Discovery and pattern classification of large scale harmonic measurements using data mining

Ali Taher M. Asheibi
University of Wollongong

Discovery and Pattern Classification of Large Scale Harmonic Measurements using Data Mining

A thesis submitted in fulfilment of the

requirements for the award of the degree

Doctor of Philosophy

from

University of Wollongong

by

Ali Taher M. Asheibi, BSc(Eng), MSc(Eng)

School of Electrical, Computer and Telecommunications Engineering

March 2009

Dedicated to my parents...

Acknowledgements

It is my pleasure to thank the many people to whom I am indebted for the development of this thesis. First and foremost, thanks go to my supervisors, Dr David Stirling, Professor Danny Sutanto and Dr Duane Robinson. Their dedication, knowledge and experience could not have been surpassed.

Thanks to Sean Elphic, Neil Brown and Dr Vic Smith of the Integral Energy Power Quality and Reliability Center who have responded to many technical, administrative and software related requests for assistance.

Tim Brown, Ahsan Lateef, Matthew Field and Praboda Paranavithana, presently and previously with the Integral Energy Power Quality and Reliability Center , have been the sources of many interesting discussions which have contributed to the PhD experience.

Thanks to family who exercised considerable patience even in the face of typical thesis consequences. Particular thanks go to my wife Faesa Netfa who has suffered from many such consequences and been very kind to admit it.

Certification

I, Ali Taher M. Asheibi, declare that this thesis, submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Electrical, Computer and Telecommunications Engineering at the University of Wollongong, is wholly my own work unless otherwise referenced or acknowledged. The document has not been submitted for qualification at any other academic institution.

.....

Ali Taher M. Asheibi

12 March 2009

Abstract

Harmonic monitoring is an important issue for electricity utilities and their customers. Continuous monitoring of voltage and current are required to identify any substantial harmonic events before they occur. This monitoring results in large volumes of multivariate data. Although researchers have realised that such large amounts of power quality (PQ) data hold much more information than that reported using classical statistical techniques for PQ monitoring, few have taken the opportunity to exploit this additional information. This hidden information might be of assistance in the identification of critical issues for diagnoses of harmonic problems such as, predicting failures in advance and giving alarms prior to the onset of dangerous situations.

Utility engineers are now seeking new tools in order to extract information that may otherwise remain hidden, especially within large volumes of data. Data mining tools are an obvious candidate for assisting in such analysis of large scale data. Data mining can be understood as a process that uses a variety of analytical tools to identify hidden patterns and relationships within data. Classification based on clustering is an important utilisation of unsupervised learning within data mining, in particular for finding and describing a variety of patterns and anomalies in multivariate data through various machine learning techniques and statistical methods. Clustering is often used to gain an initial insight into complex data and particularly in this case, to identify underlying classes within harmonic data.

The main data mining methodology used in this work is that of mixture modelling based on the Minimum Message Length (MML) algorithm which essentially searches for a model which best describes the data using a metric of an encoded message. This method of unsupervised learning, or clustering, has been shown to be able to detect anomalies and identify useful patterns within the monitored harmonic data set. Anomaly detection and pattern recognition in harmonic data can provide engineers with a rapid, visually oriented method for evaluating the underlying operational information contained within the data set.

A case study from power quality data upon which the MML method has been ap-

plied, was taken from a harmonic monitoring program installed in a typical 33/11kV MV zone substation in Australia that supplies ten 11kV radial feeders. Several patterns have been identified from using the MML technique on the harmonic data, such as significant high harmonic disturbances, footprints of the monitored sites, unusual harmonic events (capacitor switching, turn on televisions, air conditioners and the off peak hot water system) and detection of different abstractions (super-groups), each of which comprise similar clusters. The C5.0 supervised learning algorithm has been used to generate expressible and understandable rules which identify the essential features of each member cluster, and to further utilize these in predicting which ideal clusters any new observed data may best described by.

One difficulty with the MML algorithm when used to derive various mixture models is the difficulty in establishing a suitable stopping criterion to secure the optimum number of (mixture) clusters during the clustering process. A novel technique has been developed to overcome this difficulty using the trend of the exponential of message length difference between consecutive mixture models. First, the proposed method has been tested using data from known number of clusters with randomly generated data points and also with data from a simulation of a power system. The results from these tests confirm the effectiveness of the proposed method in finding the optimum number of clusters. Second, the developed method has been applied to various two-weekly data sets from the harmonic monitoring program used on this thesis. The optimum number of clusters has been verified by the formation of super-groups using Multidimensional Scaling (MDS) and link analysis. Third, the method was benchmarked against a commonly used fitness function technique, which has underestimated the optimal number of cluster in the measured harmonic data. This resulted from the theoretical maximum entropy equation used in calculating the fitness function that assumes the attributes are independent which is not the case in the correlated nature of the harmonic attributes. Finally, generated rules from the C5.0 algorithm were used for classification and prediction of future events to determine which cluster any new data should belong to.

List of Symbols and Abbreviations

ANN	Artificial Neural Network.
Aom	Accuracy of measurement.
DFT	Discrete fourier Transform.
D	Data set.
DWT	Direct Wavelet transform.
CT Fund.	Fundamental current.
CT Harm 3	Third harmonic current.
CT Harm 5	Fifth harmonic current.
CT Harm 7	Seventh harmonic current.
CT Harm 19	Nineteenth harmonic current.
CT Harm 49	Forty-ninth harmonic current.
CT THD	Total harmonic current distortion.
f	Frequency.
FT	Fourier transform.
FFT	Fast Fourier transform.
IEC	International Electrotechnical Commission.
K	Mixture of clusters model.
KDD	Knowledge Discovery in Databases.
KL	Kulback Lieber distance.
LV	Low voltage.
MV	Medium voltage.
MDS	Multidimensional scaling.
MML	Minimum Message Length.
MVAr	Reactive power Q.
PCA	Principle Component Analysis.
PQ	Power Quality.
Ph Fund	Fundamental voltage.
Ph Harm 3	Third harmonic voltage.
Ph Harm 5	Fifth harmonic voltage.
Ph Harm 7	Seventh harmonic voltage.
Ph Harm 19	Nineteenth harmonic voltage.
Ph Harm 49	Forty ninth harmonic voltage.
Ph Total H Dist	Total harmonic voltage distortion.
rms	Root mean square.
SOM	Self Organising Map.
ST	S-transform.
SVM	Support Vector Machine.
WT	Wavelet transform.

Publications arising from this Thesis

1. A. Asheibi, D. Stirling, and D. Sutanto and D. Robinson, *"Clustering, classification and explanatory rules from harmonic monitoring data"*, Book Chapter in *"Theory and Novel Applications of Machine Learning"*, Men Joo Er and Yi Zhou, Eds., I-Tech Education and Publishing, Vienna, Austria, February 2009.

2. A. Asheibi, D. Stirling, and D. Sutanto, *Analyzing Harmonic Monitoring Data using Supervised and Unsupervised Learning.*, IEEE Transactions on Power Delivery, Vol. 24, No.1, pp. 293-301, January 2009.

3. A. Asheibi, D. Stirling, and D. Sutanto, *Classification and Explanatory Rules of Harmonic Data*, Proc. Australasian Universities Power Engineering Conference (AUPEC 2008), 14-17 December 2008 Sydney, Australia, Paper ID: 259.

4. A. Asheibi, D. Stirling, and D. Sutanto, *Determination of the Optimal Number of Clusters in Harmonic Data Classification*, Proc. of the 13$^{\text{th}}$ International Conference on Harmonics and Quality of Power (ICHQP 2008), 28 September-1 October 2008, Wollongong, NSW, Australia, Paper 1045.

5. A. Asheibi, D. Stirling, and D. Sutanto, *Analyzing Harmonic Monitoring Data using Data Mining*, Proc. Fifth Australasian Data Mining Conference(AusDM06), 29-30 November, 2006, Sydney, NSW, Australia, pp: 63-68.

6. A. Asheibi, D. Stirling, and D. Sutanto, *Analyzing Harmonic Monitoring Data using Data Mining*, Conferences in Research and Practice in Information Technology (CRPIT), 61. Peter, C., Kennedy, P.J., Li, J., Simoff, S.J. and Williams, G.J., Eds., Australian Computer Society Inc. (ACS), 2006, pp: 63-68.

7. A. Asheibi, D. Stirling, and D. Robinson, *Identification of Load Power Quality Characteristics using Data Mining. Proc. of the Canadian Conference on Electrical and Computer Engineering, 2006. CCECE '06.*, 7-10 May 2006, Ottawa, Canada, pp: 157-162.

8. A. Asheibi, D. Stirling, S. Perera and D. Robinson, *Power quality data analysis using unsupervised data mining*, Proc. Australasian Universities Power Engineering Conference (AUPEC 2004), 26-29 September 2004, Brisbane, Australia, Paper ID: 187.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem statements and background

With the increased use of power electronics in residential, commercial and industrial distribution systems, combined with the proliferations of highly sensitive microprocessor controlled equipment, distribution customers are becoming increasingly concerned about power quality problems, in particular problems due to harmonics in power supply waveform. Electricity utilities need to provide high quality of power supply waveform to customers to guarantee operation of sensitive electrical equipment. In order to meet utility and customer requirements, continuous monitoring of voltage and current is essential. This monitoring results in large volumes of multivariate data. Although researchers have realised that such large amounts of power quality (PQ) data hold much more information than that reported using classical statistical techniques for PQ monitoring [1], few have taken the opportunity to exploit this additional information. This hidden information might be of assistance in the identification of critical issues for diagnoses of harmonic problems such as, predicting failures in advance and giving alarms prior to the onset of dangerous situations. For example data arriving from sensors in a substation may indicate an impending failure of expensive equipment due to excessive harmonics, such as power capacitors. Such

information could be made available to the operation engineers before failure occurs so that proper action can be undertaken.

Since the value of any data relies on the information that can be obtained from it, utility engineers are now seeking new tools in order to extract information that may otherwise remain hidden, especially within large volumes of data. Data mining tools are an obvious candidate for assisting in such analysis of large scale data. Data mining can be understood as a process that uses a variety of analytical tools to identify hidden patterns and relationships within data. Classification based on clustering is an important utilisation of unsupervised learning within data mining, in particular for finding and describing a variety of patterns and anomalies in multivariate data through various machine learning techniques and statistical methods. Clustering is often used to gain an initial insight into complex data and particularly in this case, to identify underlying classes within harmonic data.

1.2 Thesis objectives and methodology

The aim of this thesis is to develop a methodical approach using the classification tools of data mining techniques to classify harmonic data into distinct clusters, which can assist utility engineers or data analysts in the electricity utilities to quickly analyse large volumes of measured harmonic monitoring data. These techniques allow utility engineers to gain physical insights into the occurrence of various harmonics derived from the clustering outcomes in order to support decisions of assessing the security of operational distribution systems.

The objectives of the thesis are as follows:

1. To investigate the use of the Minimum Message Length (MML) for classifying large harmonic measurement data set into clusters.

2. To develop a novel method to determine the optimum number of clusters.

3. To apply the MML algorithm to a data set from a harmonic monitoring program

in a distribution system in Australia.

4. To determine the specific operating condition associated with each cluster obtained from the MML unsupervised learning algorithm from visual observation.

5. To apply the novel method, of determining the optimum number of clusters, to the measured harmonic data and to validate this method by super-groups formation using link analysis and Multidimensional Scaling (MDS) algorithm.

6. To interpret the obtained optimum clusters and to predict the occurrence of unusual clusters from future data using unsupervised learning algorithm.

In conjunction with the above there is an overarching need to provide the engineers with a rapid, visually oriented method of evaluating the underlying operational information contained within the clusters.

1.3 Thesis outline and summary of original contributions

In this section, a brief description of each chapter is given. In addition a flow chart of the thesis outline is shown in Figure 1.1.

Chapter 2: In this chapter a literature review is presented of the existing techniques that are used in analysing harmonic and other power quality monitoring data such as statistical technique, signal processing and classification techniques. Advantages and limitation of these techniques are also discussed. This chapter also presents a review of data mining and its applications in several areas including power quality and harmonics.

Chapter 3: In this chapter a data mining method based the Minimum Message Length (MML) technique to classify a large volume of data into clusters is presented. The MML technique has been chosen for the research work on harmonic classification in this thesis. The MML technique has been applied within the successful AutoClass [2] and the Snob research programs [3], [4]. The advantages and disadvantages of using an MML algorithm over other existing techniques in harmonic and other

power quality data analysis such as Fourier transform (FT) and Wavelet transform (WT) are also explained. The advantages and disadvantages of MML compared with alternative classical clustering techniques, such as, K-means and Fuzzy C-means are discussed and demonstrated.

Chapter 4: In this chapter a novel methodical approach based on the Minimum Message Length (MML) has been developed to determine the optimum number of clusters (or mixture model size). The proposed method was tested using data from a known number of clusters of randomly generated data points. This method was also tested with data from simulation of power system using the PSCAD®/EMTDC™ electromagnetic transient software program. The results from the tests confirm the effectiveness of the proposed method in finding the optimum number of clusters. The method was benchmarked against a commonly used fitness function technique where it was found to produce a similar number of clusters using data of independent variables.

Chapter 5: In this chapter, a harmonic monitoring program conducted in Australia between August 1999 and December 2002 to measure the harmonic currents and voltages in a medium distribution system [5] is presented. The data from this harmonic monitoring program is used in this research. Details of the method used for the monitoring process, the data captured, as well as the selected data for analysis are presented. This is followed by an introduction of how the data was prepared, or transformed, in order to suit the particular data mining algorithm based on MML.

Chapter 6: In this chapter, the data mining MML clustering algorithm is applied to the measured harmonic data of the harmonic monitoring program explained in Chapter 5. The clusters obtained can be used to identify distinct patterns of harmonic events and also to identify anomalous events associated with unusual operating conditions that often only occur for short periods of time. Supergroups of these clusters can be formulated using link analysis, wherein each Super-group is comprised of similar clusters as determined by the Kullback Leibler Distance (KL). Further basic characteristics, or rules, of each super-group are subsequently generated by the

decision tree algorithm C5.0.

Chapter 7: Using a range of mixture models (set of clusters) derived from the MML clustering algorithm of the harmonic monitoring data, specific operating condition, such as peak load, off-peak load, capacitor switching operation are detected. These operating conditions, represented by the generated clusters, can be analysed and confirmed by the operation engineers. Once the data has been classified as clusters, the C5.0 algorithm that implements supervised learning is then used to describe the essential influences/factors that form the various clusters. How to predict the occurrences of unusual clusters in future measurement data is also explained.

Chapter 8: The novel method to determine the optimum number of clusters described in Chapter 4, is applied to the measured data from the harmonic monitoring system described in Chapter 5, in a distribution system in Australia. To verify the validity of the proposed method for determining the optimum number of clusters, a large number of clusters is initially formed. Using the link analysis described in Chapter 6, the clusters are joined together to form super-groups. The result shows that similar clusters are obtained using the two techniques, although the proposed method of obtaining the optimum number of clusters provide a more robust method of obtaining the desired results. In addition, explanatory rules are generated from the C5.0 algorithm that are useful for classification and prediction of future data.

Chapter 9: The significant conclusions from this thesis and suggested recommendations, for future work are summarised and discussed.

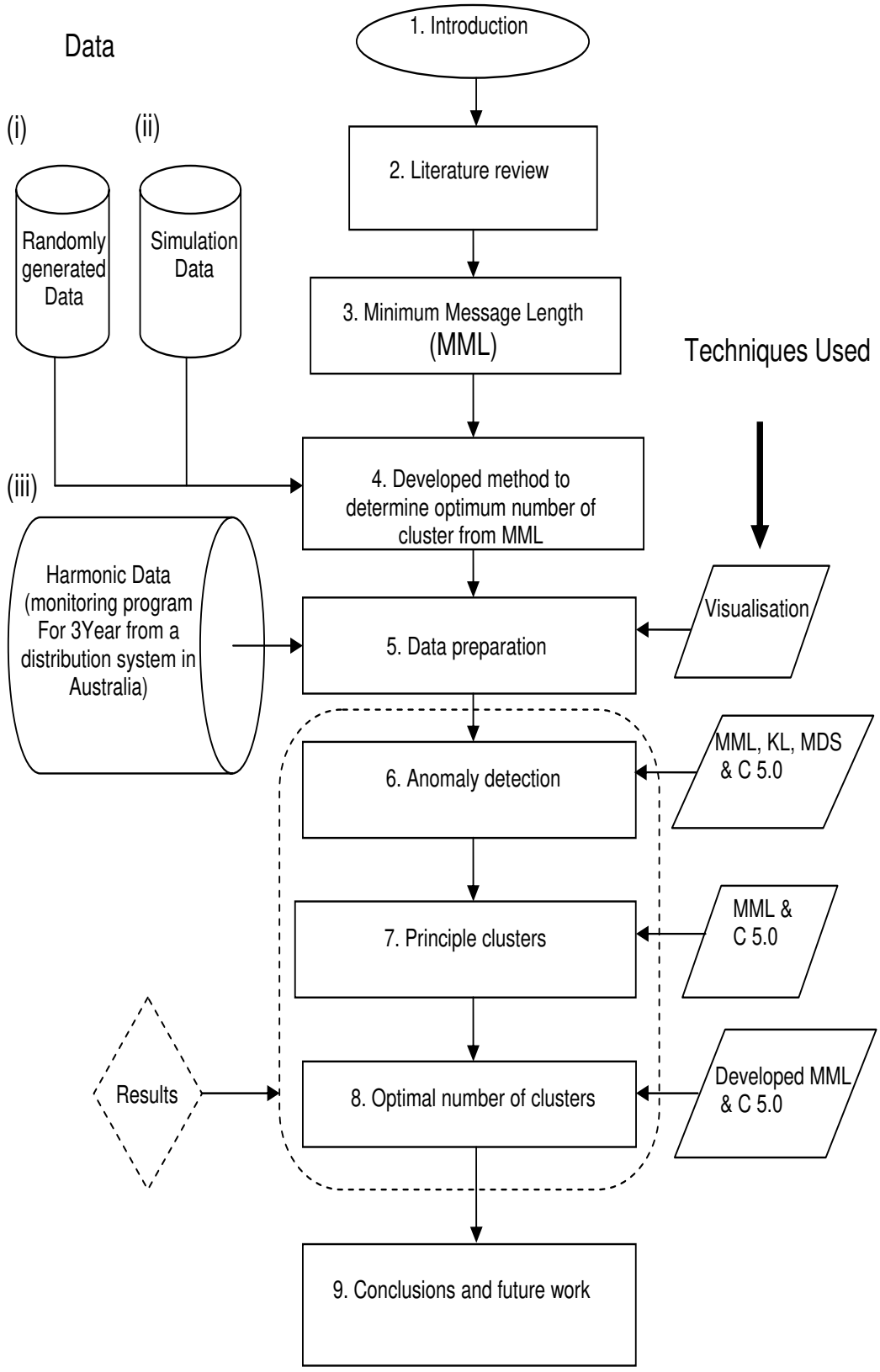

Figure 1.1: A comprehensive understanding of major building blocks of this thesis.

Chapter 2

Literature Review

2.1 Introduction

This chapter initially describes why power quality monitoring is needed and why power quality classification is required. The chapter then reviews the literature on the techniques that have been used to analyse power quality data obtained from such monitoring system, in particular statistical techniques, signal processing and classification techniques. The applications of signal processing techniques, such as Fourier, Wavelet and S-transforms when applied to harmonic data and other power quality data, are discussed. The applications of data mining in different domains including power quality area is also discussed. The applications of classification techniques to classify the power quality data including harmonic will be discussed. Finally the use of clustering technique based on unsupervised data mining is presented.

2.2 Power Quality Monitoring

Power quality (PQ) monitoring which includes harmonic monitoring is an important issue for electricity utility customers due to the increasing use of power quality distorting loads, increasing penetration of equipment susceptible to power quality disturbances and the competition in the distribution networks [6]. Power quality dis-

turbances resulting from distorting loads can cause significant financial impact due to loss of production, damage to equipment, and disruption on related manufacturing processes [7]. For these reasons, large industrial and commercial customers are becoming proactive with regards to PQ monitoring. For proper operation of highly sensitive equipment, such as computers and other electronic systems high quality ,reliable power supply is required that needs to be continually monitored using power quality monitoring system. The deregulation in the utility industry requires that utilities carry out extensive PQ monitoring programmes to retain current customers and also target new customers by ensuring disturbance levels remain within predetermined limits [8]. Power quality monitoring is often undertaken for a number of other reasons, such as verifying the adopted planning methodology, ensuring that power quality levels are within standard limits, the identification of various disturbances in the power system highlighting any new or unknown disturbances, as well as the prediction of future disturbances [6]. For both the utility and the customer, extensive PQ monitoring will eventually involve the storage and analysis of significantly large amounts of data.

2.2.1 Power quality monitoring campaigns

Several power quality surveys have been conducted to monitor power quality problems. The Canadian Electrical Association (CEA) for example, have carried out a power quality survey in 1991 for three years, participated by 22 utilities and monitored 55 sites to study and compare the frequency of voltage sags at industrial and commercial users [9]. In New York, the Niagra Mohawk Power Corporation had carried out a comprehensive power quality monitoring in 1989 for two years to monitor two 13.2 kV distribution feeders to detect the effect of power quality problems on residential customers and to generate a power quality database for further data analysis [10]. In Australia a power quality campaign has been accomplished to compare the power quality disturbances between 16 distribution companies [11].

2.3 Power quality reporting and data analysis

Power quality reporting is the utilisation of the database obtained from power quality monitoring. This power quality reporting results in abstracted information about the power quality levels for each monitored site and provides warnings about those sites that exceed the standard limits. Further advanced analysis can also be undertaken using the power quality monitoring database, like trend analysis, factor analysis and state estimation of unmonitored sites. Power quality indices are also typically generated and used in power quality monitoring databases and are explained in the next section.

2.3.1 Power quality indices

Power quality indices are indices that model various power quality levels adopted by different standards, such as International Electrotechnical Commission (IEC) 61000-3-6 ; 61000-3-7 ; 61000-4-7 ; 61000-4-30 and European Standard (EN) 50160 System performance can be assessed by comparing theses indices with the limit values from the standards. There are two levels of index reporting site indices and system indices based on the part of the network to be investigated. One or more power quality indices are used for each power quality disturbance, such as harmonic, flicker, voltage variation, voltage dips and voltage unbalance. Furthermore, specific indices such as the crest factor, transformer K-factor and telephone interference factor might also be required for more in depth investigation by a utility engineer to gain a specific insight into the monitored equipment in site or system [12]. This often results in many indices required to determine power quality levels in different parts of the network. A significant amount of research has been undertaken to reduce this number of indices to a minimum. In [13] for example, the four disturbances, namely, voltage sags, voltage variations, distortion and balance, were assigned only one global index each, combined from the other indices. Subsequently the maximum of these indices was selected to be the global index for all disturbances. A unified power quality index

suggested by Gosbell [14], is used to report on both variations and events of power quality disturbances for the three main levels in the power system: site, network and utility, so that the severity of each disturbance can be unified within each level. A comparison can then be made between the unified indices for the same level.

2.4 Signal processing in power quality data analysis

Signal processing techniques in the area of power quality are generally used to analyse the digital signals obtained from power quality monitoring data, and to extract features and information from these signals. The common features in using these techniques is the estimation of various parameters within the measured signals either in the time domain or in the frequency domain, or both. These techniques are explained in the following sections.

2.4.1 Fourier transform (FT)

The Fourier transform has been a prominent signal processing technique for a number of decades. In applying a Fourier transform, the signal is transformed from the time domain to the frequency domain, in the form of the frequency amplitude spectrum as shown in Figure 2.1.

An earlier form of the Fourier transform is the Discrete Fourier transform (DFT). Its application in power quality analysis is limited since the length of the signal being processed is assumed to be infinite $(-\infty, \infty)$, which is not the case in power quality measurement data [12].

