, some customers were detected as abnormal only once. Numbers of in common customers detected as abnormal during these weeks are shown in Figure 24.
Figure 21: Examples of abnormal electricity consumption patterns

Figure 22: Performance of the method while increasing the noise in data
<table>
<thead>
<tr>
<th>Weeks/Numbers of customers detected as Abnormal in</th>
<th>March</th>
<th>June</th>
<th>September</th>
<th>December</th>
</tr>
</thead>
<tbody>
<tr>
<td>W1</td>
<td>123</td>
<td>146</td>
<td>142</td>
<td>142</td>
</tr>
<tr>
<td>W2</td>
<td>159</td>
<td>168</td>
<td>166</td>
<td>124</td>
</tr>
<tr>
<td>W3</td>
<td>138</td>
<td>140</td>
<td>171</td>
<td>145</td>
</tr>
<tr>
<td>W4</td>
<td>115</td>
<td>166</td>
<td>142</td>
<td>171</td>
</tr>
</tbody>
</table>

Figure 23: Numbers of abnormal detected customers in each week of 4 months

<table>
<thead>
<tr>
<th>Weeks</th>
<th>In common Abnormal customers</th>
</tr>
</thead>
<tbody>
<tr>
<td>16 Weeks</td>
<td>9</td>
</tr>
<tr>
<td>12 Weeks</td>
<td>10</td>
</tr>
<tr>
<td>8 Weeks</td>
<td>11</td>
</tr>
<tr>
<td>4 Weeks</td>
<td>24</td>
</tr>
</tbody>
</table>

Figure 24: Numbers of in common abnormal customers detected during 16 weeks
CONCLUSIONS AND FUTURE WORK

6.1 CONCLUSION AND FUTURE WORK

In this work, a method for clustering consumption data and detection abnormal load profiles is described and evaluated. This method used only one month data of weekly consumption to create dictionary of load patterns. Therefore, some of the customers which may not have anomaly consumption load shape were also clustered as abnormal. If we consider more than a month e.g. one season of consumption data to create the load pattern dictionary, the clusters would be less sensitive to small changes. Furthermore, the customers who have the same usage pattern every day but completely different from others would not be clustered as abnormal.

In the evaluation method we compared a customer with itself in four different month of a year. This was mainly because of the fact that the selected four months were the only months in the year 2013 with 4 weeks that starts with Monday. But they are from different seasons and this would affect the results.

Unfortunately we could not find the type of the customers which are clustered as abnormal. This is very important for evaluating the method and to continue the work it needs to be considered.

For future work, there is a possibility to use the yearly data set instead of one month. Dictionary can be created by yearly load profiles and customers monthly load profile can be cluster according to that. Dictionaries can also be created by using seasonal consumption such as winter or summer.

We mainly evaluated the method on the weekly consumption load profile shape but it can also be used for daily consumption data clustering, abnormal detection, and evaluation.
Hierarchical clustering was the method we used but we can try other clustering methods and compare the results.


[9] Yong-hao HUANG, Chong-qing KANG, Qing XIA, Bai-qing DU, Yuan-jing MENG, Nan-Qiang HE, Jin-cheng


KBM Consultants
301 Al Latif Center, Main Boulevard
Gulberg III, Lahore

Halmstad University
Kristian IV:s väg 3, 301 18,
Sweden

Page 1/1

Invoice No. 1429
Date: 09/08/2017
Our Ref. HH003

Invoice for the Commission for the students enrolled in autumn 2017 intake in Halmstad University Sweden.

<table>
<thead>
<tr>
<th>STUDENT NAME</th>
<th>APPLICATION NO</th>
<th>FEE(SEK)</th>
<th>COMMISSION</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Main Syed Bilal</td>
<td>9789579</td>
<td>61,500</td>
<td>6,150</td>
</tr>
<tr>
<td>2 Naveed Ullah</td>
<td>10016331</td>
<td>61,500</td>
<td>6,150</td>
</tr>
<tr>
<td>3 Arslan tahir</td>
<td>9664459</td>
<td>61,500</td>
<td>6,150</td>
</tr>
<tr>
<td>4 Asad Rehman</td>
<td>9626827</td>
<td>61,500</td>
<td>6,150</td>
</tr>
<tr>
<td>5 Sohaib Bilal</td>
<td>9679315</td>
<td>61,500</td>
<td>6,150</td>
</tr>
<tr>
<td>6 Ehsan Haider</td>
<td>10148835</td>
<td>61,500</td>
<td>9,225</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>39,975</strong></td>
<td></td>
</tr>
</tbody>
</table>

Kindly remit your payment in below mentioned Account details
A/C Title: KBM CONSULTANTS LTD
Bank: BARCLAYS BANK
A/C #: 03971449 Sort Code: 209260
IBAN: GB13BARC20926003971449
BIC: BARCGB22
Currency: GBP ACCOUNT Branch
Address: WEMBLEY AND PARK ROYAL BRANCH,
ACOM HOUSE, 36-38 PARK ROYAL
LONDON NW10 7JA London
UK
Kindly remit your payment in above mentioned Account details.
I, Soniya Ghorbani declare that this thesis and the work presented in are my own and has been generated by me as the result of my own original research.

*Sweden, September 2017*

____________________________

Soniya Gorbani,
September 16, 2017