The Fast Fourier transform (FFT) is a fast algorithm to evaluate the frequency spectrum of sampled signals. Using this algorithm, a full spectrum of the fundamental amplitude, as well as the amplitude of various harmonic frequencies of signal, can be estimated. The efficient computation of the FFT has resulted in its wide use in power system harmonic analysis of stationary signals. There are some cases of power quality disturbances, such as capacitor switching, where detailed information about

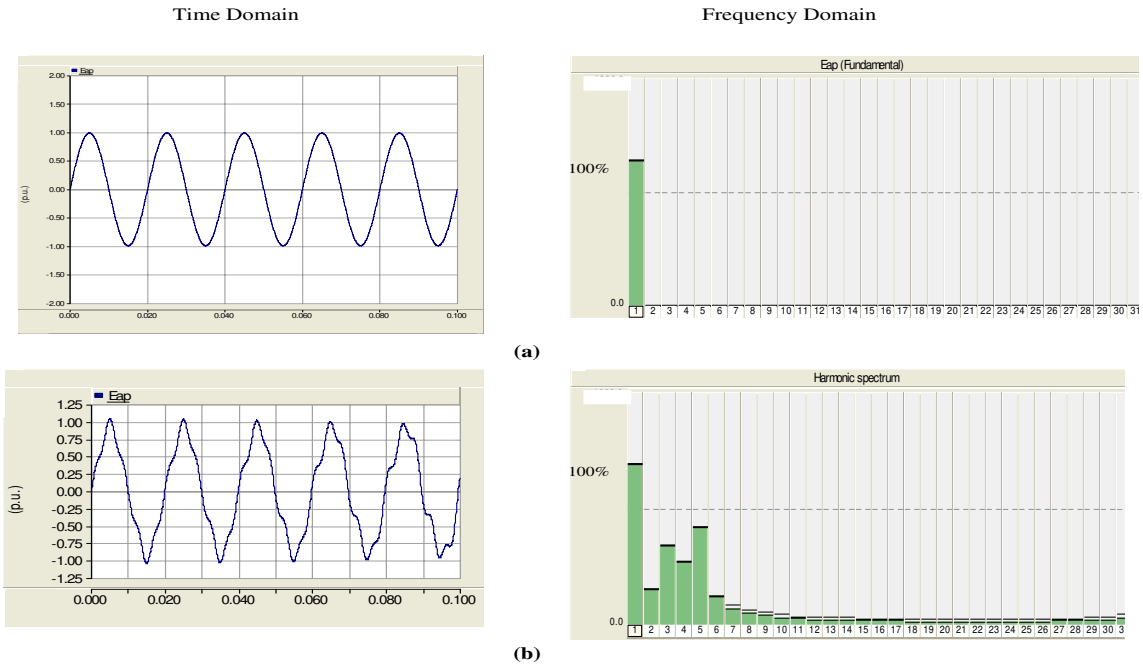

Figure 2.1: Transforming the signal from time domain to frequency domain using Fourier transform for (a) pure sine wave and (b) distorted sine wave.

the transient frequencies that occur in a specific time of the signal are needed. FFT, however, is unable to provide any information about the time instant that certain frequency component occurs as in capacitor switching, for example. In other words, FFT is inappropriate for signals with time varying frequencies, which are known as non-stationary signals.

In contrast, the Short time Fourier transform (STFT) is an attempt to overcome the disadvantage of FFT by dividing the non-stationary signal into pseudo windows of stationary signals. That is, applying a time scale for the different frequencies that exist in the measured signal, so that information of the frequencies in the signal, as well as the time they occur, can be determined. The basic formula for the STFT is:

$$STFT_x^{\omega} = \int_{-\infty}^{\infty} [x(t) \cdot w^*(t - t')] \cdot e^{(-j2\pi ft)} dt \tag{2.1}$$

where:

w - denotes the window function,

$x(t)$ - the signal to be transformed,

t - time,

t' - time shift,

f - frequency.

However, a problem of resolution arises from the size of the selected data window. If a wide window is used, then the resulting time resolution may become inadequate; alternatively selecting a narrow window will results in poor frequency resolution. Depending on the required resolution of time or frequency, there are many window types such as rectangular, Hamming, and Hanning windows [15] that can be used.

Although the choice of the size of the window can partly solve the resolution problem, the desired frequency resolution is not always maintained due to Heisenberg uncertainty principle [16], which implies that the value of frequency and time can not both be known with arbitrary precision, that is, the more precisely the frequency variable is known, the less precisely the time variable can be known and vise versa. Using a filter at a selected frequency range with a determined window size can largely improve the frequency resolution problem.

Another advantage of using a filter over STFT is that the frequency of concern is located at the center of bandpass filters, which will further facilitate a study of the frequency near this frequency of interest. Many filters are used in the literature in processing signals to analyse power quality disturbances such as Filter banks, Adaptive filters and the Kalman filter [12]. The Kalman filter, for example, is used for identification of harmonic sources and the optimal location arrangement of harmonic monitoring devices [17].

2.4.2 Wavelet transform (WT)

The Wavelet transform (WT) was first used by Grossman and Morlet in 1984 to model seismic signals [18]. From 1994 onward the WT has become a well known technique

in analysing harmonic distortion of power systems [19]. A Wavelet is a mathematical function used to divide a given function into different frequency components and study each component with a resolution that matches its scale. A wavelet transform is the representation of a function by wavelets. The wavelets are scaled and translated copies (known as "daughter wavelets") of a finite-length or fast-decaying oscillating waveform (known as the "mother wavelet"). Wavelet transforms have advantages over traditional Fourier transforms for representing functions that have discontinuities and sharp peaks, and for accurately de-constructing and reconstructing finite, non-periodic and/or non-stationary signals.

The Wavelet transform is used to detect and localise high frequency waveforms generated as a result of power quality events such as capacitor switching at power system networks [12]. However, the wavelet transform in power quality applications does not give the exact features of signals, and more computation is required to obtain accurate features [20], besides the appropriate type of the Wavelet function should ideally be chosen before being applied to a specific power quality disturbance [15].

The basic formula of the WT of the square integrable function $f(t)$ with respect to a mother wavelet $\Psi(t)$ is as follows:

$$w(s,\tau) = \int_{-\infty}^{\infty} f(t) \frac{1}{\sqrt{|s|}} \Psi^*(\frac{t-\tau}{s}) dt \tag{2.2}$$

where:

w- denotes the wavelet transform of f

s - is the scale which represent frequency parameter

τ - is the translation which represent time shift or location of the window

$*$ - denotes complex conjugation of real s and τ

A further form of this is the Discrete wavelet transform (DWT) which is a special case of the WT that provides a compact representation of a signal in time and frequency that can be computed efficiently. The DWT is used in analysing and detecting of power quality events, such as voltage sags and transients [12].

2.4.3 S-transform (ST)

Another technique that has been used in the classification of power quality disturbances is the S-transform [21]. The type of functions used in this transform are Gaussian modulated cosinusoidals. The advantage of this technique over the Wavelet transform is that the window function is a function in both time and frequency. The output of this transform is a matrix of complex number, where the rows of this matrix represent the frequencies that exist in the signal and the columns are their associated time. The S-transform is similar to the Fourier transform in that it is a reversible transform, that is, it can be transformed from time domain to frequency domain and vice versa. Using time frequency plot of S-transform contours, the power quality disturbances such as voltage sag, swell and momentary interruptions can be clearly visualised. The S-transform countours also show a better visualisation than using Wavelet transform when classifying high frequency power quality events such as capacitor switching [21]. The complete form of the ST of the continuous function $h(t)$ of time t is given by:

$$S(\tau, f) = \int_{-\infty}^{\infty} h(t) \cdot (\frac{|f|}{\sqrt{2\pi}}) \cdot e^{-\frac{(\tau-t)^2 f^2}{2}} \cdot e^{-j2\pi ft} dt \tag{2.3}$$

where:

S- denotes the S-transform of h

τ - is the parameter which controls the position of the Gaussian window on t-axis

f - represents frequency parameter

2.5 Data mining

Data mining is a process that uses a variety of data analysis tools to identify hidden patterns and relationships within data [22]. These tools are a mixture of machine learning, statistics and database utilities [23]. Data mining has recently obtained popularity within many research fields over classical techniques in analysing data due

to the following reasons:

1. Vast increase in the size and number of databases,
2. The decrease in storage device costs,
3. An ability to handle data which contains distortion (noise, missing values, etc.),
4. Continuous progress in the implementation of automatic learning techniques,
5. The rapid increase in computer technology [24].

The ultimate goal of data mining is the discovering of useful informative patterns from large amounts of data using different processes and techniques, such as classification, associations, clustering and visualisation.

2.5.1 Data mining versus statistics

A strong relationship between data mining and statistics exists as many of the techniques used in data mining algorithm are originally drawn from applied statistics. Consequently, one might say that statistics formed the basis of data mining. However, data mining processes differ from classical statistical methods in that solutions from statistical methods focus only on model estimation, while data mining techniques focus on both model formation and its performance; that is, the ability of the model to predict. Another significant difference is that statistical methods fail to analyse data with missing values, or data that contains a mixture of numeric and qualitative forms. In contrast data mining techniques can be readily used to analyse and cope intelligently with records containing missing values, as well as a mixture of qualitative and quantitative data, without tedious manual manipulation [25]. Searching for patterns in the data, where only parts of the observations are used, is another merit of data mining over statistical methods. Unlike statistical methods, where sampling is essential to avoid large scale data, the algorithms used in data mining are scaled to accommodate massive data sets. This has led to several new and successful techniques for the application of data mining domains such as text mining and web mining.

2.5.2 Data mining versus machine learning

Although basic statistical methods are routinely utilised, machine learning is considered to be the core of data mining or, Knowledge Discovery in Databases (KDD), however there are some differences between the two areas. In machine learning the focus is only on the learning process, whereas in KDD the ultimate goal is to find knowledge from the data with the help of learning. In machine learning the model or the process that initially generated the data is the most essential thing, unlike KDD where, the emphasis is on the whole data. Things that are difficult for the humans such as highly repetitive or highly computational tasks are the main concern of machine learning, whereas in KDD, interaction of an expert in the domain is crucial to evaluate the usefulness of any discovered knowledge or information. Also most machine learning algorithms used in KDD are able to scale appropriately in order to accommodate large volumes of data. The complexity of both machine learning and KDD occurs when the number of attributes increases, as the number of likely patterns increase dramatically with the number of attributes [23].

2.5.3 Data mining applications

Data mining is a global approach incorporating many techniques that have been utilised in many applications across a wide range of domains, as explained in the following subsections:

Data mining in marketing

Commercial applications, such as sales and marketing form one of the largest areas of utilisation for data mining. Due to the highly competitive nature of marketing, and the constant goal to increase profits, data mining is extensively used by large number of corporations. Typically data mining might be used to improve the advertisements of certain products to reach targeted customers in the optimal time. Customer behaviour is also investigated to develop new customer services and to identify the

most frequent buyers [26]. Clustering, classification models and association rules are generally the most used techniques in the marketing area.

Data mining in financial data analysis

Data records about customers in banks and financial institutions are generally comprehensive and of a large scale. This encourages the use of data mining to improve services such as loan and investment services. Fraud detection and forecasting stock and commodity prices are also other areas where data mining is extensively used [27]. Data mining is also used in forecasting financial disasters, such as bankruptcy to assist avoiding occurrence of such situations [26] . Decision trees, artificial neural networks and time series analysis are used in the classification and prediction in such areas.

Data mining in health care and biomedical research

As the historical records of patients are available, information about different types of diseases are now improving by the use of data mining techniques. This information is used to improve health services and to provide better diagnoses as well as improving therapy. Analysing the DNA-data sequences using data mining has led to the identification of the causes of many genetic diseases and hence the development of treatments for those diseases [28]. Visualisation and classification are the most used techniques in these areas.

Data mining for the telecommunication industry

Following deregulation in the last decade, and the rapid increase in services offered by the telecommunication industry, communication companies now compete to retain costumer loyalty by using customer service databases and telemarketing data [26]. In the telecommunication industry data mining is used to model and analyse wireless telecommunication networks by using visualisation techniques. TV program

organisers use data mining to decide the best time to show a certain program and to determine the most likely audience for that program [28].

Data mining in science and engineering

Data mining has also been successfully applied in molecular biology, astronomy, and chemical engineering. In molecular biology data mining is used to determine the macro molecular structure and protein sequencing. Neural networks has been shown to be a promising techniques in this area [28]. Data mining is also used in astronomy to categorise the large volume of collected image data. The specific techniques used in these areas are clustering and classification. In chemical engineering domains, predictive models are often built to describe the various state conditions within a chemical process. Such models facilitate the improvement of quality and also increase the productivity of a plant. A combination of artificial neural networks, fuzzy logic and statistical methods have been proved to be efficient techniques [28].

Data mining in power engineering

Due to the large amount of multidimensional data such as voltage, current and impedance, that can be generated during the operation of power system networks. Utility engineers are now beginning to rely on the classification tools of data mining techniques to support decisions of assessing the security of operation of power system [29]. Data mining is used to identify anomalies that occur as a result of a complex network or load operation, which may not be acknowledged by standard reporting techniques. Data mining methods might be used in load forecasting to build predicting models and to discover relationships between input and output variables such as weather parameters, seasonality and load profiles [30]. The induction of decision trees, being an elementary tool for defining rules and of pattern recognition in data mining, can be used to discover new unseen rules for short and long term load forecasting and the probable demand surplus/deficit arising from unusual weather

can be predicted from these rules [31].

Data mining in other areas

Data mining has been employed in many other challenging areas and is often seen as complementing the existing techniques in many disciplines [27], [32], [33].

2.5.4 Data mining in power quality data analysis

The rapid increase in computer technology and the availability of large scale power quality monitoring data should now motivate distribution network service providers to attempt to extract information that may otherwise remain hidden within the recorded data. Such information may be critical for identification and diagnoses of power quality disturbance problems, prediction of system abnormalities or failure, and provide an early warning of critical system situations. Data mining tools are an obvious candidate for assisting in such analysis of large scale power quality monitoring data. Data mining can provide answers to the end-users about PQ problems by converting raw data into useful knowledge. Essentially applying data mining tools to power quality data provides the ability to identify the various underlying contexts or classes associated with the sites monitored, and power quality disturbances of interest. There are two important learning strategies in machine learning and data mining techniques: these are Supervised Learning (SL) and Unsupervised Learning (USL). These two strategies in data mining and machine learning will be explained in the next section.

2.6 Unsupervised learning and supervised learning

Unsupervised learning generally amounts to the discovery of a number of patterns (labels), subsets, or segments within the data, without any prior knowledge of the target classes or concepts, that is learning without any supervision as illustrated in Figure 2.2(a). In supervised learning, each instance of data is mapped to align with

its associated pattern (label) in order to find the interpretation of these pattern labels as shown in Figure 2.2(b). If labels are not already identified, they can be estimated using unsupervised learning. Clustering and association techniques belong to unsupervised learning, whereas classification and estimation techniques are examples of supervised learning. Supervised learning applications like neural networks have received more attention in the power quality research studies than clustering methods using unsupervised learning. This is due to the assumption that the power quality disturbances can be identified and prelabelled in the data by the power system experts. Unsupervised learning is also often useful in this and other real domains, in order to automate labelling of new or unfamiliar data into distinct modes or classes where an expert is unavailable.

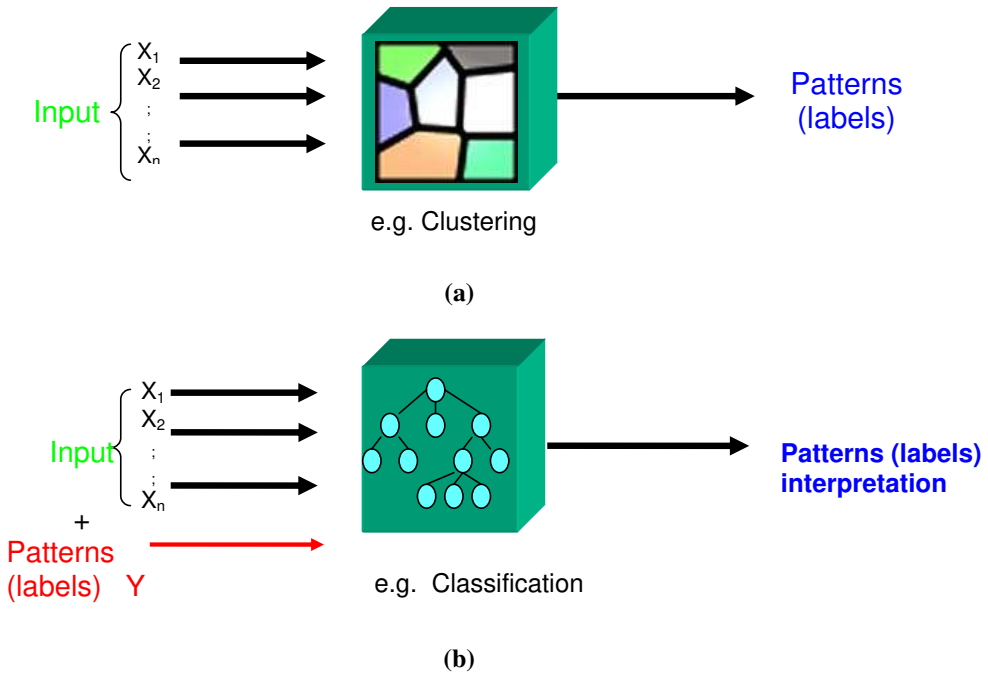

Figure 2.2: (a) Unsupervised learning and (b) Supervised learning.

2.7 Classification of power quality events using supervised learning

Classification techniques of power quality events using supervised learning are defined as learning from power quality data or from its features to identify which class an object belongs to. For supervised learning, the classes related to particular input data need to be defined first. This can be achieved using feature extraction by applying techniques described in Section 2.4, such as FFT or WT. Once the features are extracted, the supervised Learning techniques are used to learn about the relationship between input data and the defined classes. In these techniques the power quality data attributes are used as an input. Many different classifications techniques have been and are used in power quality event classification, based on the various types of the classifiers, that perform the mapping function from the input attributes to the output classes such as an Artificial Neural Network (ANN), using, for example, Multi Layered Perceptron (MLP) which is inspired by a model of the biological human neural system. Statistical based classifiers, such as Bayesian and Support Vector Machine (SVM), can also be used when the relationship between the input variables and the output classes is non-deterministic. In a deterministic relationship, non statistical based classifiers, such as rule-based expert systems classifiers are then used. In rule based expert systems, the knowledge acquired by the expert in the domain are codified as a set of "if....then....else" rules and the relevant information is automatically acted upon by the classifier when appropriate cases arise in the input variables. The above mentioned supervised learning classification methods and their applications in power quality data are explained in the following sections.

2.7.1 Supervised learning classification of power quality events using artificial neural network (ANN)

An artificial neural network (ANN) is an abstract information processing model inspired by the structure of the human brain in which a significant number of nodes (neurons) are interconnected by various directed links (axons). The strengths of these links express the information learned from the training examples. Artificial neural networks used in classification applications offer a significant capacity in handling noise, as well as, recognising patterns in data.

ANNs have been applied in power engineering fields, such as power system security and load forecasting. ANN has also been used in power quality applications, to classify different power quality disturbances. In an earlier study [34], the performances of feed-forward and time delay neural networks in the classification of power quality disturbances such as impulse, voltage sag, current and voltage distortions were compared. In another application of ANN to classify power quality [35], the wavelet transforms of the power quality data is first used to extract various features of power quality disturbances. These features are subsequently used as inputs to a multiple layer neural network classifier. This method using wavelet preprocessing has been shown to improve the ultimate accuracy of the classification.

One major disadvantage of the ANN is the over fitting problem, as the accuracy obtained in the training phase may be significantly higher than that obtained by testing a model on other data. This problem is called a problem of poor generalisation or local optimum conversion of the classifier position [28]. A second major disadvantage is the lack of transparency, or interpretability of the model structure that has been learnt. This is often contrasted to the expressive rules and decision trees readily obtained from symbolic machine learning algorithms [36].

2.7.2 Classification of power quality events using Bayesian classifiers

A Bayesian classifier is a probabilistic method in which a model is developed to update the information of an uncertain event. Proportional variables and the conditional probabilities between these variables are used in the classifier to compute the posterior probability of the event of concern. The Bayesian network illustrated in Figure 2.3 is used when the concern is in the causality of the event. Bayesian networks have been used in power quality data analysis to recognise the causes and effects of voltage sag in electrical installation. The study in [37] shows that a Bayesian network can help the plant engineer to recognise the most probable fault that causes voltage sag in its system as well as to identify the most susceptible equipment that will be affected by this sag. The reason why the Bayesians are not generally applied in power quality data analysis is because the prior probability density function (pdf) of events are not usually known.

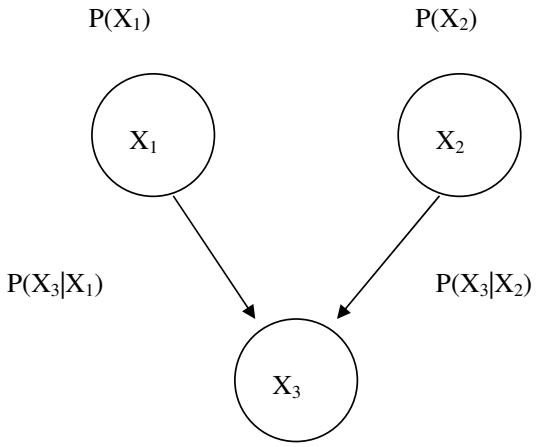

Figure 2.3: Bayesian network where x1 and x2 are independent and x3 is dependent variables. (adopted from [12]).

2.7.3 Classification of power quality events using Support Vector Machines (SVM)

In a Support Vector Machine (SVM), a classifier is constructed such that it is able to separate points belonging to two given sets in n-dimensional space into their relevant classes by transforming data into a new coordinate system which also maximises the decision boundaries. An SVM is a nonlinear classifier that can generally achieves better classifications compared to linear based classifiers, especially when the classes are not linearly separable by nature, such as an *exclusive-OR* function shown in Figure 2.4, where it is not possible to draw a straight line to separate the two classes A and B.

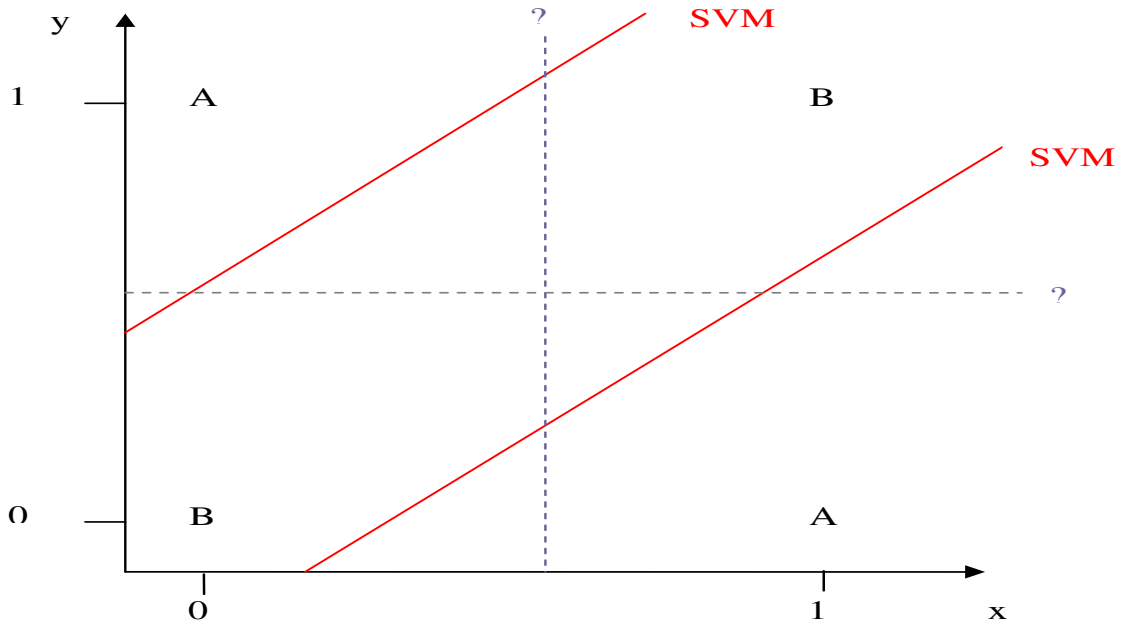

Figure 2.4: The XOR problem where the class A is defined if and only if x or y equal 1 but not both.

Unlike ANN classifier in Section 2.7.1, the SVM is able to generalise well such that it can often fit unseen data with high accuracy, resulting in high generalisation performance [38]. In power quality applications the SVM classifier has been used to

recognise different signals within PQ disturbances, such as voltage sag, spike, swell, notching, transient, harmonic and fluctuation. This is achieved by first extracting the features from the signals using techniques such as S-transforms, and the extracted features are then trained with the SVM classifier [38].

2.7.4 Classification of power quality events using expert systems

Expert systems are software systems that are used to express the knowledge of experts in the domain in the form of the following production rule:

if antecedent i; and

if antecedent j; and

if antecedent k; and

.....

then consequent C.

Expert systems have also been used in the classification of the power system events. Using this technique [39], different types of the voltage dips and interruptions in medium voltage level were classified with respect to the causes of these events. A segmentation algorithm [39] of unsupervised learning is used to divide the signal into different recognised segments, then the features for each segment are extracted. Rules derived from the experts in the field of power engineering are then used to classify these events.

2.8 Clustering as unsupervised learning

Clustering is a process that divides or segments an initial unlabeled collection of data with various attributes into a certain number of groups or clusters. As a result, the data residing in each cluster are similar, whereas data across different clusters are dissimilar. Clustering can, in part, be considered as a learning process, and as an analytical method for analysing large volumes of data which is hard to analyse as a whole, however once separated into clusters, the data in each cluster can be anal-

ysed separately. Clustering is a useful tool used in many different areas, such as in business, climate, biology, psychology and medicine [40]. It is also a useful tool for analysis of complex data sets, such as for lossy image compression in communication systems [41]. Other terms used for clustering are 'unsupervised learning', 'segmentation' and 'partitioning'. Clustering is an important technique in data mining, machine learning and communication systems. It is a powerful approach that can discover underlying and meaningful groups of data from a large database. Once clustered, supervised learning is usually used to map each instance of data to align with its associated class.

It should be noted that an expert familiar with the data and its source in the field is generally needed to interpret the discovered segmentations. Further analysis is also needed, such as experimental work or simulation to verify any derived knowledge.

2.8.1 Why clustering?

There are many benefits in learning from unlabeled samples of data, or clustering. First, features and information in the data set can be used to categorise and label the data set into separate clusters. Second, exploring and examining data as a first step is useful in designing or selecting a good classifier. Third, in some cases lablelling data in a large data set might be expensive or time consuming. Therefore clustering unlabeled data might be unavoidable. Fourth, clustering can be initially used to group large amounts of unlabeled data into a number of clusters and consequently supervised learning can be successfully used to abstract the obtained clusters using their cluster ID into natural groups. Finally, several changes in time series data lead to different output classes, if theses changes are learned by the classifier, without seeing the output classes, then more insight will be gained from the data. In summary, learning from unlabeled data is essential in pattern recognition, data exploration and classifier selection, especially, when labeled data is costly to gather, time consuming or unavailable [42].

2.8.2 Clustering objectives

It is worth noting that the objective of clustering should be predetermined. Clustering has many purposes, such as structuring, description, association and generalization. Structuring is discovering how the uncovered clusters should be organised with respect to each other, such as, hierarchical, partitioned, exclusive, non-exclusive, complete, partial, and fuzzy clusters [40]. Describing each cluster via a profile of the input attributes is useful when a prediction of a new data is needed, as whether the object can be assigned to a cluster is based on the profile of that cluster. If association or correlation between some factors is being sought, then finding a specific cluster that comprises these factors is evidence of affect or causality. The above three objectives (structuring, description, association) can be collectively attained through generalisation, in which the main features of data is sought [43].

2.8.3 Clustering algorithms and types

There are a variety of clustering algorithms used in the literature based on the proximity or similarity measure between clusters. K-means [44], for example, uses the distance measure to assign each object to the nearest cluster based on the proximity to its mean, whereas hierarchal clustering algorithms use distance measures between other prototypes of each cluster, such as the minimum, maximum, average or centroid distances. The Self Organising Maps (SOM) algorithm [45] uses competitive learning in neural networks to differentiate between clusters. Reducing the dimensionality of the data is another method used in clustering algorithms such as Principle Component Analysis (PCA) and Factor Analysis (FA) [40]. Describing clusters through probability density functions of the data is another criterion used within clustering methods. This results in the formation of mixture models, such as Gaussian Mixture Models (GMM) [40], [43]. There also exists a rich and diverse range of other clustering approaches reported in the literature, such as hierarchical (nested), partitioned (un-nested), exclusive (each object assigned to a cluster), non-exclusive (an object

can be assigned to more than one cluster), complete (every object should belong to a cluster), partial (one or more objects belong to none), and fuzzy (every object has a membership weight to belong to a cluster) [40]. The following sections consider some of the most important of these.

K-means algorithm

K-means is a simple algorithm, where the number of clusters is determined by selecting K points as initial centroids of the clusters. K-means essentially assumes that the data comes from spherical Gaussian distributions, and hence, other types of statistical distributions may not clustered correctly using the K-means. Finally K-means algorithm does not identify the attributes that are more significant in the clustering process as it assumes that all attributes have the same weight. A flow chart of the K-means algorithm is shown in Figure 2.5 [46].

The Euclidean distance is then used to assign each data point to the nearest cen-

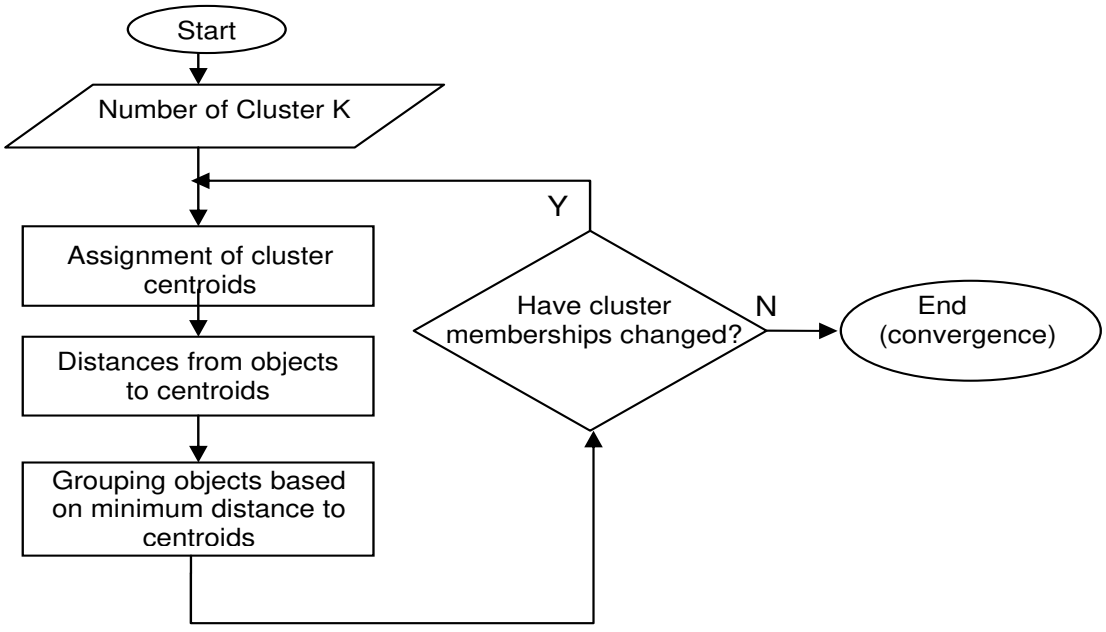

Figure 2.5: Flow chart of K-means algorithm.

troid. Once this is done the location of each centroid is then recomputed based on the members of each respective cluster. This process is iterated until the convergence is achieved [47].

The weakness of this algorithm is that the initial centroids cannot guarantee to produce the natural clusters embedded in the data, even though this process is repeated several times. Sampling and pre-clustering using another method of clustering, such as hierarchical clustering, is one solution to poor initial centroids [40]. This method also has other limitations, such as the sampling size should be small and the number of clusters should be less than the sampling size. Another disadvatage is that the K-means algorithm clusters all data points including outliers. This will significantly increase the Sum of the Square Errors (SSE) which should be kept to a minimum in order to obtain optimum clustering [44].

Fuzzy Clustering and Fuzzy C-means

In Fuzzy Clustering algorithms a weighting function is used to assign objects to all or some of the clusters with a degree of membership between 0 and 1. This degree of membership provides fuzzy clustering algorithms with an advantage over hard clustering algorithms such as K-means especially in overlapping clusters where some objects belong to more than one cluster [43]. The closest objects to the centroids of each cluster have high degrees of membership whereas the furthest objects have low degrees. Fuzzy clustering algorithms such as Fuzzy C-means, Adaptive Fuzzy Clustering, and the Gustafson-Kessel algorithm have been extensively used in pattern recognition [48]. The structure of Fuzzy C-means is similar to the K-means algorithm in the iteration steps of updating the clusters centroids and assigning objects to these centroids. In Fuzzy C-means, however any object can be assigned to any centroids not just the nearest one. The sum of the square error (SSE) is the same criterion used in the K-means algorithm except that in Fuzzy C-means a weighting function (w) is included in the calculation of SSE. This weighting function does however add to the computation time in comparison to the K-mean algorithm. Apart from this, other

advantages and limitations of Fuzzy C-mean are the same as that for the K-means algorithm mentioned earlier [40].

Self Organising Map (SOM)

A neural network method used as a form of clustering is the Self Organizing Maps (SOM) technique which was developed by Teuvo Kohonen in 1989. A self organizing map utilises a competitive learning strategy that is based on Hebian learning [45]. In this context, weights of certain network links are modified based on the similarity of the input patterns of data. As a result, the map layer of neurons respond only to the various patterns contained in the training data and essentially this is a form of unsupervised learning algorithm as the output classes are unknown. The type of unsupervised learning used in SOM is called competitive learning where the neurons compete together until one neuron wins the competition and produce an output layer (Kohonen layer) or grid formed by nodes shown in Figure 2.6 [45]. This competition is created by the lateral connections between neurons in the output layer (see Figure 2.7 adopted from [45]).

An advantage of SOM is that it generates from n dimensional data sets a two dimensional output layer (map layer) as a visual depiction of the clustering. The limitation of SOM is that the number of clusters should be determined before hand by the user.

Principal Component Analysis (PCA)

Principle Component Analysis (PCA) is a dimensional reduction technique used to discover variability in data that might be hidden. It is also used as an intermediate process prior to many clustering algorithms, as some clustering algorithms are inefficient in partitioning high dimensional data. In PCA algorithm the mean of each variable is eliminated from the data to create new orthogonal components that carry the variability of the data sorted in ascending order. The first component captures

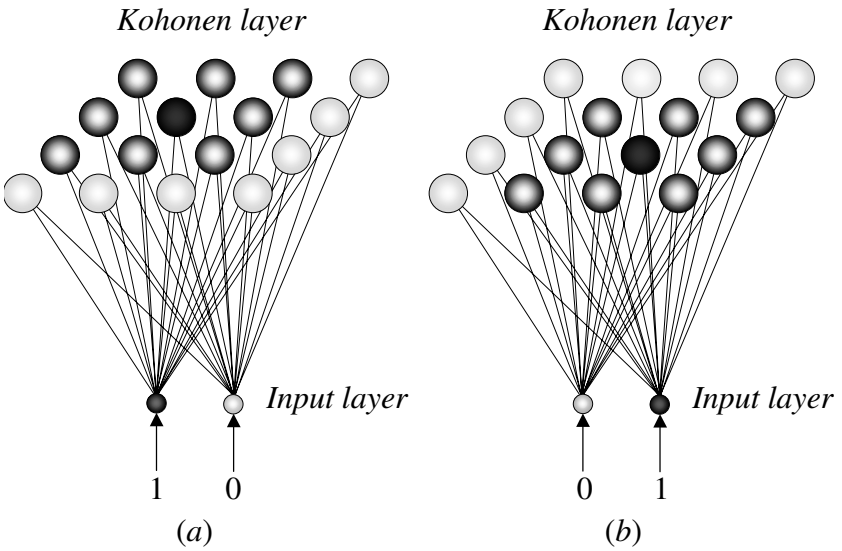

Figure 2.6: Kohonen model of feature-mapping (adopted from [45]).

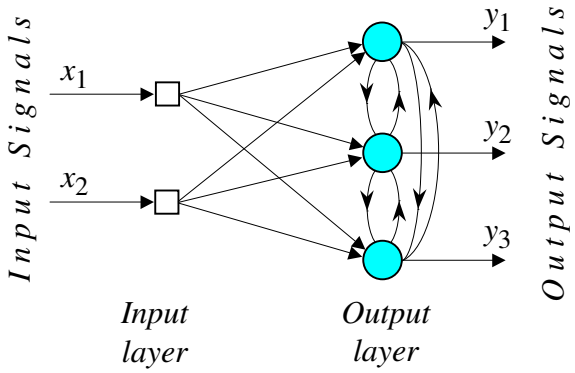

Figure 2.7: Input and output layers of SOM (adopted from [45]).

the highest variability and the second is perpendicular to the first with less variability as shown in Figure 2.8. The number of these components depends on the nature of the data and the percentage of variability needed to be captured.

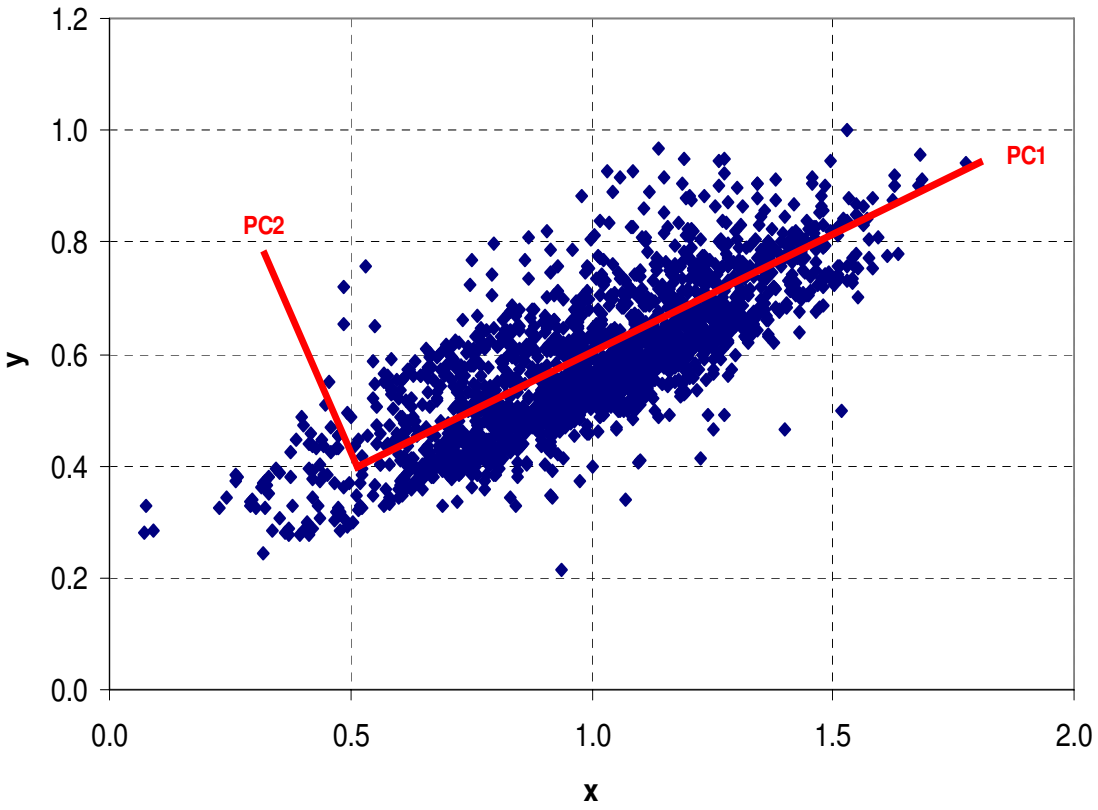

Figure 2.8: Principle components PC1 and PC2 of two dimensional data.

Normally, the variability is unevenly distributed between the variables, and hence one or two variables have the most variability. In this case the PCA works well in capturing nearly all variability in the data while reducing the dimensionality to a minimum. However, PCA fails to reduce data to a low dimensional form in the case of evenly distributed data [40]. Another criticism of the PCA is that it presents its outcomes only in terms of the transformed data, and not the raw data. PCA also assumes that the variables are correlated with each other, and so the data has a diagonal shape. Consequently this makes PCA unsuitable for independent variables. The mathematical background of PCA can be found in many statistical references

(see, for example [40]).

Mixture Modelling

Mixture modelling involves the development of an abstraction concept from several statistical distributions and is often referred to as clustering. It assumes that the distribution of data under study was generated from various simpler distributions, representing the number of clusters within the data. Any data sample, x_1, x_2,..., x_n is assumed to come from a distribution of the form

$$f(x) = \sum_{j=1}^{K} p_j \times f_j(x) \tag{2.4}$$

where

$f_j(x)$ - is a distribution in a simpler form

p_j - is the relative weight of $f_j(x)$

An example that visualises the concept of mixture modelling using Gamma distribution is shown in Figure 2.9 which explains a mixture model that encompass five gamma clusters representing the source of the data.

The normal distribution is a common distribution used to represent any random variable. The parameters of this distribution are the mean μ and standard deviation σ. To form a mixture model, the number of clusters K is firstly estimated and then the parameters (μ_j, σ_j, p_j) for each cluster are also estimated. Here the p_j represents the probability that the data comes from the j distribution. The Minimum Message Length (MML) method is an information theoretic criterion for parameter estimation and model selection which is based on mixture modelling is used in this thesis to estimate the number of clusters in data. The origin of this method and how it works for mixture modelling of Gaussian distribution along with its benchmarks against other common clustering algorithms is explained in the next chapter.

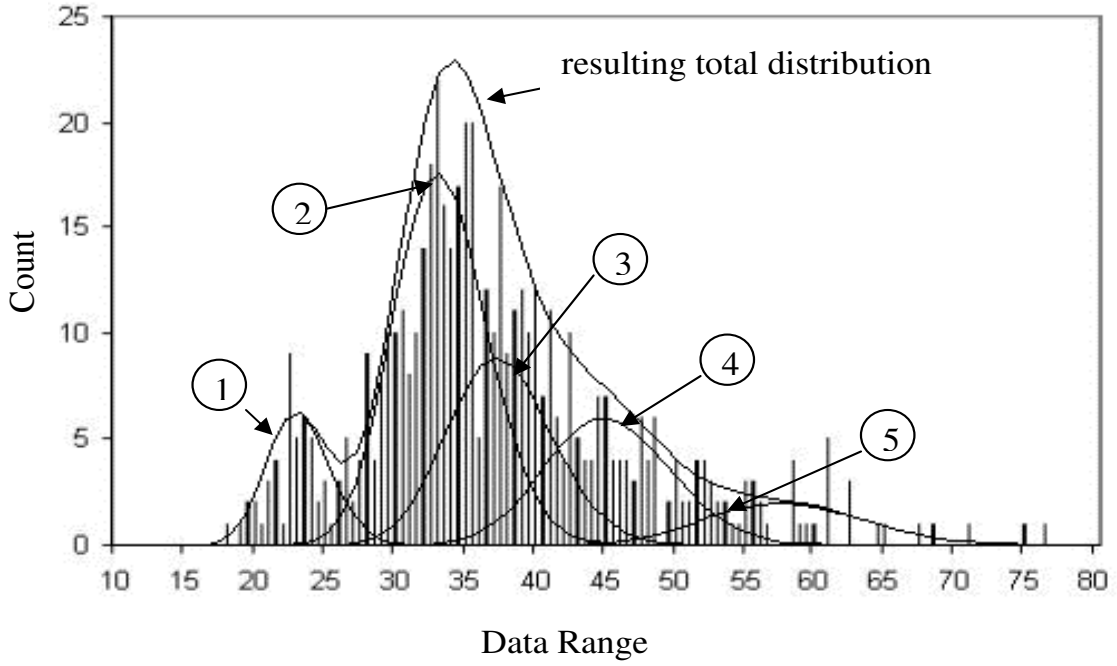

Figure 2.9: Typical mixture modelling of five Gamma distributions using MML (adopted from [54]).

2.9 Summary

In this chapter, the reasons behind the interest in power quality monitoring have been reviewed. To fully utilise the results from power quality monitoring data, a proper reporting and data analysis method has to be adopted. Indices of power quality disturbances have been used to differentiate among different sites and/or systems in terms of power quality levels.

Different methods used in signal processing, such as Fourier, Wavelet and S-transforms to extract information from power quality waveforms have been discussed and compared. The limitations, as well as the benefits of each method, have been addressed.

Due to the large database that can be generated from a power quality monitoring system, data mining is often needed to cluster and classify the data using unsupervised and supervised learning. Supervised learning or classification techniques within data

mining are powerful techniques that have been used to classify new data sets, provided that training sets of (initially) labelled data are available in advance.

Different supervised learning techniques applied in the power quality area have been compared and contrasted. Supervised learning has been used more often in the power quality area when compared with unsupervised learning due to the assumption that the data can be readily labelled by the experts. However, there are cases where unsupervised learning is needed to identify the various underlying classes associated with data without labels. A review of several unsupervised learning techniques with their limitations and advantages have been presented and discussed.

Chapter 3

Mixture Modelling Method using Minimum Message Length (MML) Technique

3.1 Introduction

This chapter presents an unsupervised learning method often referred to as the Minimum Message Length (MML) technique. The technique is based on the mixture modelling method and will be used extensively in this thesis for classifying the available power quality data from the harmonic monitoring system in a typical Australian distribution network. First, the mechanisms employed by the mixture modelling algorithm in performing the clustering through parameter estimation and model selection, with the Expectation Maximisation (EM) algorithm are presented. The background of the Minimum Message length (MML) encoding algorithm used in the mixture modelling method is then explained. The MML algorithm is compared and contrasted against other commonly used clustering algorithms.

3.2 Mixture modelling

Mixture modelling can be described as an unsupervised learning method which constructs a model based on a mixture of statistical distributions that have been learned from the data. It assumes that the distribution of the studied data is generated from a mixture of simpler statistical distributions, representing the number of clusters within the data. Any statistical distribution can be used in the mixture models, but the normal or Gaussian distribution is the most commonly used [40].

The normal or Gaussian distribution is characterized by a mean (μ), which represents the centre point of the distribution, and a standard deviation (σ), which depicts the variance of data values on both sides of the mean. If a random variable X has a normal or Gaussian distribution with mean and a standard deviation, the following shorthand notation is commonly used: (μ , σ^2). The normal distribution can be defined with its probability density function as follows:

$$P(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)}{\sigma^2}} \tag{3.1}$$

where:
x is a continuous variable between $-\infty$ and ∞ .

If x in Equation (3.1) is an n dimensional vector then the distribution is called the multivariate distribution which is defined by the equation [40]:

$$P(x) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\Sigma}} e^{-\frac{1}{2}} \cdot (x-u)^T \cdot \Sigma^{-1} \cdot (x-u) \tag{3.2}$$

where:
μ - is the mean vector
Σ - is the covariance matrix of the distribution

Perhaps the simplest example of a finite mixture model can be considered when a data population of only one numerical attribute can be represented as by a mixture model of two statistical distributions (clusters), each of which is normally distributed.

Unless each data point is labelled to denote which of the two distribution each of them belongs, it is difficult to determine the parameters of these two distributions. Visually, one can only observe the data samples which come from two combined distributions as one distribution. The simplest example of combining two normal distributions is when they have the same standard deviation and same proportion of the complete population. Figure 3.1(a) shows two normal distributions whose means (μ_1, $\mu_2 = \pm 2$) are four units apart, providing a bimodal population density curve. Changing these respective means to smaller values (μ_1, $\mu_2 = \pm 1$) will halve the distance between them, and produce uni-modal population density curve with a slightly vague combined distribution as shown in Figure 3.1(b). The probability density function of an object generating a single Gaussian distribution is given by Equation (3.1). For m objects producing a mixture of K distributions, Equation (3.1) becomes:

$$P(x/\theta_j) = \sum_{j=1}^{m} w_j \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{(x-\mu_j)^2}{2\sigma_j^2}} \tag{3.3}$$

where: w_j - is the weight that the jth distribution is chosen to generate an object.

3.2.1 Parameter estimation and model selection

The two main steps used in the mixture modelling method are parameter estimation and model selection. First, the parameters of each constituent statistical distribution and its relative weights are estimated and then the mixture model of the optimum number clusters that best represents the data will be selected.

In order to identify a suitable model, one should first estimate the values of distribution parameters that could have plausibly generated the same data values. The equation that calculates the probability of all data points, assuming these points are generated independently is given by:

$$P(x/\theta) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{\sqrt{2\sigma^2}}} \tag{3.4}$$

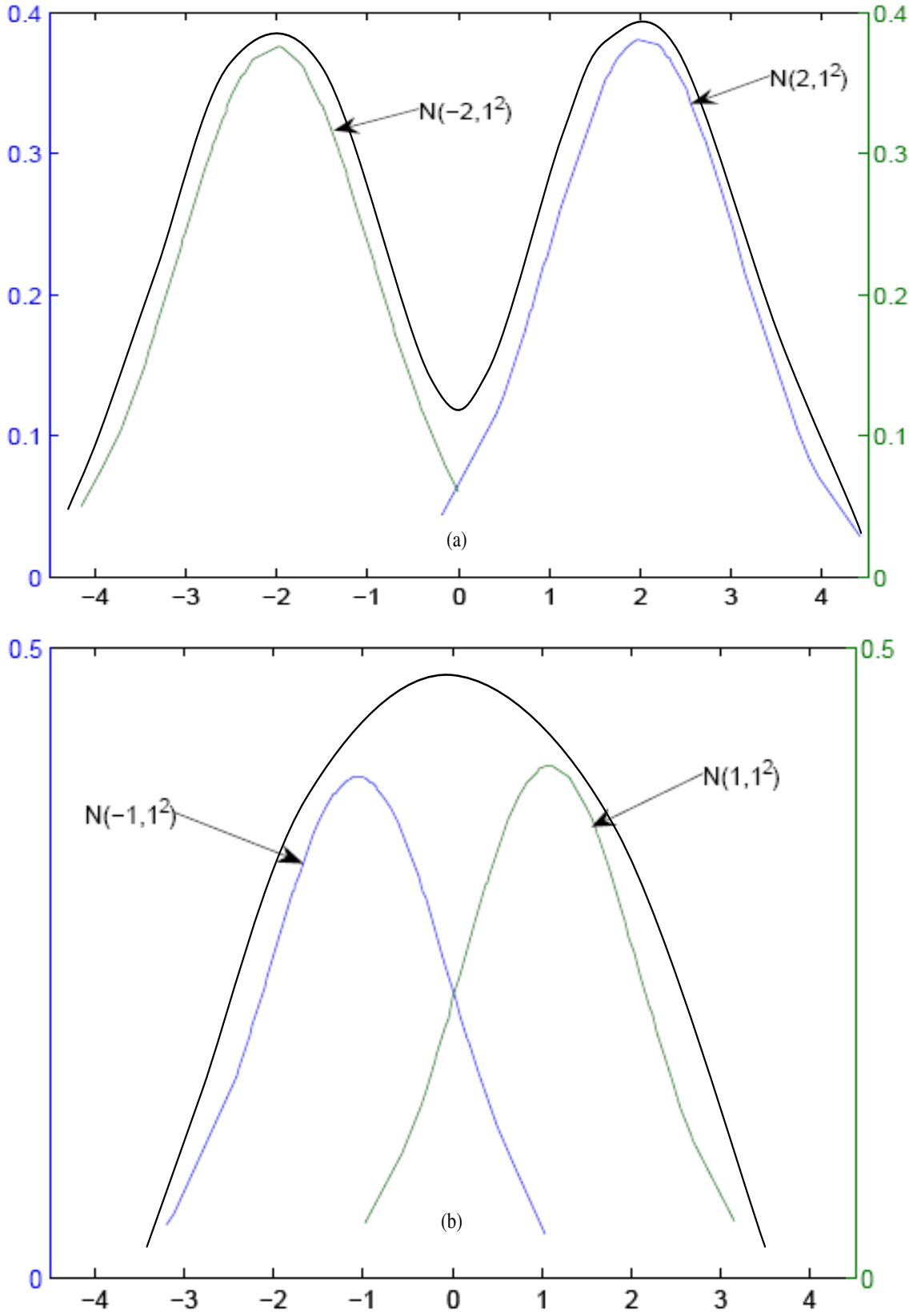

Figure 3.1: Two normal distributions of similar means, standard deviations and proportions a) $\mu 1$, $\mu 2 = 2$ and b) $\mu 1$, $\mu 2 = 1$(adapted here from various Matlab plots).

Because the probability from Equation (3.4) would be a very small value, it is preferable to work with its log value as [40]

$$\log(P) = -\sum_{i=1}^{m} \frac{(x_i - \mu)^2}{2\sigma^2} - 0.5 \log 2\pi - m log\sigma \tag{3.5}$$

In a single distribution the values of these parameters that makes the data most likely to have arisen from this distribution, can be calculated using the maximum likelihood estimation (MLE) by differentiating Equation (3.5) twice, First by differentiating it with respect to the variable μ and second with respect to the variable σ then equating the result to minimum (zero). For a normal distribution the mean and the variance can be derived directly from the data sample. In a mixture of distributions, however, the probability of each data point from Equation (3.3) cannot be calculated, since it is not known which distribution has generated which data point. The Expectation Maximisation algorithm is the solution to this problem which is explained in the next section.

3.2.2 The Expectation Maximisation (EM) algorithm [49]

It is stated in Section 3.2 that the distribution of the studied data is assumed to be generated from a mixture of simpler statistical distributions, representing the number of clusters within the data. Each cluster has a different distribution and each instance in the data is assigned to each cluster with a certain probability value. Each instance should ultimately belong to only one cluster that is associated with the highest probability value. In the case of mixture distributions it is hard to compute the probability of each data point as it is unknown which data points produce which distribution. This problem can be solved using the expectation and maximization (EM) algorithm [49]. It is an iteration algorithm that calculates the probabilities of each data point belonging to each candidate distribution within the data as well as estimating the parameters of these distributions. The two steps of the EM algorithm can be explained as follows. First, in the expectation step, initial values of distribution

parameters are selected, and the probabilities of each point, with respect to their distribution, are calculated from Bayes rule as follows:

$$P(distribution_j|x_i,\theta) = \frac{w_j p(x_i|\theta_j)}{\sum_{j=1}^{m} w_j(x_i|\theta_j)} \tag{3.6}$$

Second, the calculated probabilities are then used to estimate the distribution parameters (maximisation).

There are many mixture modelling software packages in the literature that use a combination of parameter estimation and model selection, such as Autoclass (which uses a Bayesian network) by Cheeseman and Stutz [2] , EMMIX [50] (using the maximum likelihood criterion), and Snob (using MML) by Wallace and Boulton [51], [3]. Several other software packages exist that are also based on the same two processes: parameter estimation and model selection.

3.2.3 Fitting a model to a mixture of statistical distributions

As stated above, any statistical distribution can be used within mixture models, but the normal or Gaussian distribution is the most commonly used [40]. For a Gaussian distribution $\mathrm{N}(\mu\ ,\ \sigma^2)$, there are three important areas that represent the percentage of values in the population within a given interval [52]. These percentages of values along with their related intervals are listed in Table 3.1. Two of these areas are shown in Figure 3.2 in a standard normal curve ($\mu = 0$, $\sigma = 1$).

Table 3.1: Percentage of values in the population within a given interval.

Population (%)	Interval Range
68	(μ-σ , μ+σ)
95	(μ-2σ , μ+2σ)
98	(μ-3σ , μ+3σ)

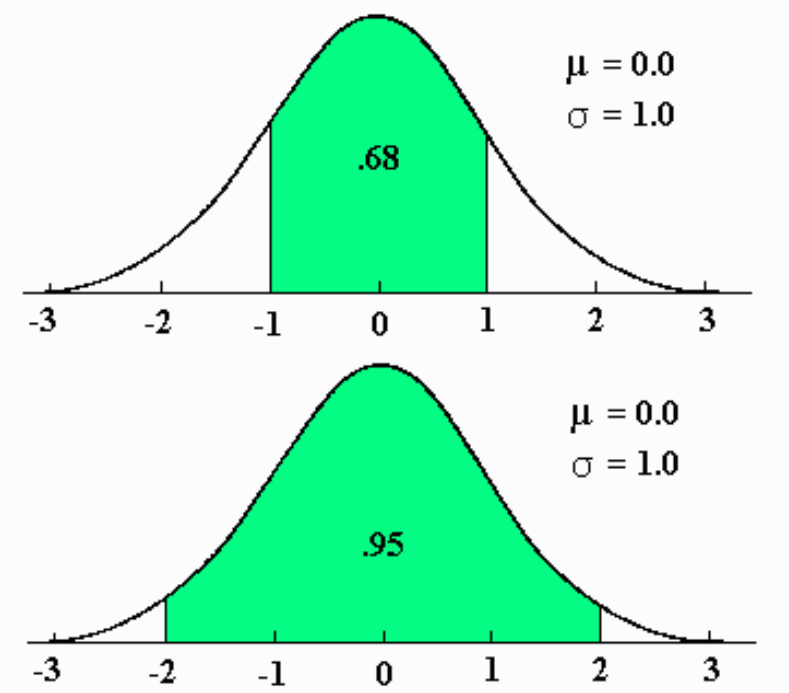

Figure 3.2: Most important areas in normal distribution (a) 68% and (b) 95% of values in population.

Fitting a single distribution, such as a Gaussian distribution, to a certain data values can be achieved by comparing the superimposed probability density function (pdf) with the histogram of the data as can be seen in Figure 3.3 with 68% of the population is constrained between its limits (μ-σ, μ+σ) and the probability density function curve. For bivariate distributions the area that includes 68% of the data (plus and minus one standard deviation) is expected to be the projection of these two areas which is an elliptic shape centred at (μ_x, μ_y) with radii of σ_x and σ_x as shown in Figure 3.4. It can be seen from Figure 3.4, however, that the ellipse does not encompass the 68% of the data coloured in red dots from both variables due to the extreme points on at least one variable which can be explained from the distributions in Figure 3.5. A rectangular shape with its length being μ+σ should be used to cover this area for 68% of the data coloured in red dots from both variables. The green dots inside the rectangle are these instances that lie within the range (μ-σ ,μ+σ) for only one variable but not both. Similarly, a rectangular shape of μ+2σ length will include 95% of the data and likewise for the 98% of the population a square of μ+3σ length

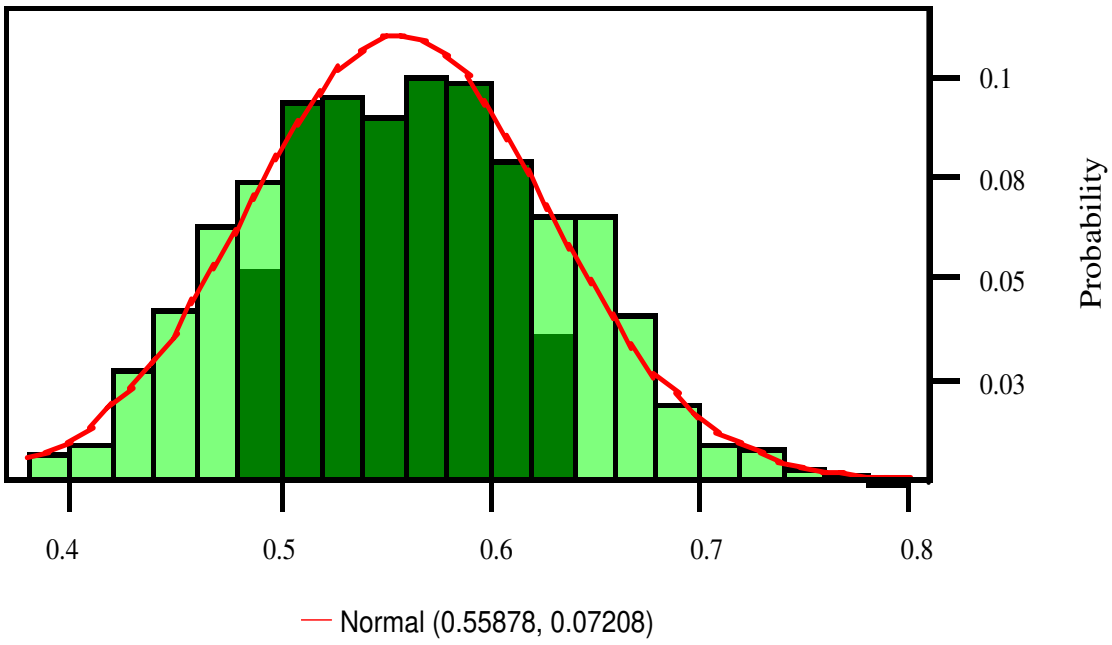

Figure 3.3: Fitting normal distribution to a data with 68% of population is highlighted.

is needed. In the multivariate distributions with three or more variables the square shape evolves into a hyper-cube which include all such red datum points instead of the hyper-ellipse which fails to surround all these points as illuatrated in Figure 3.6.

3.3 Minimum Message Length (MML) Technique in Mixture Modelling Method

The first application of mixture modelling method using the MML technique was suggested by Wallace and Boulton in 1968 and based on this proposal a classification program called Snob was written [53]. The program was successfully used to classify groups of six species of fur seals. Since then, the program has been extended and utilised in different areas, such as psychological science, health science, bioinformatics, protein and image classification [54]. Mixture Modelling Method using the MML technique has also been applied to other real world problems such as human

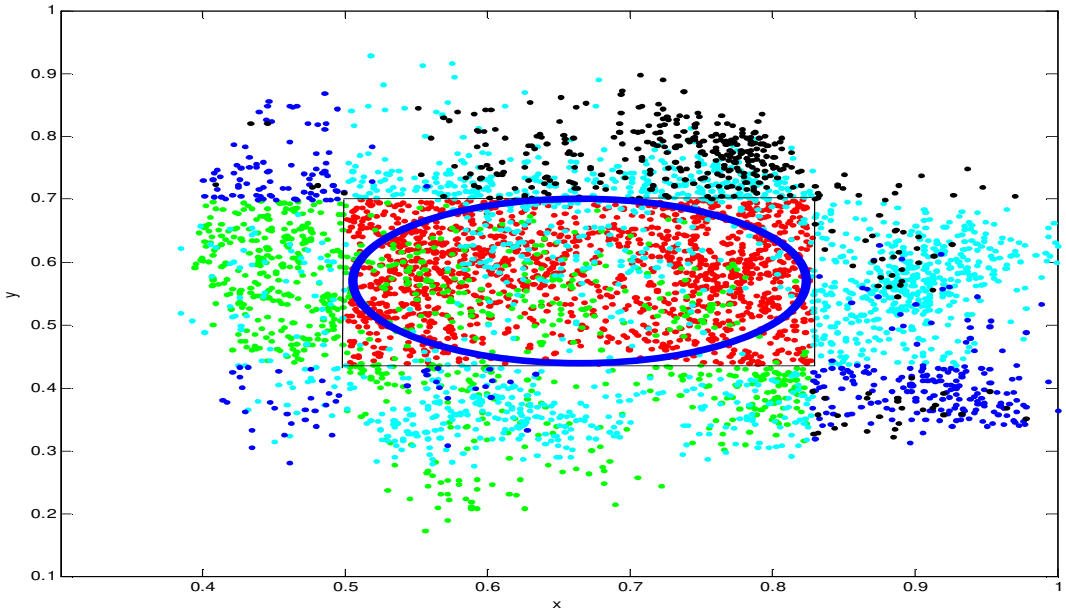

Figure 3.4: Two variables x and y with the area of 68% population in red colour represents the intersection of the two single distributions.

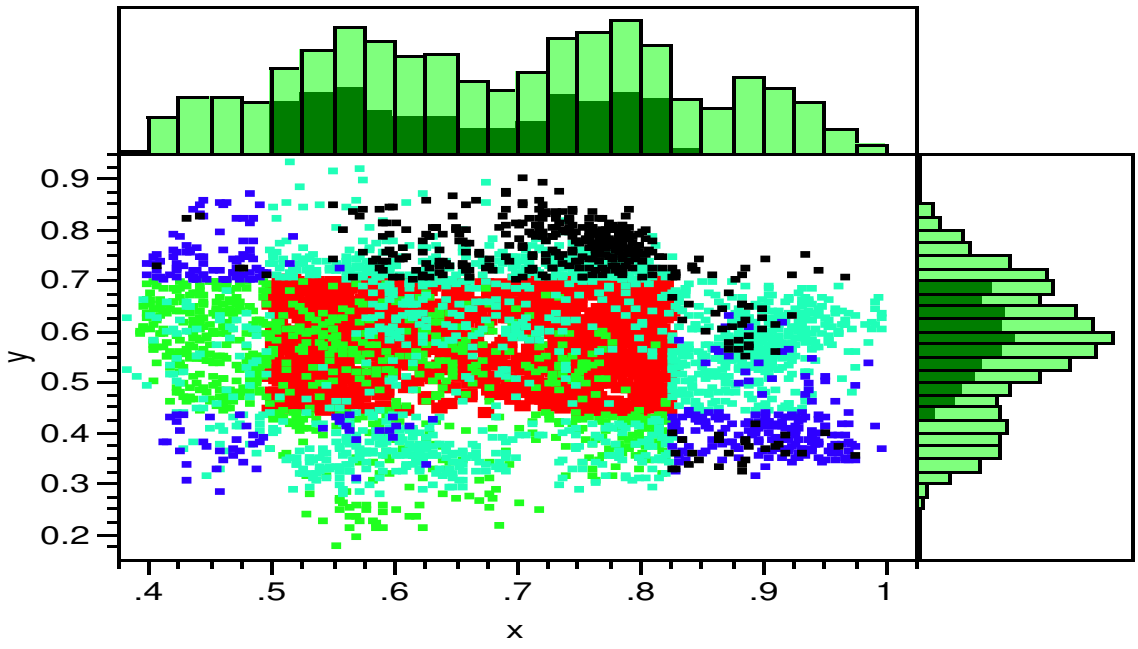

Figure 3.5: The area of one standard deviation (68% of population) generate square shape from bivariate distribution.

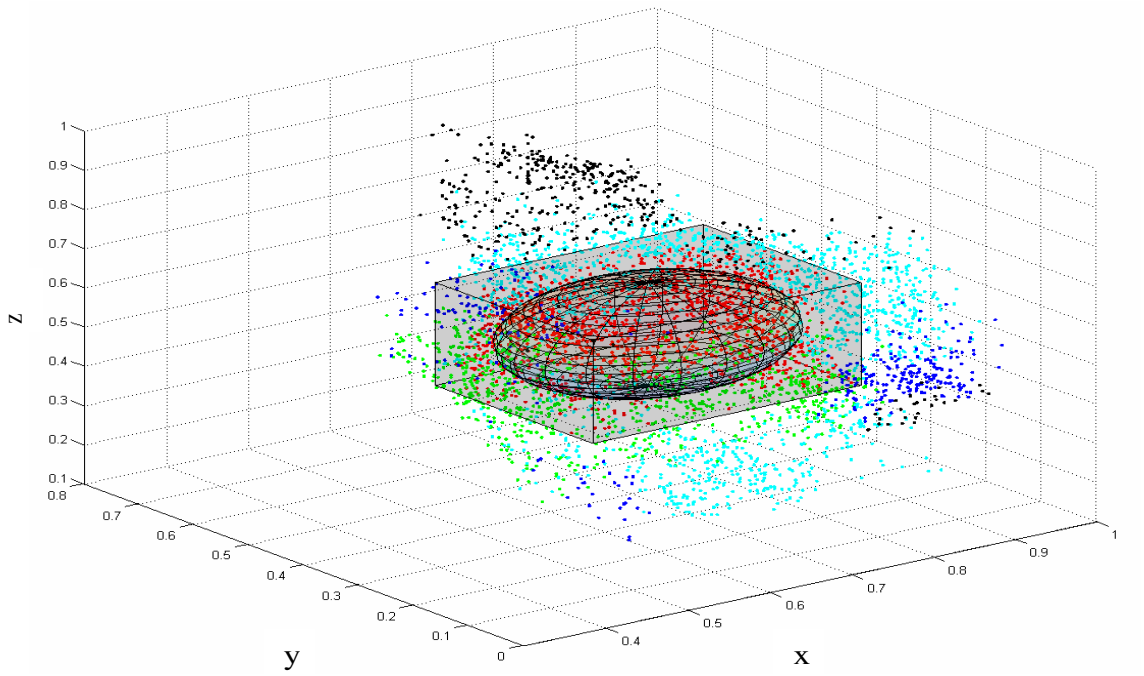

Figure 3.6: The hyper-cube shape, unlike the ellipsoide can cover the one standard devation area.

behaviour recognition and the diagnosis of complex issues in industrial furnace control [55]. In psychological science, a study was done in Melbourne among 1500 families to examine the intensity of grief, psychological morbidity, and social adaptation 13 months after the death of a parent. Using Mixture Modelling Method based on MML technique, five unique patterns were recognised from that study [56]. In health science MML was utilised in an attempt to identify plausible relationships between DNA strings and Autism, and specifically to establish if this was a single or multiple species relationship [57]. Human behaviour recognition is another area of applying Mixture Modelling Method using MML technique where the similarity between several human behaviours has been measured and represented as real numbers with the help of the Kullback Leibler distance and Smith Waterman Sequence alignment technique [58].

Mixture Modelling Method using the MML technique has also been used to segment human postures based on Euler angles and to depict motion phases by classification of each frame of motion as a particular pose given by a set of Euler angles [59].

Another application of Mixture Modelling Method using MML technique was to detect the serial correlation between the clusters in collections of known proteins by evaluating the hidden Markov model [60]. As far as the author is aware, the Mixture Modelling Method using MML technique has never been applied to power quality classification from harmonic monitoring data.

3.3.1 Minimum Message Length

The Minimum Message Length inductive inference, as the name implies, is based on evaluating models according to their ability to compress a message containing data. Compression methods generally attain high densities by formulating efficient models of the data to be encoded. The encoded message consists of two parts. The first of these describes the model and the second describes the observed data given that model. The model parameters and the data values are first encoded using a probability density function (pdf) over the data range and assume a constant accuracy of measurements (Aom) within this range. The total encoded message length (two parts) for different models is then calculated and the best model (shortest total message length) is selected. The MML expression is given as:

$$L(D, K) = L(K) + L(D|K) \tag{3.7}$$

where:

D : data set

K : mixture of clusters in model

$L(K)$: the message length of model K

$L(D/K)$: the message length of the data given the model K

$L(D, K)$: the total message length.

Initially given a data set D, the range of measurement and the accuracy of measurement for the data set are assumed to be available. The message length of a mixture of clusters each having Gaussian distributions with its own mean (μ) and

variance (σ) can be calculated as follows [61].

$$L(K) = \log_2 \frac{range_\mu}{AOPV_\mu} + \log_2 \frac{range_\sigma}{AOPV_\sigma} \tag{3.8}$$

where:

$range_\mu$: range of possible μ values

$range_\sigma$: range of possible σ values

$AOPV_\mu$: accuracy of the parameter value of μ

$$AOPV_\mu = \overline{s}\sqrt{\frac{12}{N}} \tag{3.9}$$

where:

$\overline{s}$: unbiased sample standard deviation

N: number of data samples

$$\overline{s} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{3.10}$$

$\overline{x}$: the sample mean

x_i : data points

$AOPV_\sigma$: accuracy of the parameter value of σ.

$$AOPV_\sigma = \overline{s}\sqrt{\frac{6}{N}} \tag{3.11}$$

The message length of the data using Gaussian distribution model can be calculated from the following Equation [61]:

$$L(D/K) = N \log_2 \frac{\overline{s}\sqrt{2\pi}}{Aom} + N\frac{s^2 + \frac{\overline{s}^2}{N}}{2\overline{s}^2}\log_2(e) \tag{3.12}$$

where:

Aom: accuracy of measurement

s: sample standard deviation

$$s = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{3.13}$$

Data segmentation by MML

Given a data set D and a given accuracy of measurement, Aom, the assumed statistical distribution is initially chosen as a Gaussian distribution. Starting from having all the data in one cluster (K=1) with a sample mean $\overline{x}$ and standard deviation s the parameters μ , σ and π (mean, variance and abundance) of this model can be estimated using the Expectation Maximisation algorithm (EM) to fit the mixture of Gaussian distribution model [40]. The abundance value π, for each cluster represents the proportion of data that is contained in the cluster in relation to the total data set. For a single cluster, the abundance value will be 100%. The abundance value can provide an indication of the importance of each of the clusters. Small abundance values may mean the cluster represents data of anomalous or rare occurrences and this may point out instances when the system needs to be observed more carefully. Once μ and σ are obtained, $range_\mu$ and $range_\sigma$ can be estimated, and $AOPV_\mu$ and $AOPV_\sigma$ can be calculated from Equations (3.9) and (3.11). The total message length L (D, K) can then be calculated using Equations (3.7), (3.8) and (3.12). The single cluster may subsequently be divided into a mixture of two clusters having the chosen Gaussian distributions (K = 2), each with its own sample mean and standard deviation. EM is then used to optimise the parameters μ , σ and π (mean, variance and abundance) of each of the new clusters. The total message length of the two clusters is recalculated and compared with the message length of the one cluster. If the total message length of the two clusters is smaller than the message length of one cluster, the splitting is assumed to be successful. However if the message length of the two clusters is higher than or equal to the message length of the one cluster, the single cluster is retained and the splitting process is repeated until a smaller message length

is obtained. In the program, an optimisation algorithm has been developed to find the best two clusters that yield the largest reduction of message length. The next step is to divide one of these clusters into two (K=3), and the above process is then repeated.

By itself, the splitting method has been found to be deficient to find the minimum message length in that the minimum message length is often not found. To overcome this problem other tactics are used in Snob (one of the clustering programs used in this thesis) such as, merging, reclassifying and swapping [51].

A conceptual flow chart of the Mixture Modelling Method using MML technique clustering algorithm is given in Figure 3.7.

An Illustrative Example

An example of how the Mixture Modelling Method using MML technique works, can be illustrated by applying the method to a small data set as shown in Figure 3.8, in which a set of 30 points is generated randomly from three normal distributions of 10 points each. The data for the 30 points is given in Table 3.2. Assuming the use of a model having a mixture of Gaussian distributions and an Aom of 1, for all data values, the range of is initially chosen as 0 - 25 to cover the whole range of data and for σ not to exceed $\mu/4$, the range of σ is chosen to be (0 - 5).

The segmentation process can be described in the following steps:

- Step 1

  Considering the whole data set as one cluster and calculating the message length of this cluster from Equations (3.7 - 3.12) yields 123.6 bit.

  m = 5.48, s = 3.68, $\overline{s}$ = 3.74 and N = 30.

  $AOPV_\mu = 3.74\sqrt{\frac{12}{30}} = 2.36$

  $AOPV_\sigma = 3.74\sqrt{\frac{6}{29}} = 1.7$

  Message Length = $\log_2\frac{25}{2.36}$ +$\log_2\frac{5}{1.7}$+30*$\log_2\frac{3.74\sqrt{2\pi}}{1.0}$+30*$\frac{3.68^2+\frac{3.74^2}{30}}{2*3.74^2}\log_2(e)$

  = 3.42 + 1.58 + 96.86 + 21.67

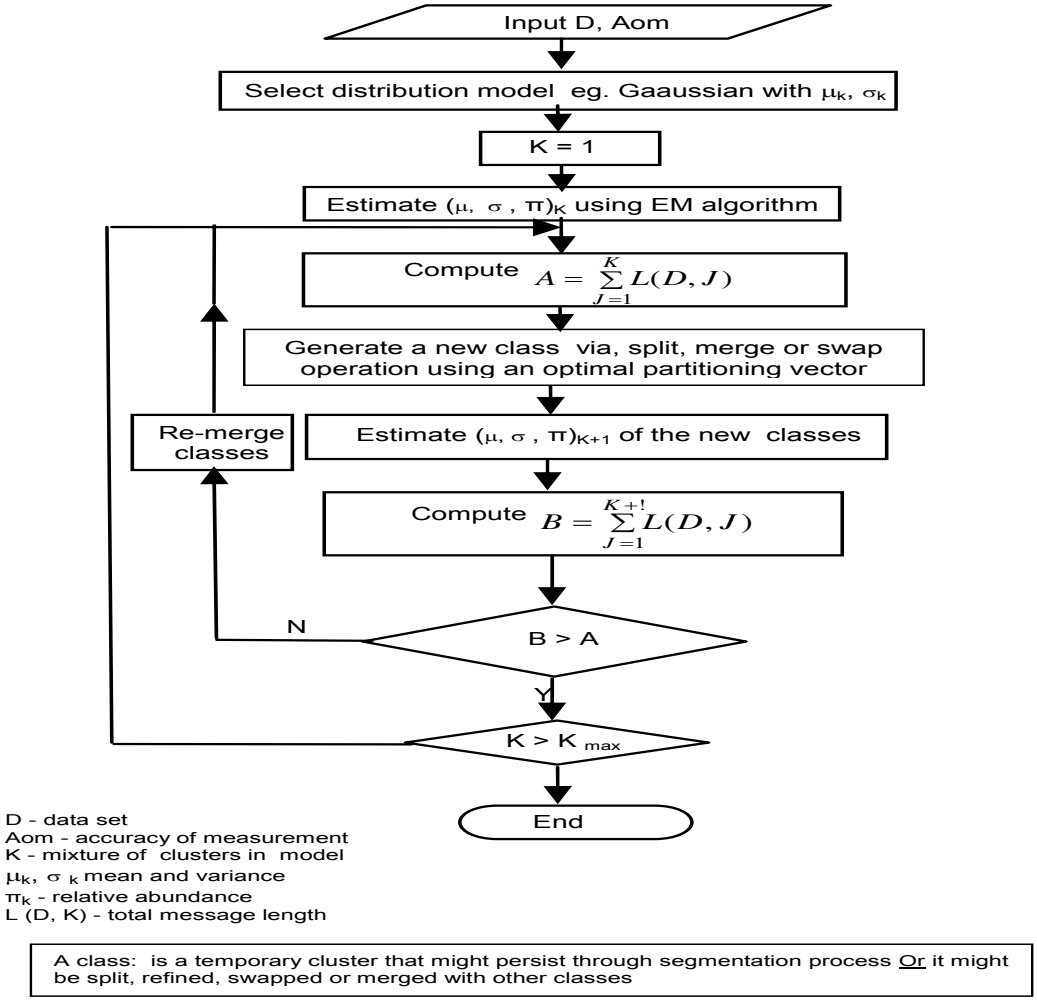

Figure 3.7: Conceptual flow chart of clustering algorithm of Mixture Modelling Method using MML technique.

$= 5 + 118.53 = 123.53$ Bits

- Step 2

  By splitting the data into two clusters (N1=16, N2=14) and using optimal partitions [62] the total message length is 104.7 bit.

  Cluster 1: m = 4.691, s = 1.58, $\overline{s} = 1.71$ and N = 16.

  $AOPV_{\mu} = 1.71\sqrt{\frac{12}{16}} = 1.48$

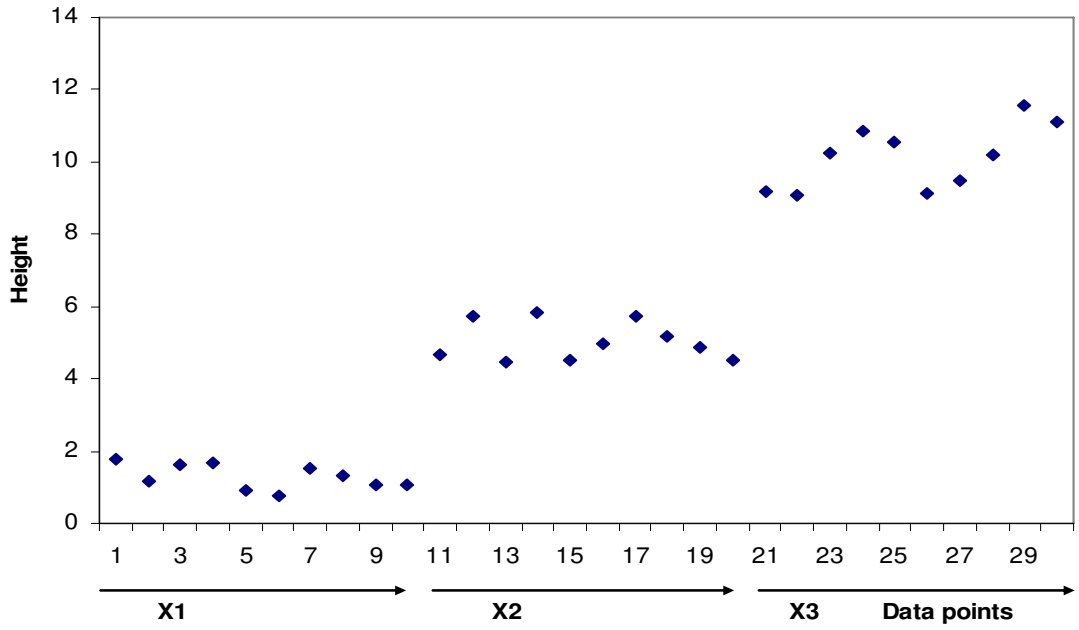

Figure 3.8: Three cluster (30 data point) generated randomly from X1 (cluster1), X2 (cluster2) and X3 (cluster3).

$AOPV_\sigma = 1.71\sqrt{\frac{6}{15}} = 1.08$

Message Length = $\log_2\frac{25}{1.48}$ +$\log_2\frac{5}{1.08}$+16*$\log_2\frac{1.71\sqrt{2\pi}}{1.0}$+16*$\frac{1.58^2+\frac{1.71^2}{16}}{2*1.71^2}\log_2(e)$

= 4.1 + 2.2 + 33.6 + 10.6

= 6.3 + 44.2 = 50.5 Bits

Cluster 2: m = 6.01, s = 2.17, $\overline{s}$ = 2.25 and N = 14.

$AOPV_\mu = 2.25\sqrt{\frac{12}{14}} = 2.08$

$AOPV_\sigma = 2.25\sqrt{\frac{6}{13}} = 1.53$

Message Length = $\log_2\frac{25}{2.08}$ +$\log_2\frac{5}{1.53}$+14*$\log_2\frac{2.25\sqrt{2\pi}}{1.0}$+14*$\frac{2.17^2+\frac{2.25^2}{14}}{2*2.25^2}\log_2(e)$

= 3.6 + 1.7 + 34.9 + 10.1

= 5.3 + 45.0 = 50.3 Bits

Table 3.2: Data points shown in Figure 3.8.

N	$X(1)$	$X(2)$	$X(3)$
1	1.8	4.66	9.17
2	1.19	5.74	9.06
3	1.61	4.45	10.26
4	1.65	5.83	10.87
5	0.89	4.49	10.54
6	0.76	4.95	9.13
7	1.5	5.72	9.48
8	1.31	5.16	10.18
9	1.05	4.87	11.56
10	1.04	4.5	11.09

- Step 3

  The segmentation process is continued to three clusters (N1=N2=N3=10) by splitting either one of two clusters and transferring between the existing cluster, the total message length is found to be 108.85 bit which is greater than the message length of two clusters (104.7 bit).

  Cluster 1: m = 2.285, s = 2.953, $\overline{s}$ = 3.112 and N = 10.

  $AOPV_{\mu} = 3.112\sqrt{\frac{12}{10}} = 3.41$

  $AOPV_{\sigma} = 3.112\sqrt{\frac{6}{9}} = 2.54$

  Message Length = $\log_2\frac{25}{3.41}$ +$\log_2\frac{5}{2.54}$+10*$\log_2\frac{3.112\sqrt{2\pi}}{1.0}$+10*$\frac{2.954^2+\frac{3.112^2}{10}}{2*3.112^2}\log_2(e)$

  = 2.87 + 0.98 + 29.63 + 7.21

  = 3.85+ 36.84 = 40.69 Bits

  Cluster 2: m = 4.691, s = 1.312, $\overline{s}$ = 1.383 and N = 10.

  $AOPV_{\mu} = 1.383\sqrt{\frac{12}{10}} = 1.52$

  $AOPV_{\sigma} = 1.383\sqrt{\frac{6}{9}} = 1.13$

  Message Length = $\log_2\frac{25}{1.52}$ +$\log_2\frac{5}{1.13}$+10*$\log_2\frac{1.383\sqrt{2\pi}}{1.0}$+10*$\frac{1.312^2+\frac{1.383^2}{10}}{2*1.383^2}\log_2(e)$

  = 4.04 + 2.15 + 22.61 + 7.21

  = 6.19+ 29.82 = 36.01 Bits

Cluster 3: m = 9.475, s = 1.835, $\overline{s}$ = 1.934 and N=10.

$AOPV_{\mu} = 1.934\sqrt{\frac{12}{10}} = 2.21$

$AOPV_{\sigma} = 1.934\sqrt{\frac{6}{9}} = 1.58$

Message Length = $\log_2\frac{25}{2.21}$ +$\log_2\frac{5}{1.58}$+10*$\log_2\frac{1.934\sqrt{2\pi}}{1.0}$+10*$\frac{1.835^2+\frac{1.934^2}{10}}{2*1.934^2}\log_2(e)$

= 3.56 + 1.66 + 22.77 + 7.23

= 5.22+ 30 = 35.22 Bits

Total message length = 40.69 + 36.01 + 35.22 = 111.92 Bit

- Step 4

  The previous generated clusters are re-merged and step 3 is repeated until the smallest message length is found which represents the original three generated clusters in Figure 3.8. The total message length for these three clusters is calculated to be 57.31 bits. The result of the above clustering steps is shown in Table 3.3. It can be observed that, as the number of clusters is increased from 1 to 3 clusters, the model message length is increased accordingly from 5 Bits to 23.96 Bits (10.1 + 6.41+7.45) to allow for the additional description of the new clusters. On the other hand, the data length is significantly reduced from 118.5 Bits to 33.35 Bits (5.42 + 10.1 + 17.83), as the new model was able to compress the message containing the data, resulting in less total message length i.e. 57.31 bits compared to 123.6 bits for the original single cluster.

Mixture Modelling software ACPro

ACPro has been chosen as the MML software used in this thesis. ACPro is an enhanced version of Autoclass, a data mining software tool [2]. It is an application of the Bayesian theory and minimum message length of information theory. ACPro is used in clustering data into similar segmentations. Predictions can also be made by

Table 3.3: Segmentation process of data points in Table 3.2.

Step	N	Model L(K) (Bits)	Data L(D/K) (Bits)	L(D,K) /cluster L(D,K) (Bits)	L(D,K) total L(D,K) (Bits)
1	30	5.00	118.50	123.53	123.53
2	16	6.30	44.20	50.50	
	14	5.3	45.00	50.30	100.80
3	10	3.85	36.84	40.69	
	10	6.19	29.82	36.01	
	10	5.22	29.99	35.22	111.92
4	10	10.10	5.42	15.52	
	10	6.41	10.10	16.51	
	10	7.45	17.83	25.28	57.31

ACPro based on the mixture model of clusters it developed through previous training phase. ACPro has been used in many applications in diverse range of areas such as telecommunication data, market segmentation, DNA Intron data [63] geology [64], image classification and spectral analysis of rock samples [65] [66]. ACPro has several advantages such as [2]:

1. can estimate the number of clusters (or classes) automatically,
2. can use mixed discrete and real valued data,
3. can handle missing values,
4. can have linear processing time with respect to the amount of the data,
5. can have probabilistic class membership,
6. can allow correlation between attributes within a class,
7. can generate extensive reports describing the classes found.

3.4 Comparison between Mixture Modelling Method using MML technique and other clustering and feature extraction algorithms

Mixture Modelling using the MML technique, unlike several other clustering techniques, uses encoded data and probability measures to assign each class to its object. These make it advantageous over other classification techniques based on distance measures such as K-means and Fuzzy C-means and feature extraction methods used in signal processing, such as Fourier or Wavelet Transforms. Comparisons between the mixture modelling using MML and these alternative techniques are presented in the following sub-sections.

3.4.1 Comparison between the Mixture Modelling using MML technique and traditional feature extraction methods based on signal processing techniques

The use of Mixture Modelling using the MML technique for harmonic classification has several advantages over traditional feature extraction method based on signal processing techniques in power quality data analysis. One advantage of applying the Mixture Modelling Method using MML technique in harmonic monitoring data is that it does not require the full waveforms of the studied signals to do the classification, unlike signal processing techniques, such as, Fourier transform (FT) or Wavelet transform (WT) which first require the complete waveform in order to perform the transformation to the relevant domain, and only then can the classification process be initiated. Some signal processing techniques are time consuming in analysing power quality monitoring data, for example Wavelet transform does not provide the exact features of the signal and hence more computation is required to obtain the accurate features. However, Mixture Modelling using MML technique is able to find the simpler parameters (mean, variance and abundance) of each of the generated clusters. In

harmonic data, for example, each cluster can represent a specific operating condition, such as peak load, off-peak load, capacitor switching operation [67].

3.4.2 Comparison between Mixture Modelling using MML with other distance base clustering methods.

In this section, the Mixture Modelling using MML is compared with other distance-based algorithms, such as K-Means and Fuzzy C-means.

MML Mixture Modelling versus K-Means

The Mixture Modelling using the MML technique used here is often known as intrinsic classification [44], [68]. In this form Mixture Modelling using the MML technique typically performs better than those based on an a priori distance measures, such as the nearest neighbour algorithm, like K-means [69]. Mixture Modelling using the MML is actually a generalized form of K-means because it can use other types of distributions beside Gaussian distributions, and with variously shaped clusters. K-means is also known to be unable to obtain acceptable clusters when the clusters have different sizes, shapes, or co-variances [40]. For comparison purposes, three randomly generated clusters are shown in Figure 3.9.

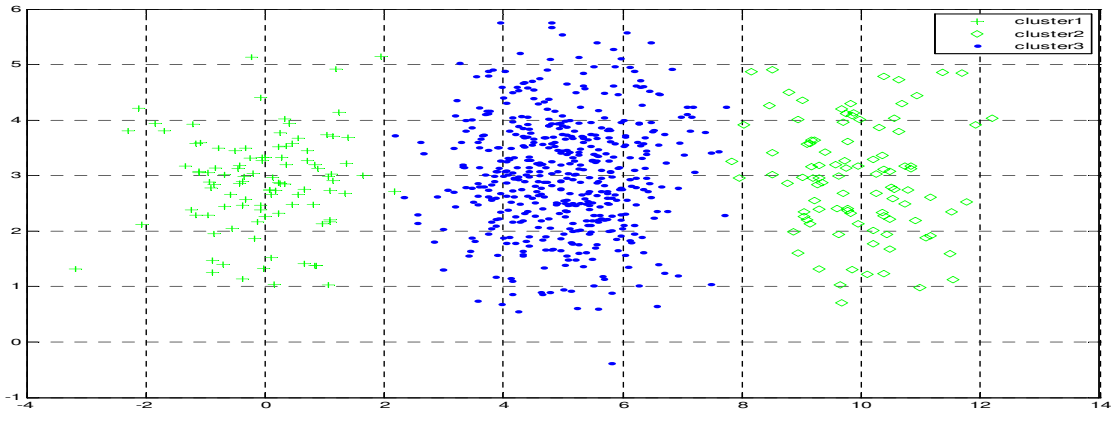

Figure 3.9: Three randomly generated Clusters.

Mixture Modelling using MML can classify correctly these three randomly generated clusters as shown in Figure 3.10, whereas K-means fails to detect the right cluster as shown in Figure 3.11, where the centre cluster is larger and denser than other two. Further, in many clustering algorithms based on distance measures, such

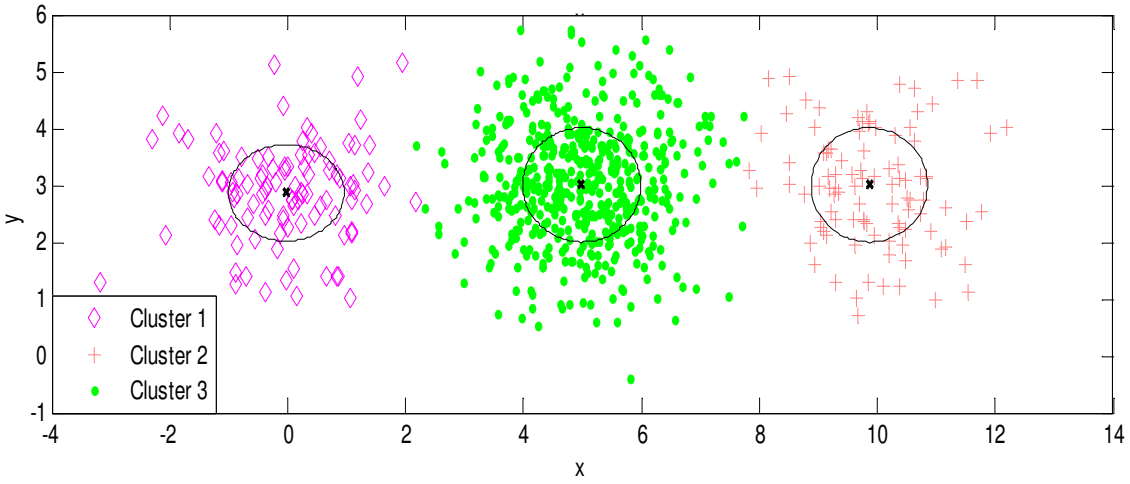

Figure 3.10: Correctly clustering of the clusters shown in Figure 3.9 using MML.

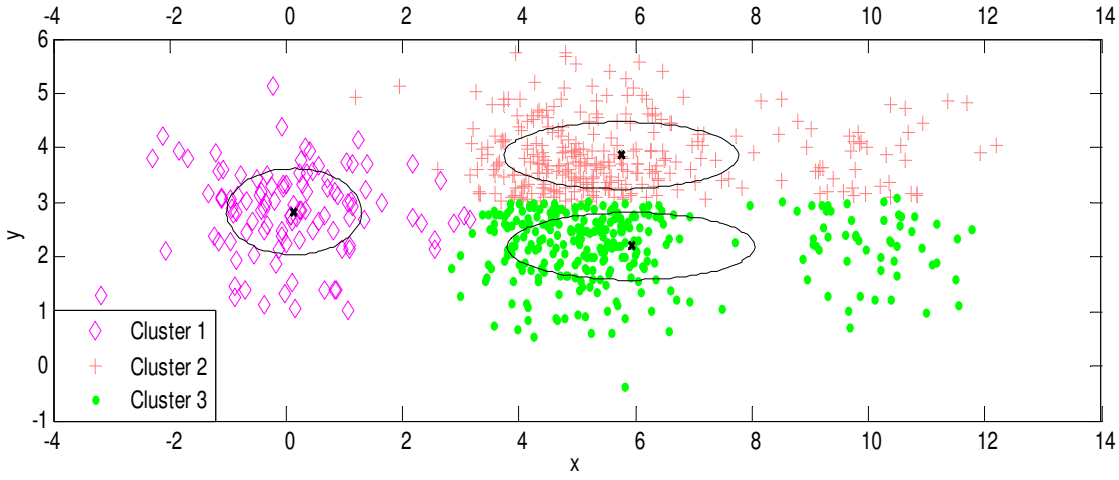

Figure 3.11: False Clustering of the clusters shown in Figure 3.9 using K-means.

as K-means and Fuzzy C-means, the value of the expected numbers of clusters needs to be specified as an input parameter to the algorithm. This means that one needs

to know the number of clusters in advance, or to carry out a trial and error method to determine the optimum number of clusters.

In the Mixture Modelling using MML approach however, the number of the clusters can be automatically generated in iterative steps, representing the shortest message length as described in Section 3.3.1.

MML Mixture Modelling versus Fuzzy C-means

The Fuzzy C-means algorithm, which also assigns each data point to all or some of the clusters based on distance measures, is the same as K-means except that there is a membership value between 0 and 1 for each data point belonging to that cluster. The closer a data point is to the centre of the cluster, the more certain it is that this data point belongs to that cluster by having a high membership value: equal to, or close to 1.0. Similarly, the further a data point is from the centre of the cluster the less probable it is that this data point is a member of the cluster: its membership value is close to or equal to 0. Fuzzy C-means has also been applied to the randomly generated data shown in Figure 3.9 and the result is shown in Figure 3.12.

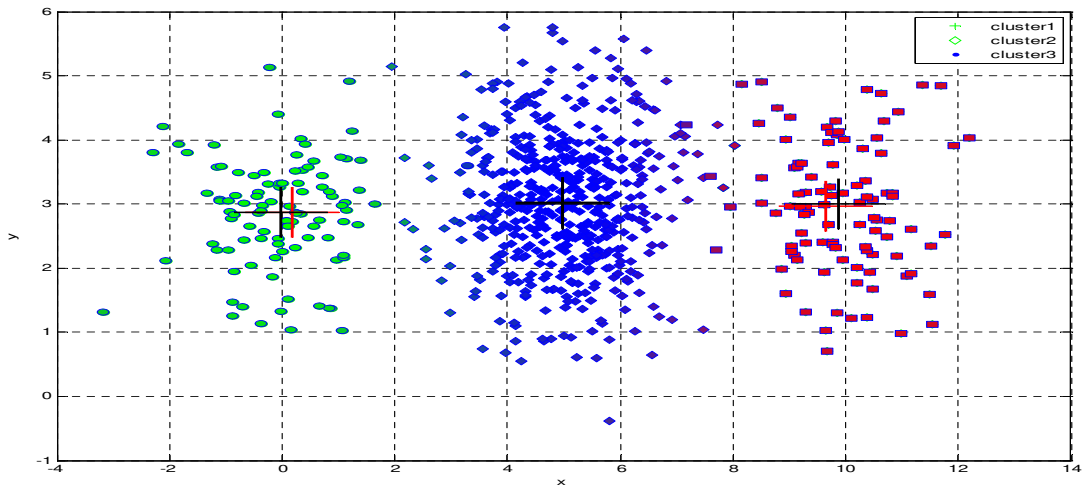

Figure 3.12: Centre displacements of clusters 1 and 2 of the clusters shown in Figure 3.9 using Fuzzy C-means.

From Figure 3.12, it can be observed that Fuzzy C-means clusters the data better than K-means, however the centres of clusters 1 and 2 were not accurately specified.The dislocation of these centres results from the effect of weak members which overlap with members generated from cluster 3 in the middle. Mixture modelling method using MML technique, initially 'offers' each data point to all generated clusters and then the clusters virtually compete to acquire their members using the probability values. Each data point is assigned to a cluster with the highest probability value.

3.5 Summary

In this chapter, the mixture modelling using MML techniques has been presented. Limitations and weakness of some clustering algorithms such as K-means and Fuzzy C-means have been addressed. The mixture model of MML is benchmarked with distance K-means and Fuzzy C-means. It is also compared with signal processing techniques such as FFT and WT. The successful application of MML in several disciplines, such as psychological science, health science, bioinformatics, human behaviour recognition and synergistic reactive control skills, has created the motivation to use MML in harmonic monitoring data to discover underlying natural groups in this data. This will allow utility engineers, as experts in the domain, to make ready use of the clustered data to quickly interpret these patterns, and in particular, to detect unusual power quality events.

Chapter 4

Optimal Number of Clusters

4.1 Introduction

One difficulty with the MML algorithm used in the mixture modelling method is the difficulty in establishing stopping criterion to secure optimum number of (mixture) clusters during the clustering process. During the investigation, it was discovered early that a method has to be found to determine the optimum number of clusters using the MML technique, since overestimating the number of clusters will produce a large number of clusters each of which may not necessarily represent truly unique operating conditions, whereas underestimation leads to only small number of clusters each of which may represent a combination of specific events.

In this chapter, a novel technique has been developed to overcome this difficulty using the trend of the exponential of message length difference between consecutive mixture models. The proposed method was tested using data from known number of clusters with randomly generated data points and with data from a simulation of a power system. The results from the tests confirm the effectiveness of the proposed method in finding the optimum number of clusters. The method was benchmarked against commonly used fitness function technique and found to produce in similar number of clusters using data of independent variables.

4.2 Determination of the optimal number of clusters

Determining the optimum number of clusters in the mixture modelling method using MML technique is essential since the result of clustering is adversely reliant on the number of the generated clusters. Overestimating the number of clusters will produce a large number of clusters, often complicating the result by artificially forcing a split of originally well formed clusters, whereas underestimation often leads to aggregations or merges of originally well segmented clusters and thus loss of information. In order to address this trade-off, a fitness function [70] has been utilised as a criterion to find optimal number of clusters. However, it is well known that such a technique assume that the attributes are independent and hence will have difficulty if the attributes are not independent variables, such as the data from a harmonic monitoring system. To overcome this, an intelligent method to determine the stopping criterion for the cluster generation in the mixture modelling method is required.

4.2.1 Effect of the number of clusters

To analyse the effect of choosing a variable number of clusters, five clusters each of 100 data points (D's) were randomly generated (D1, D2,..., D5), each with its own mean and standard deviation sorted in ascending order as shown in Table 4.1. The normal distributions of these clusters are superimposed on the data as shown in Figure 4.1.

Table 4.1: The parameters (μ and σ) of the five generated clusters.

Data set	D1	D2	D3	D4	D5
Mean (μ)	1.02	4.04	7.95	11.94	16.06
SD (σ)	0.28	0.52	0.87	1.12	1.44

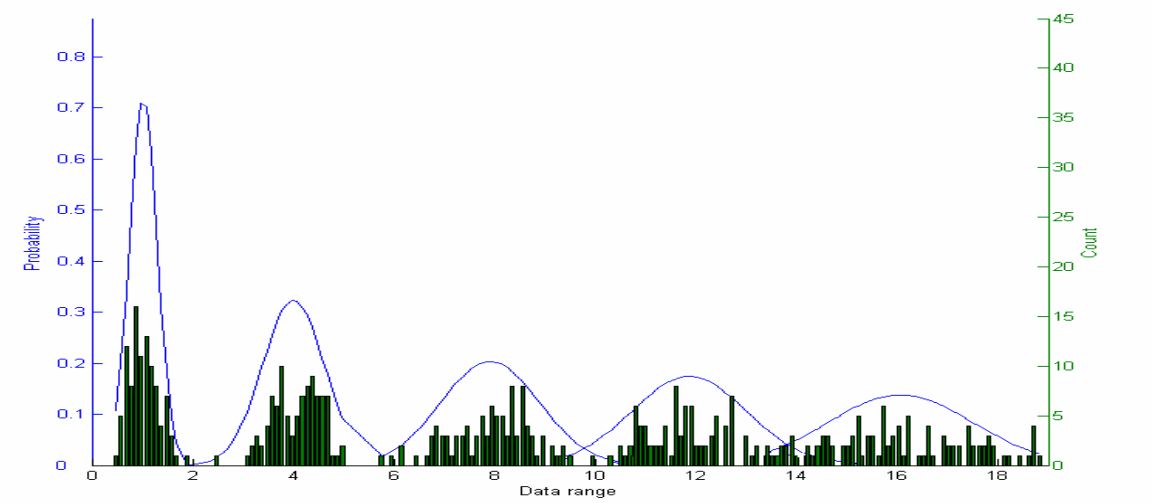

Figure 4.1: Five randomly generated clusters each with its own mean and standard deviation.

Initially two, four and five clusters were specified as input parameters to the MML data mining program. Subsequently, ACPro was allowed to determine the number of clusters itself resulting in seven clusters. The generated clusters in each case are shown in Figure 4.2(a-d).

Figures 4.2(a-b) show that underestimation of the number of clusters will result in having clusters with a combination of D's. Figure 4.2(a) shows that one of the clusters represents D1 and D2 and the other D3, D4 and D5. Figure 4.2(b) shows that D1, D2 and D3 are identified correctly, but D4 and D5 are identified as one cluster. Figure 4.2(d) illustrates that the overestimation generated by ACPro, was due to its inadequate stopping criterion, producing spurious clusters representing the data of higher variances. Figure 4.2(c) shows that ACPro correctly segments the data into the right five clusters given the correct input for the number of clusters. This identifies the need to have an optimal way of deciding the correct number of clusters from a given data set.

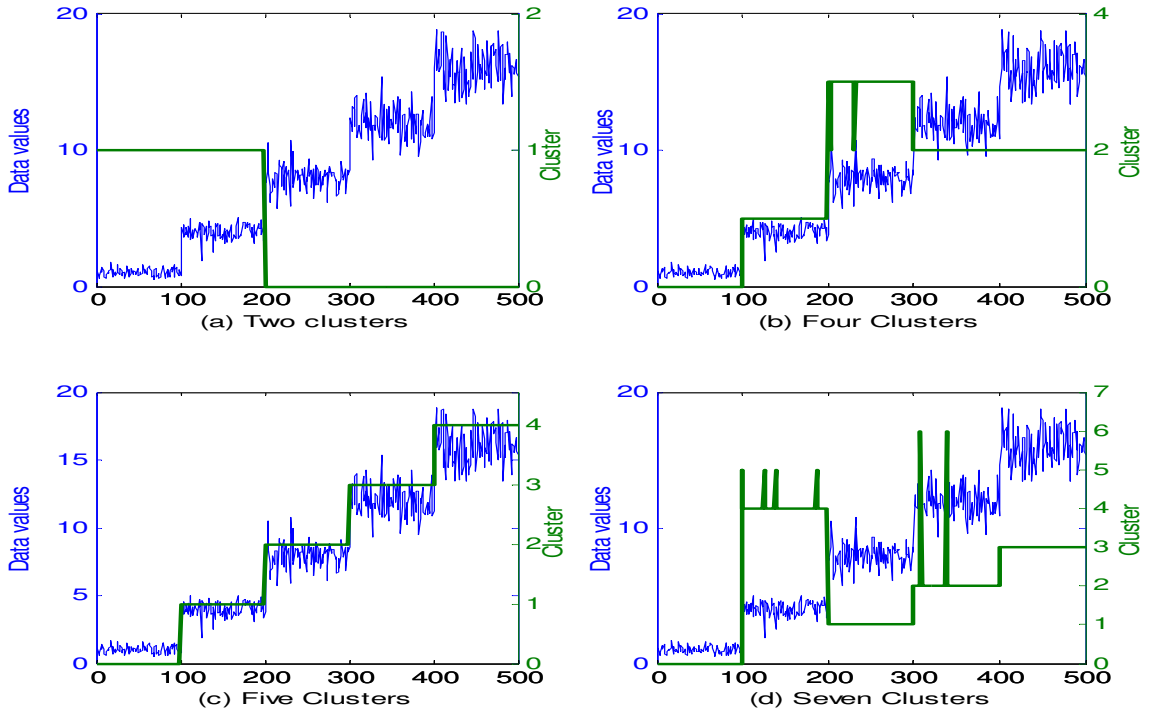

Figure 4.2: The clusters obtained superimposed on the randomly generated data.

4.2.2 Fitness function determination of the optimal number of clusters

From information theory, it is suggested that a fitness function [71] can be utilised as a criterion to determine the optimum number of clusters when the mixture modelling method is used for data fitting. The higher the fitness function value, the better the data fit. The fitness function tends to gain maximum information from data via maximum entropy. This maximum entropy is fulfilled if the data set can be modelled as a mixture of Gaussian distributions [70]. The theoretical maximum entropy H_{max} of any distribution can be calculated as follows:

$$H_{max}(C_i) = \frac{1}{2}\log(2\pi e)^n |cov(C_i)| \tag{4.1}$$

where:

C_i: a column vector containing the highest probabilities of each data point (P_i) belonging to cluster i.

cov: is the covariance matrix of C_i.
n: number of independent attributes.

The individual fitness function ef_i is dependent on the maximum entropy given in Equation 4.1 and can be calculated as follows:

$$ef_i = \frac{H(C_i)}{H_{max}(C_i)} \tag{4.2}$$

where:
$H(C_i)$ is the entropy of C_i.

$$H(C_i) = -\sum_i P_i \log_2 P_i \tag{4.3}$$

The total fitness function EF_T from ef_i can be calculated from the individual fitness function ef_i given in Equation 4.2 as:

$$EF_T = \sum_{i=1}^{k} \alpha_i |ef_i| \tag{4.4}$$

where:
k is the total number of clusters.
α is the abundance of the clusters in the whole data.

The higher the value of the total fitness function the better the data set can be modelled by a mixture of Gaussian distributions and hence the largest value of the fitness function EF_T corresponds to the optimum number of clusters in the data set.

When applied to the five randomly generated clusters, as discussed in Section 4.2.1, Figure 4.3 shows how the fitness function increases and reaches maximum when the total number of clusters is 5, suggesting that such a method is suitable to determine the optimum number of clusters. A recent study [72] shows that the entropy fitness

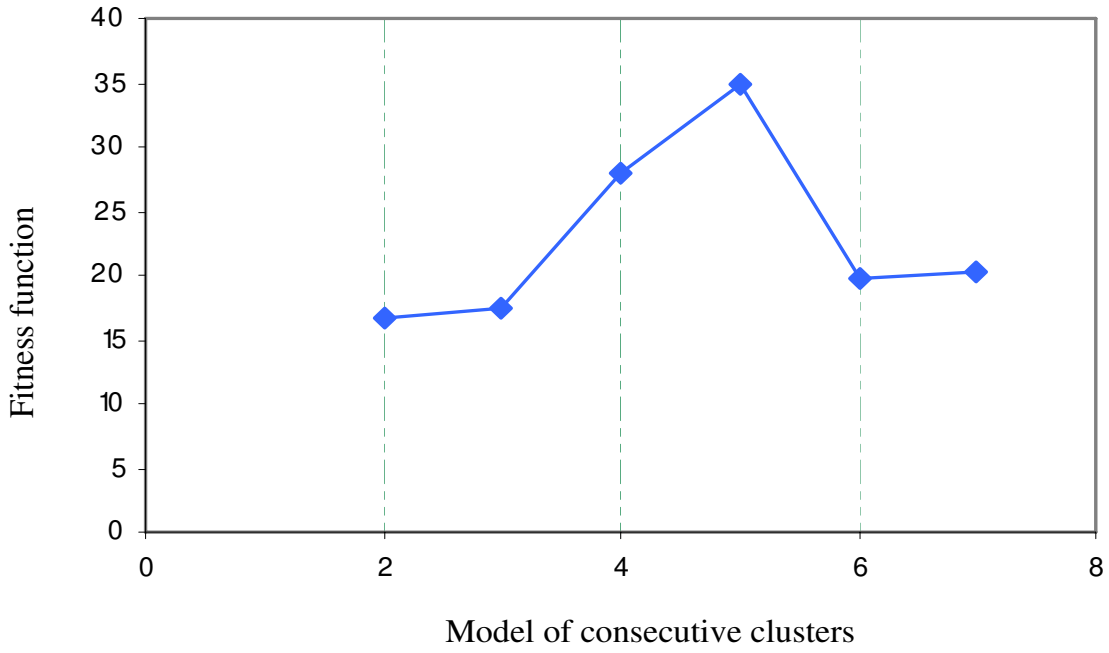

Figure 4.3: Fitness function showing five clusters in random data.

function can be used to determine an optimal number of clusters to correctly identify the anomalies in intrusion detection data. While the fitness function can be used to determine the optimum number of clusters, it has difficulties when faced with real harmonic data measured at several points in the network where the attributes at one point are correlated to the same attributes at the other part of network [73]. Since the purpose of this research work is to segment the harmonic data from the harmonic monitoring system, a different strategy is needed to overcome this problem.

4.2.3 Using Mixture Modelling based on MML to determine the optimum number of clusters

The MML states that the best theory or model K is the one that produces the shortest message length of that model and data D given that model. Minimizing this message length in an MML technique from information theory is equivalent to maximizing the posterior probability in Bayesian theory [2]. This posterior probability of Bayes'

theorem is given by:

$$Prob(D|K) = \frac{Prob(K) * L(D|K)}{Prob(D)} \tag{4.5}$$

From (3.7) and (4.5) and [61] yields:

$$L(D, K) = Prob(D|K) \tag{4.6}$$

Further, this difference can be emphasised by calculating the exponential of the change in message length for consecutive mixture models, which in essence represents the probability of the model correctness. If this value remains constant at around 1 for a series of consecutive mixture models then the first time it reaches this value can be considered to be the optimum number of clusters.

Verifying the exponential of message length difference method using randomly generated data

The proposed method which calculate the exponential of message length difference between consecutive mixture models is tested using the five randomly generated clusters described in Section 4.2.1. When the proposed method is applied to this data, the optimum number of cluster was found to be five when the first maximum value of the trend in the exponential of message length difference curve is reached and remains constant afterward at models of 6 and 7 clusters respectively as shown in Figure 4.4.

The result is the same as that obtained from the fitness function (see Figure 4.3) which, when applied to this data, has also found that five is the optimum number of cluster. This method is now tested in the next section using data from a simulation of a power system.

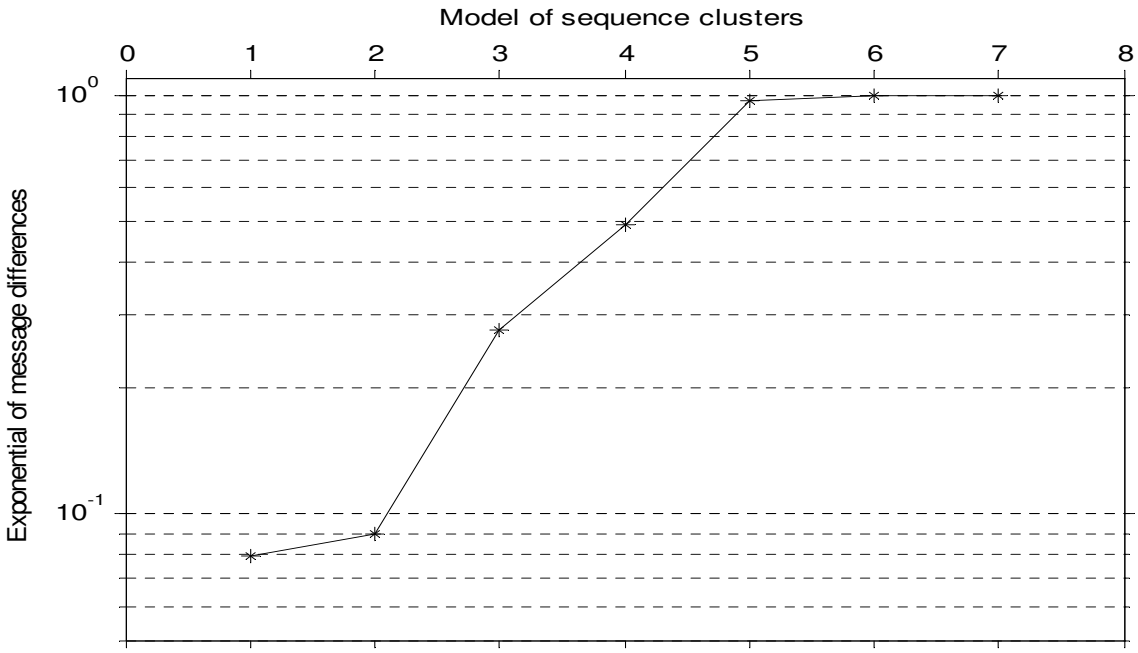

Figure 4.4: Exponential of message length difference identifying five clusters as the optimum number.

Verifying the exponential of message length difference method using simulation data

To further test the proposed method with another data type, a simulation of a simplified power system shown in Figure 4.5 is carried out using PSCAD®/EMTDC™. Three switches are used to represent 8 operating conditions of three resistive loads depending on which switch is turned ON or OFF. The switching operation, the times of switching and the cluster, generated by MML, in each case are shown in Table 4.2.

Figure 4.6 illustrates the rms voltage and current at phase 'a' at Bus 1. Using these two variables as the two input attributes for the MML algorithm, Figure 4.7 displays the trend in the exponential of the message length difference from a sequence of consecutively increasing mixture model sizes. In this case ten clusters were found to be the optimum number. The heuristic used here might be expressed as "select

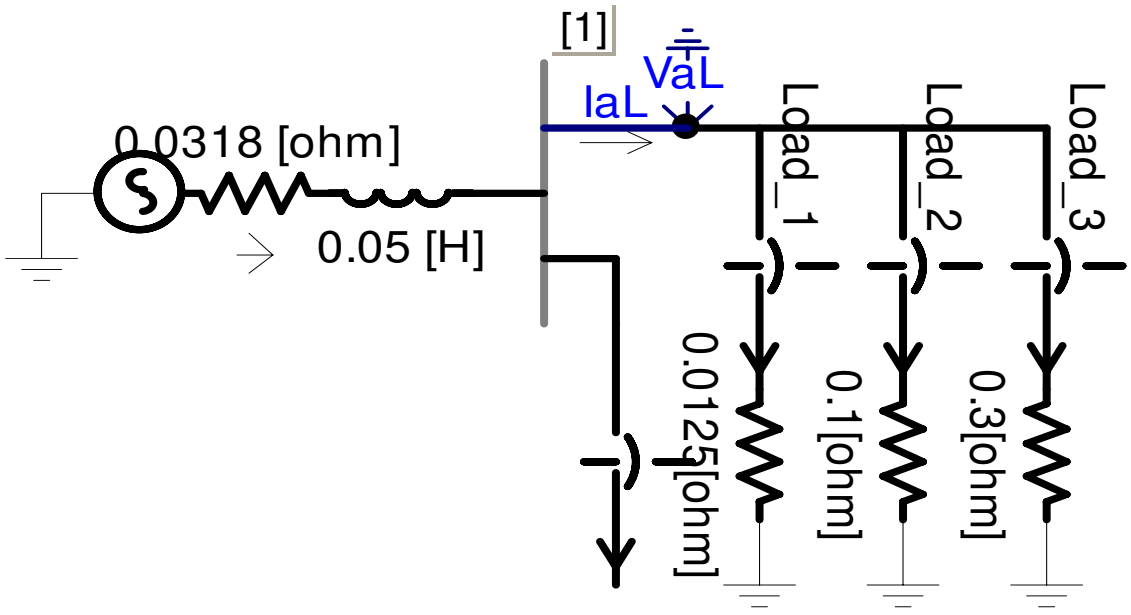

Figure 4.5: A single line diagram of a simplified power system model used in a PSCAD®/EMTDC™ Simulation.

Table 4.2: The load switching operation and timing.

Load1 on/off	Load2 on/off	Load3 on/off	Time on (Sec)	Time off (Sec)	Cluster no
0	0	0	0	0	6
0	0	1	1.25	2.50	5
0	1	0	2.50	3.75	7
0	1	1	3.75	5.00	1
1	0	0	5.00	6.25	0
1	0	1	6.25	7.50	2
1	1	0	7.50	8.75	3
1	1	1	8.75	10.00	4

the model size as the point where, the value of the exponential of the message length difference first, or most rapidly, approaches unity, and continues near unity for any further increase in size". The detailed model parameters obtained from the MML

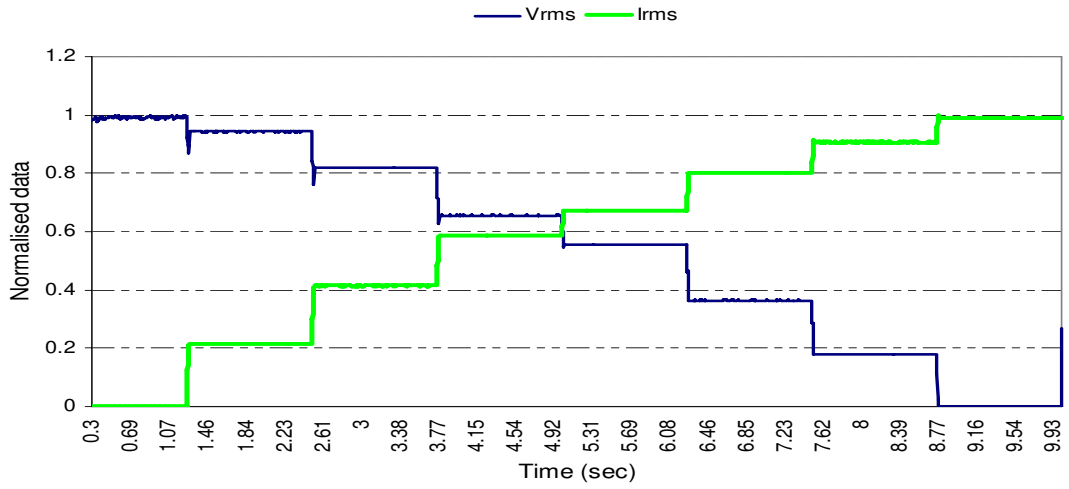

Figure 4.6: The rms values of voltage and current in phase 'a'.

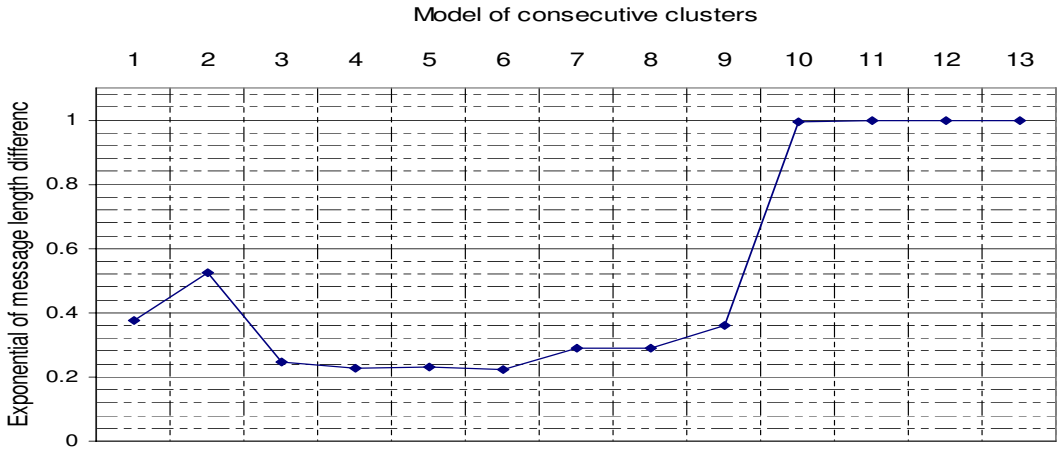

Figure 4.7: Exponential of the message length difference of consecutive clusters.

algorithm are given in Table 4.3 (mean, variance and abundance) for each of the 10 member clusters. In addition, the behaviour of this 10 cluster model (s0, s1, s2, ...,

s9) is also superimposed over the variations of the two input attributes in Figure 4.8.

Table 4.3: Ten generated clusters with different means and standard deviations.

Cluster	Abundance(π)	Va(μ)	Va(σ)	Ia(μ)	Ia(σ)
s0	0.127763	0.553912	0.01	0.671487	0.01
s1	0.127764	0.655373	0.01	0.587366	0.01
s2	0.127751	0.364667	0.01	0.801446	0.01
s3	0.12776	0.177792	0.01	0.906666	0.01
s4	0.127252	0.001109	0.01	0.991054	0.01
s5	0.126763	0.943114	0.01	0.215727	0.01
s6	0.098403	0.990902	0.01	0.000313	0.01
s7	0.126723	0.820193	0.01	0.414265	0.01
s8	0.004657	0.356497	0.188228	0.808623	0.142181
s9	0.005080	0.820948	0.077592	0.308243	0.13732

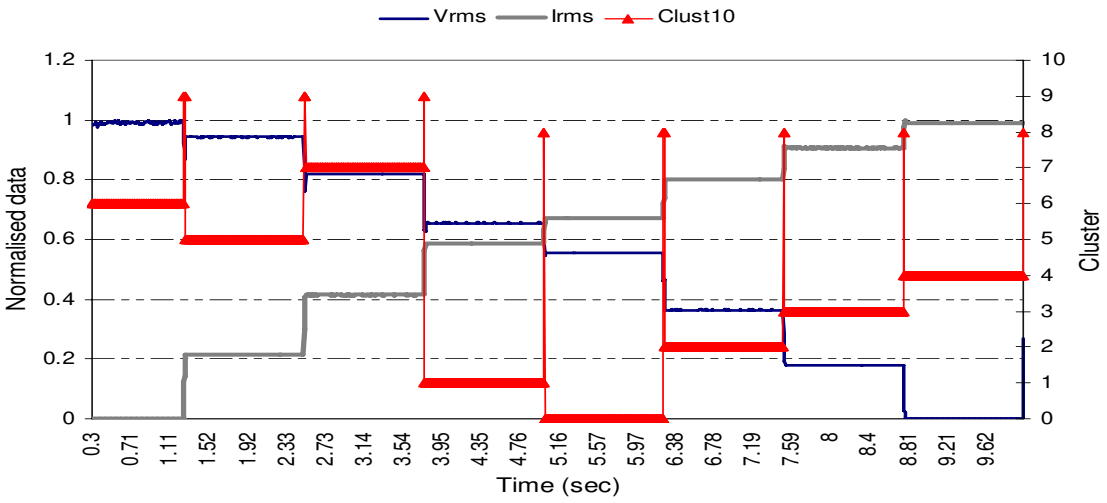

Figure 4.8: The ten generated clusters superimposed on simulation data.

This is a very interesting result, as only 8 clusters were expected, however because of the inductance in the source, transient events can be observed in Figure 4.8 at each switching point of the simulated loads, and the MML method has identified these transients as two separate clusters (s8 and s9) at the instant of switching at 1.25,

2.5 and 3.75 seconds - and s8 at the other switching times. Looking at Table 4.2, it can be observed that the s9 is associated with load 1 being OFF and s8 is due to load 1 being ON. Figure 4.8 shows that there is a distinct difference in the voltage and transients at these two different groups of switching times, while at the same time the similarity in each group of the transient events. From Figure 4.9, it can be seen that the abundances (π) of these two transient clusters (s8, s9) are the lowest (rare clusters) among other clusters and their variances (σ) are the highest (unstable clusters) which make them special clusters that need more investigation. Applying the fitness function described in Section 4.2.2 to the same two attributes, produces the same optimum number as shown in Figure 4.10. The highest fitness function is found at 10 clusters.

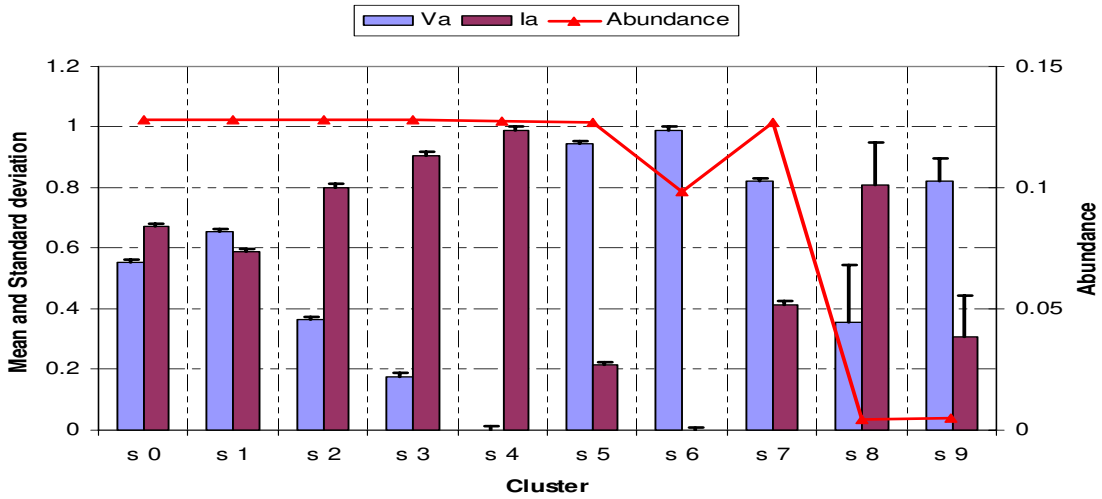

Figure 4.9: The clusters statistical parameters mean (μ), standard deviation (σ) and abundance (π).

4.3 Summary

A technique is proposed to find the optimum number of clusters when using the MML. This technique uses the trend of the exponential difference in message length between two consecutive mixture models to determine the optimal number of clusters.

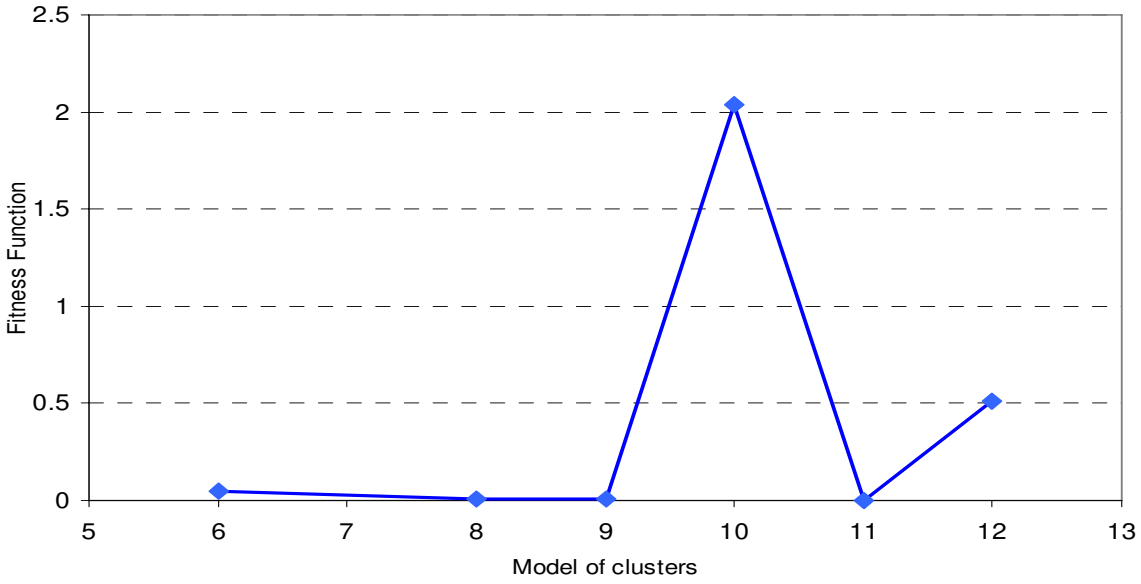

Figure 4.10: Fitness function [71] also identifying that 10 is the optimum number of clusters.

The optimal number of clusters from randomly generated data and from simulation data of power system were investigated. The technique was benchmarked against the fitness function technique and found to generate similar number of clusters using data of independent variables.

Chapter 5

Harmonic data collection and preparation for data mining techniques

5.1 Introduction

A harmonic monitoring program was conducted in Australia between August 1999 and December 2002 to measure the harmonic currents and voltages in a medium distribution system. Details of the method used for the monitoring, data captured as well as the selected data for analysis are presented. How the data has been prepared for use in the data mining application is introduced.

5.2 Harmonic monitoring program and System study

Between August 1999 and December 2002, a harmonic monitoring program was designed and implemented by a local electricity distributor in Australia in a medium voltage (MV) distribution system in Australia [74]. The data used in this thesis was obtained from a local utility distributor in Australia and the measurement from

this harmonic monitoring program. The monitoring involved simultaneous measurements of the three-phase harmonic current and voltage from residential, commercial, and industrial load sectors [5]. Simultaneous measurements of three-phase harmonic currents and voltages from these different load sectors enabled the effect on the net distribution system harmonic voltage and current to be determined. The selected zone substation is a typical 33/11 kV zone substation in a suburban area in Australia, which supplies ten 11kV radial feeders. The substation maximum demand was approximately 22 MVA (80% of capacity with N-1 transformer redundancy) and the short circuit level at the 11kV busbar was approximately 213 MVA [5]. Figure 5.1 illustrates the layout of the zone substation and feeder system for the harmonic monitoring program. As can be seen from Figure 5.1, seven monitors were installed: a monitor at each of the residential, commercial and industrial sites (site ID 5-7); a monitor at the sending end of the three individual feeders (site ID 2-4); and a monitor at the zone substation incoming supply (site ID 1). Sites 1-4 in Figure 5.1 are all within the substation at the sending end of the feeders and were identified as being of a predominant load type. Site 5 was along the feeder route, approximately 2km from the zone substation, and feeds a residential area. Site 6 supplies a shopping centre with a number of large supermarkets and many small shops. Site 7 supplies a factory manufacturing paper products such as paper towels, toilet paper and tissues.

5.2.1 Identification of load types from selected monitored sites

Load type plays an important factor when monitoring harmonic levels in distribution systems due to the diversity of equipment in each load sector. The three main load sectors measured are residential, commercial and industrial. Residential loads consist of lighting lamp and house appliances such as colour televisions, washing machines and air conditioners. Commercial loads are mostly lighting, office machines, and elevators whereas industrial loads are electric motors, power electronic apparatus and lighting. It has to be noted that the load type (it might be a mixture of two or three load types) should be specified prior to commencing harmonic monitoring program,

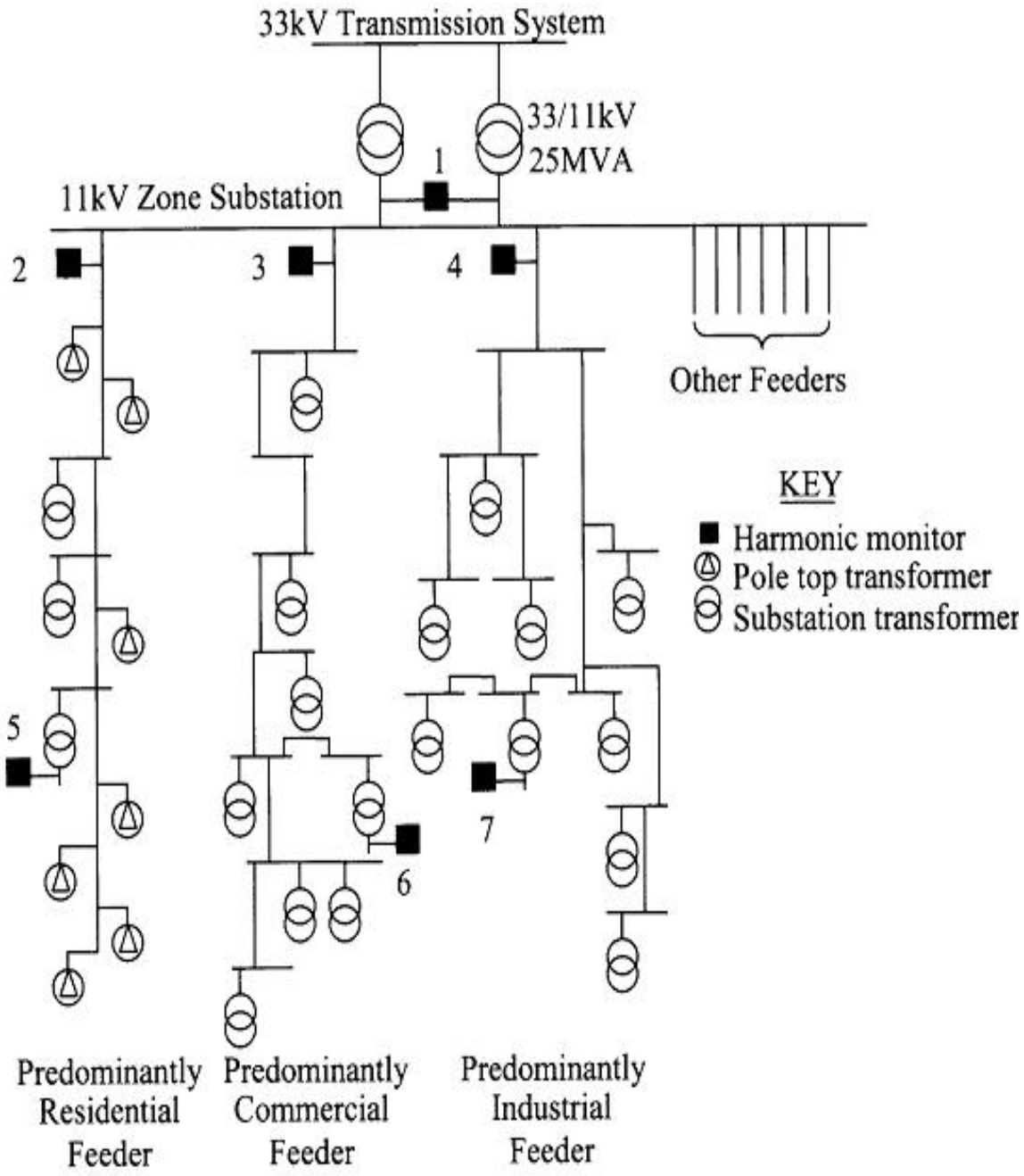

Figure 5.1: Single line diagram illustrating the zone distribution system.

as the variations of harmonic distortion are dissimilar. The load type percentages for the various monitored sites in Figure 5.1 are given in Table 5.1 below.

Table 5.1: Proportions of each MV/LV sites based on load types

Site ID	Mainly site type	Residential (%)	Commercial (%)	Industrial (%)
Site 2	Residential feeder	85	15	-
Site 3	Commercial feeder	10	90	-
Site 4	Industrial feeder	5	20	75
Site 5	Residential transformer	90	10	-
Site 6	Commercial transformer	-	90	10
Site 7	Industrial transformer	-	-	100

5.2.2 Harmonic monitoring equipment

The monitoring equipment utilised to measure the harmonic data used in this thesis are EDMI MK energy meters [75] as illustrated in Figure 5.2. The selection of this equipment was made based on its compliance with standards and cost. Further considerations were also made to this selection such as recording parameters availability, storage memory, sampling speed, data storage and transfer format [5]. The specifications of EMDI MK is shown in Table 5.2. The standard used for measurement and the reason for choice of the sampling rate of the obtained harmonic data are explained in the following sections.

5.2.3 Australian power quality standards

Australia has adopted the IEC standards in its power quality standards. Examples of these standards are AS 60038-2000 for standard voltages, AS/NZS 61000.3.6-2001 for harmonics at medium and high voltage and AS/NZS 61000.3.3 for fluctuation at

Table 5.2: EMDI MK3 energy meter specifications.

Main Specs	**Measurement ranges**
Voltages	
Nominal	57 to 240 V (phase to neutral)
Min. to Max.	45 to 290V
Burden	$<$ 10 VA / phase@ V_n (3 phase)
	(As per IEC62053-61)
Currents	
Nominal	1A(C.T.) 5A(C.T.)
Range standard	0.05A-1.2A 0.25A-6A
Range extended	0.05A-4A 0.25A-20A
Short time over-current	20 times the I_{max} for 0.5 second
Starting current	$<$ 0.10% of I_n
Burden	$<$ 0.5 VA / phase
Measurement modes	single phase (3 circuits)
	3 phase 3 wire
	3 phase 4 wire
Pulse outputs	voltage, current, pulse width, polarity
Pulse inputs	voltage
Temperature range	operating -10C to +60 C
	-40C to +85 C
Time keeping	accuracy (internal) $\pm$30 sec /month
	backup time 2 years without power
	backup type Lithium battery
Data storage	flashRAM, Indefinite storage period.
Communications	local ANSI type 2 OPTICOM
	Isolated RS485 or RS232.
	SCADA.

Figure 5.2: EMDI 2000-04XX Energy Meter.

low voltage, and these are adaptations of the IEC 60038, IEC 61000.3.6 and IEC 61000.3.3 respectively. The harmonic monitoring data used in this thesis have been measured according to AS 61000.3.6-2001 [5].

5.2.4 Harmonic data sampling

The suggested time interval according to IEC61000-3-6 for measurements of harmonic, inter-harmonic and unbalance waveforms is 3 seconds for very short interval and the 10 minutes for short interval. Due to memory limitations in the measurement equipment [75], the measurement time interval used in the harmonic monitoring system is 10 minutes following the standard (IEC61000-3-6) specification above for short interval. Each 10 minute data record or sample represents the aggregate of the 10-cycle rms magnitudes, averaged for each 3 seconds over the 10 minute period [76]. The resultant acquired data is thus synchronised at every 10 minute time stamp.

Generally speaking, for data mining application, shorter measurement time interval will be better as it will not only provide more data but also provide greater

number of repetitive patterns. However, a recent study [77] suggested that statistically, sampling at faster rate would not provide additional significant insight.

5.2.5 Harmonic data measurement

The parameters that were measured in the harmonic monitoring program were: the fundamental currents and voltages, the total harmonic distortion (THD) of voltage and current, the 3rd, 5th and 7th harmonics in the three phases. It has been shown that the 5th harmonic is the most problematic of these [78], and the 3rd and 7th harmonics are the next most problematic [5]. Other odd harmonics (such as 9^{th}, 11^{th}, 13^{th}, 15^{th}, 17^{th}, 21^{th}, ..., 47^{th} harmonics) were excluded because of the memory capability of the monitoring equipment (EMDI 2000-04XX Energy Meter) used in the harmonic measurement data. The data was recorded and stored for a three-year period from August 1999 to December 2002 consisting of 10 blocks of two weekly data for every year for each monitored site. This was planned in the monitoring program due to the limitation in the memory of the monitoring device [75]. To take into account seasonal effects, the two week measurements is carried out at different times in the three year period.

5.2.6 Harmonic monitoring data set

The data retrieved from the harmonic monitoring program spans from August 1999 to December 2002 equating to some 200Mb as the total data size to be analysed. Six harmonic currents and voltages (fundamental, 3rd, 5th, 7th, 19th and 49th) in phase A and phase C and the THD of phases A and C were recorded at the 11kV bus at site 1 - 4. Six currents and voltages (fundamental, 3rd, 5th, 7th, 19th and 49th) in phase A , B and C and the THD of phases A, B and C were recorded at the 415V bus at sites 5-7 and as a consequence 238 attributes consisting of four 28 attributes at the 11kV sites (sites 1-4), and three 42 attributes at 415V sites (sites 5 -7) have been pre-processed to form the basic data set for analysis. Figures 5.3 and 5.4, show the

typical output data from the monitoring equipment for the fundamental, 3rd, 5th and 7th harmonic currents in phase a at sites 1 and 2, recorded from the 12th to the 19th of January 2002. These plots illustrate the 10-min maximum fundamental current at 1293A and minimum 10-min fundamental current at 435 A. It is understandable that for the engineers to realistically interpret such large amounts of raw data, the data needs to be segmented into clusters in an intelligent way.

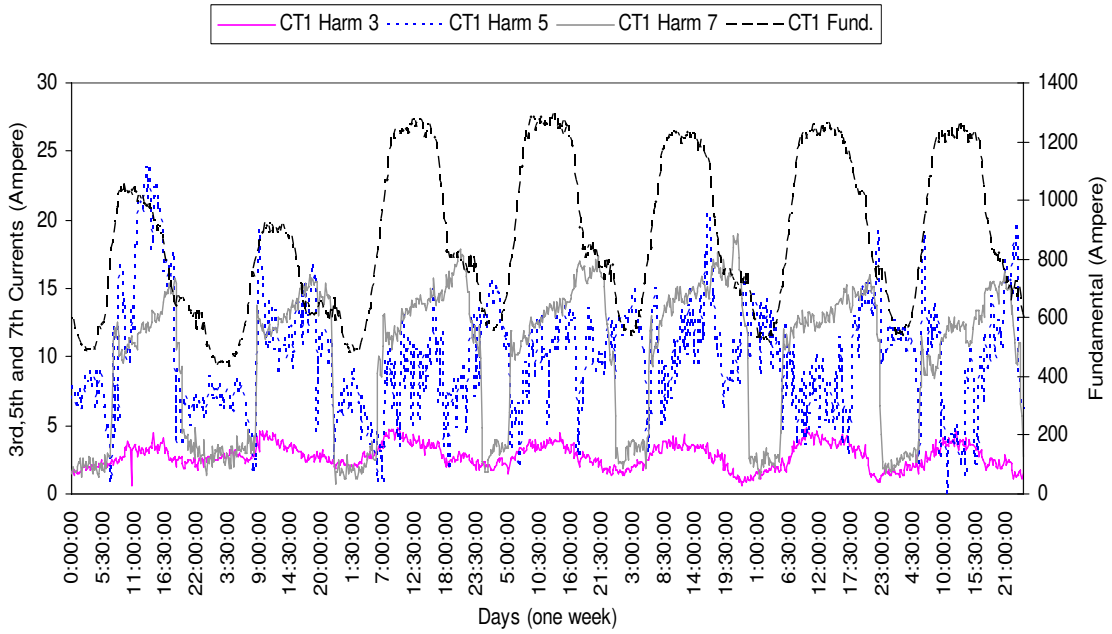

Figure 5.3: Zone substation (site 1) weekly harmonic current data from the monitoring equipment.

5.2.7 Harmonic data preparation

Data preparation is the first step for simulation purposes in which manipulation and transformation of raw data takes place. It allows the data mining analyst to determine whether the obtained data is useful and/or sufficient to build a plausible or good model among other available models. Preliminary analysis is essential, to obtain general insight of the data to be analysed. The data miner needs to under-

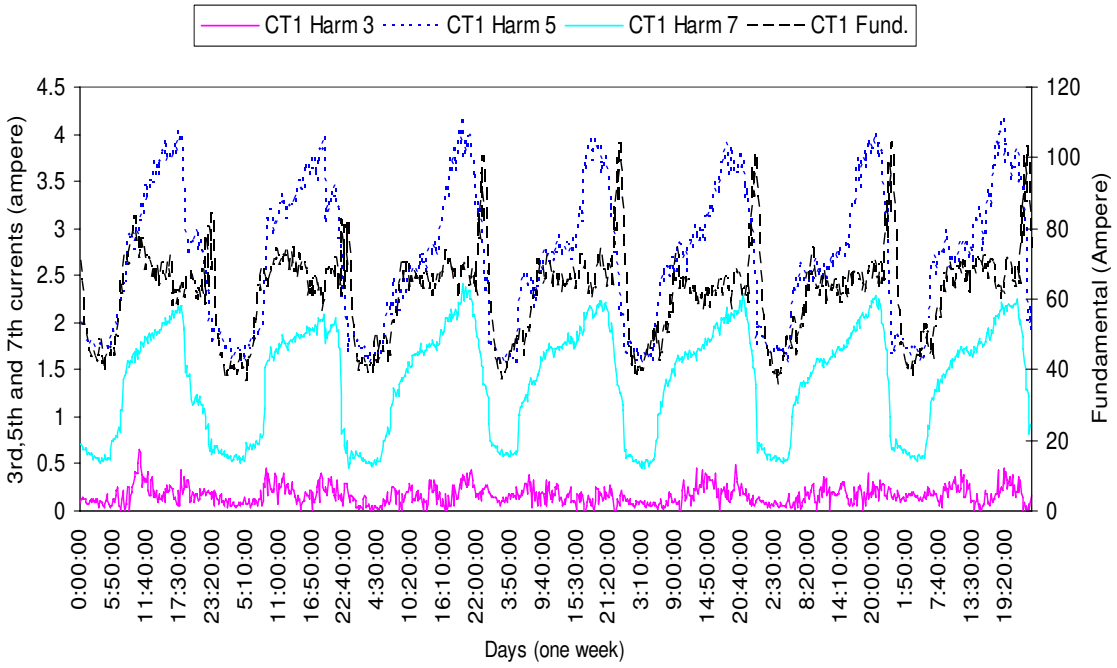

Figure 5.4: Residential feeder (site 2) weekly harmonic Current data from the monitoring equipment.

stand the data in order to select relevant sections, replace missing values and remove noise and outliers from the data. Depending on the data available, the chosen data should be transformed to an appropriate domain if required, such as frequency or time domain. Signal processing techniques such as the Fourier Transform (FT) and the Wavelet Transform (WT) are typical candidates for data transformation from time to frequency domain. The FT is efficient for the transformation of the stationary signals in which the frequency is constant over time. Non-stationary signals or time varying frequency signals can be transformed using the WT. It is fortunate that the data from the harmonic monitoring system is already in the form needed, since the measuring equipment automatically measured the fundamental, 3rd, 5th, 7th and THD harmonic data.

Re-scaling the range and/or the distribution of the data is the last step before mining the data. Changing the range of the variables in data or their distribution is called data normalization. In this thesis the data has been firstly checked to determine

if it contains any missing or corrupted values. Subsequently the resultant data was examined and analysed to uncover any observable relationships and general patterns that might exist within its members.

5.2.8 Harmonic voltage and current trends

As previously discussed, the seven harmonic attributes (fundamental, 3rd, 5th, 7th, 19th, 49th and THD) for currents and voltages of three phase (in phases a, b and c) at the 415V LV side sites (sites 5-7) were recorded at 10 minutes interval. This resulted in 42 attributes for the basic data set for the low voltage sites. The same array of harmonic measurements were acquired at the 11KV sites except for the measurements of phase (b), as the metering voltage and current transformers were unavailable for this phase. This has reduced the number of attributes at 11kV sites (sites 1-4) to 28 attributes. The seven attributes (fundamental, 3rd, 5th, 7th, 19th, 49th and THD) have been primarily analysed, visualized and examined to determine which if any are the more relevant or pre-dominant attributes within this data. No transformation is required because the monitoring equipment automatically generates the harmonic data from the measurement. Throughout this process any redundancy, noise, outliers or missing values that were detected were further manipulated appropriately. This data manipulation for each attribute variables can be explained as follows:

Fundamental voltage and Fundamental current

The fundamental currents and voltages are important attributes patterns because the fundamental currents produce information of the the dynamic operation of the load and the fundamental voltages produce information of the system states during heavy loading, switching operation and/or transient events. Varying the load current will affect the load voltage accordingly. This means that one of these attributes is sufficient to indicate the change in loads. For this reason, the fundamental current is the selected candidate in the harmonic monitoring system since the change in the

fundamental current is more noticeable than that of the fundamental voltage which usually fluctuates around the rated value. Nevertheless the fundamental voltage is still needed to verify the outcomes of the data mining process.

Third harmonic current

From Figure 5.5, it can be seen that the level of the 3rd harmonic current at 11kV is low at about 0.5 % due to the presence of Δ/Y transformers downstream, which block most of the 3rd harmonic current from flowing up. This implies that this attribute variable could be excluded from the data. However, the 3rd harmonic current values at Residential and Commercial sites (sites 5,6) are not excluded due to the presence of unbalance in the single phase loads especially at residential site, and hence, in such cases, the level of the 3rd harmonic current is relatively high as can be seen from Figure 5.6.

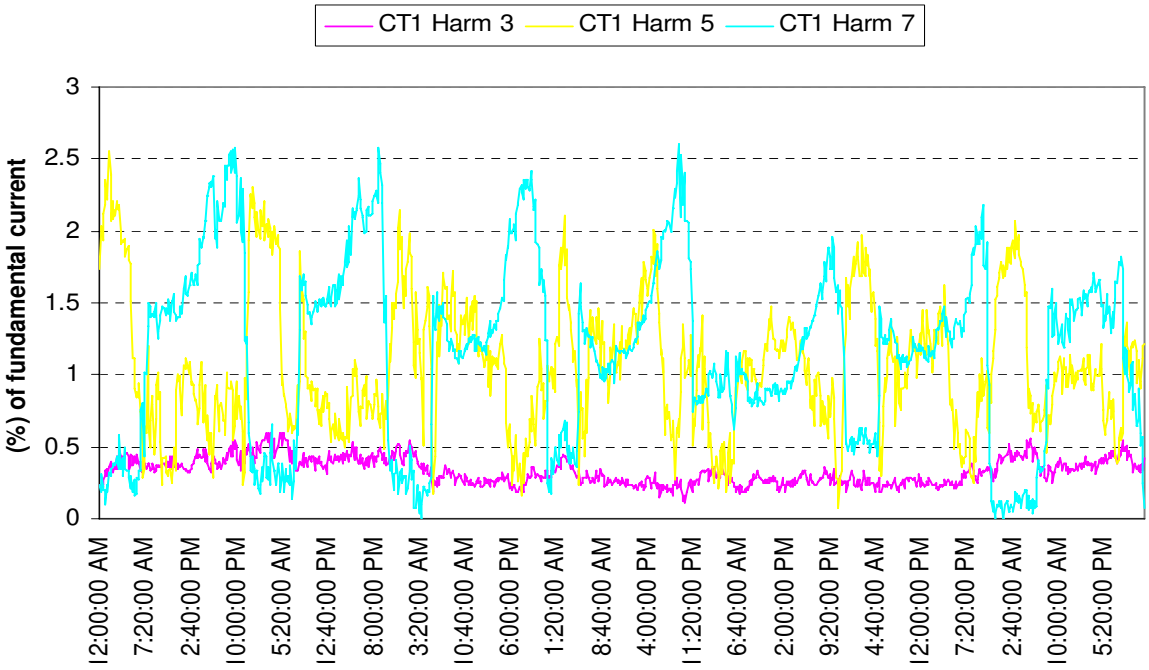

Figure 5.5: Substation (Site 1) weekly low 3rd harmonic current.

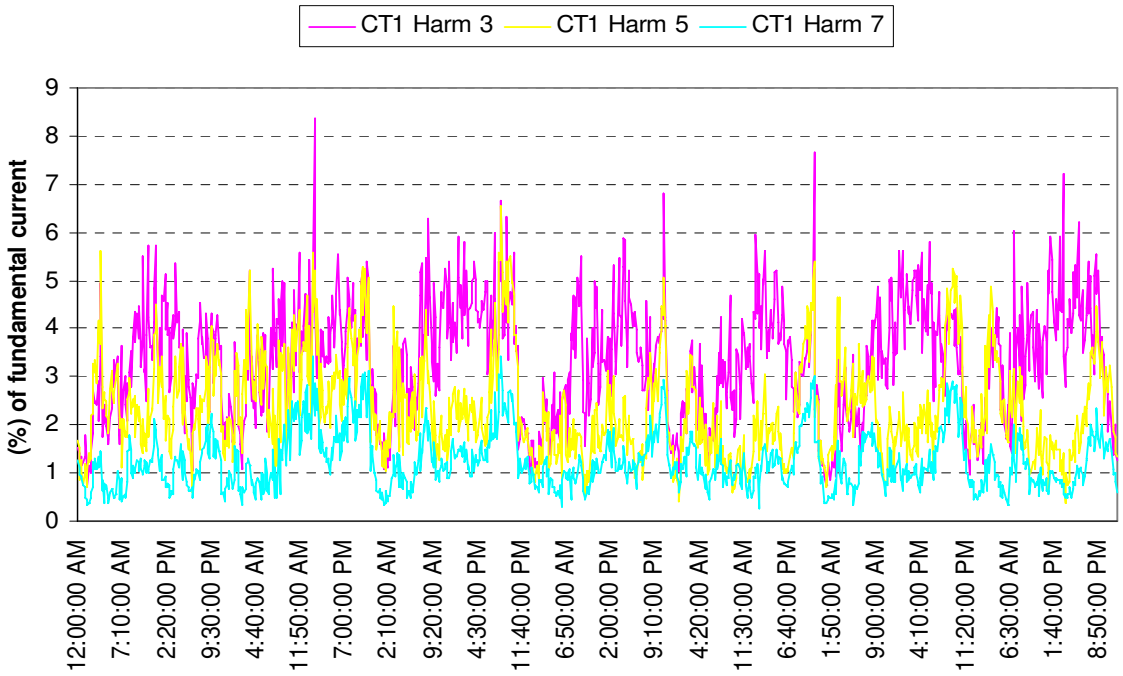

Figure 5.6: Residential site (Site 5)wit a relatively high weekly 3rd harmonic current.

Third harmonic voltage

The third harmonic voltage are found to be very small at most of the 11kV feeder except for Residential and Commercial feeders as shown in Figure 5.7. Again this attribute is not taken into consideration except when studying the Residential and Commercial feeders separately.

Fifth harmonic current

Primary analyses of the measured harmonic data reveals that the fifth harmonic current has the highest level among other individual harmonic currents as shown in Figure 5.8. This confirm previous national and international harmonic surveys [74], [78]. Thus the fifth harmonic current is the most problematic attribute in the harmonic currents and has to be included as input attributes when clustering the harmonic data using data mining.

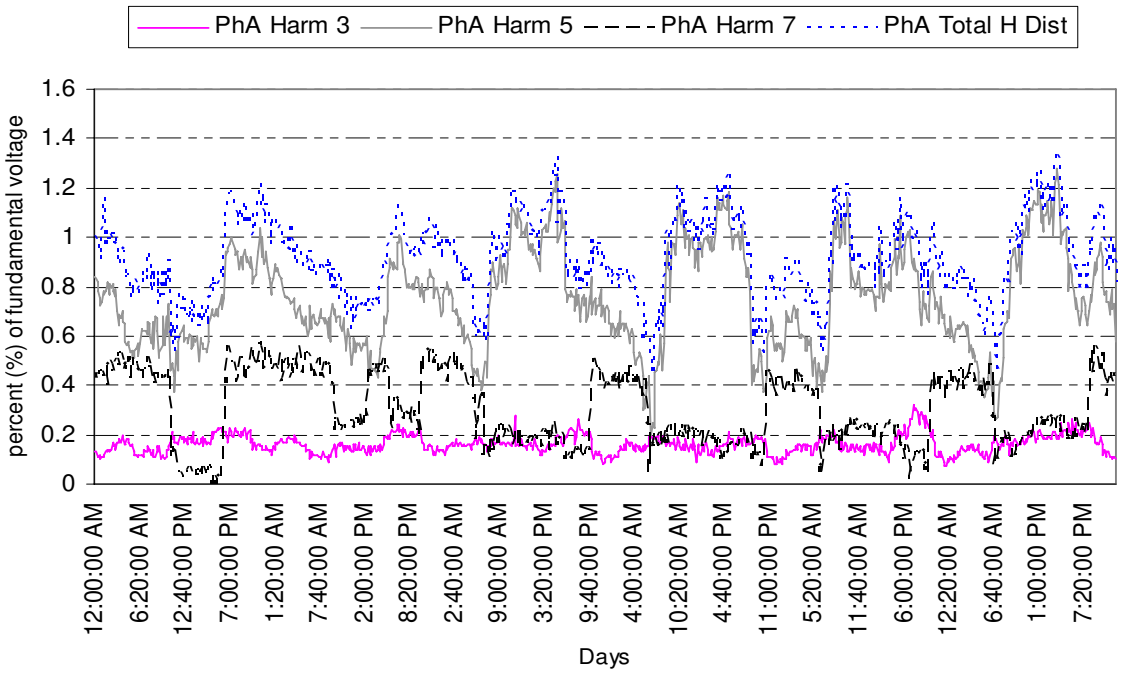

Figure 5.7: Zone Substation (Site 1) weekly high 3rd harmonic voltage.

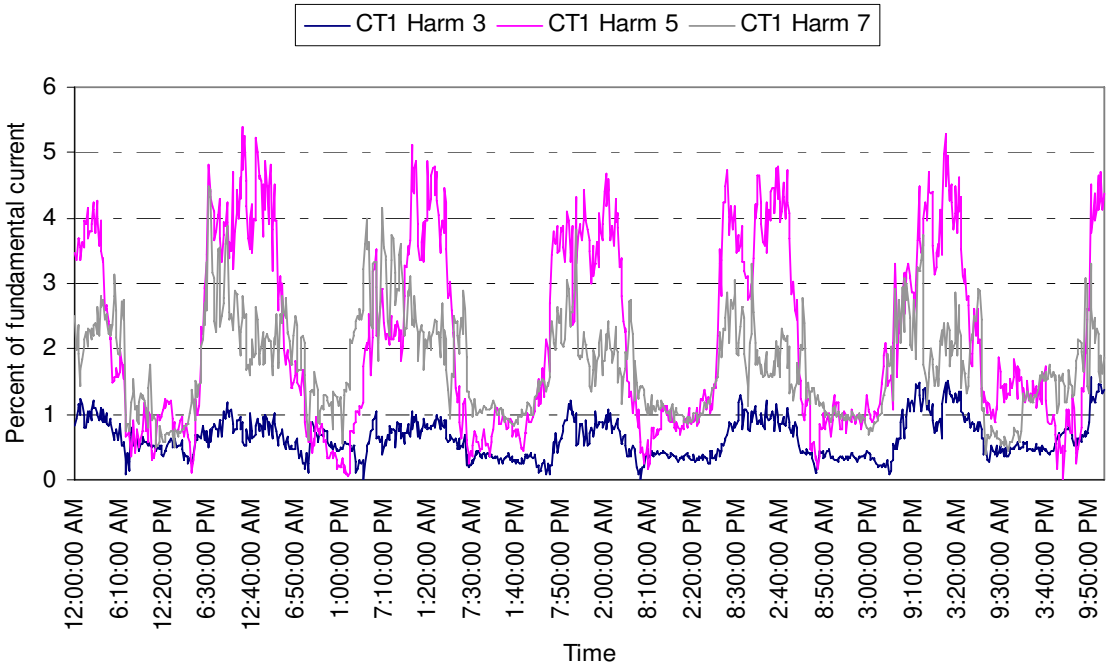

Figure 5.8: Commercial feeder site (site 3) high 5th harmonic currents.

Fifth harmonic voltage

The 5^{th} harmonic voltage is the most problematic one over other harmonic voltages, which was also confirmed in the previous surveys [74], [78] as shown in Figure 5.9.

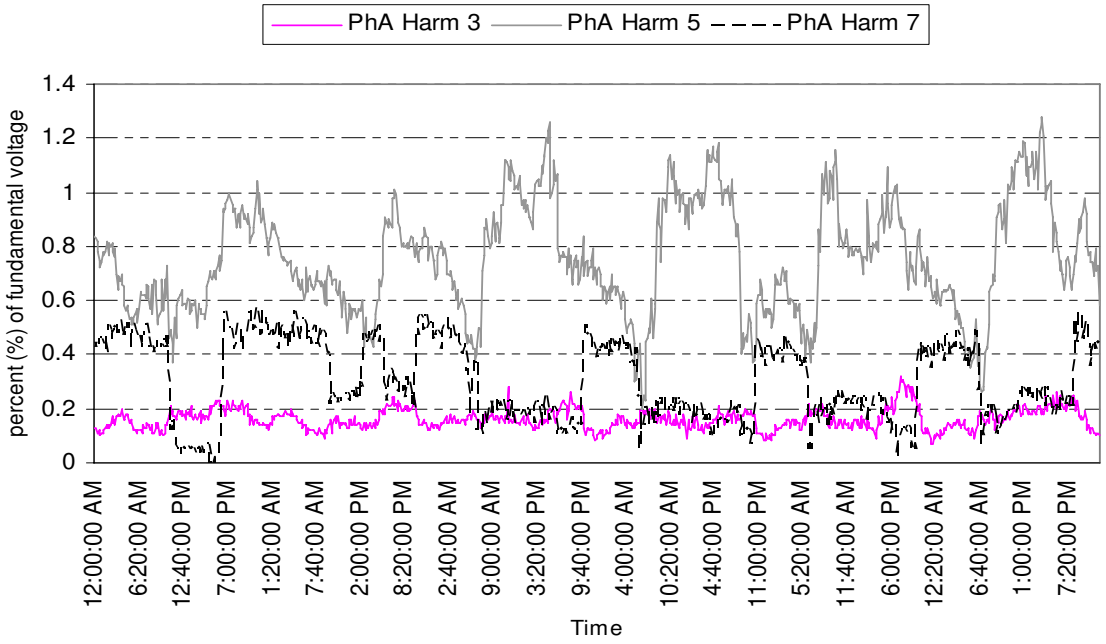

Figure 5.9: Commercial feeder site (site 3) high 5th harmonic voltages.

Seventh harmonic current and voltage

At several sites (sites 1, 2, 3) the peak of 7th harmonic current periods occurs at the time of low 7th harmonic voltage and visa versa as shown in Figure 5.10. This could conceivably be a result of a resonance near the 7th harmonic due to switching a harmonic load or a capacitor which in-turn resonates with the systems inductive impedance at that point. This phenomenon, as well as the high level of the 7th harmonic currents, needs more investigation and hence the 7th harmonic current and voltage should be included as important attributes for data mining.

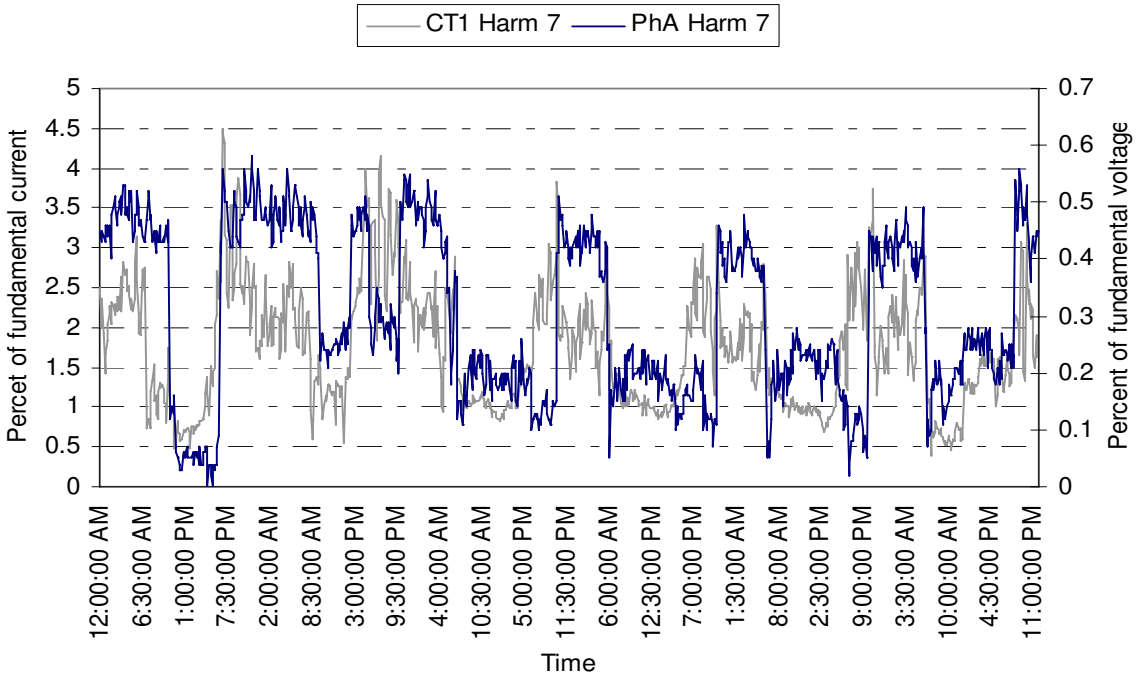

Figure 5.10: Commercial feeder site (site 3) 7th harmonic voltage and current.

19th harmonic current and voltage

From Figure 5.11, it can be seen that the 19th harmonic voltage is very low at approximately 0.1 % of fundamental voltage which is close to the limit of the resolution of the monitoring equipment (EDMI MK). This indicates that the 19th harmonic voltage may not have been measured with sufficient accuracy and hence should not be included in any further data mining or modelling steps. Similar observation has been found on 19th harmonic current where its maximum value does not exceed 0.3% of fundamental voltage.

49th harmonic current and voltage

The 49th harmonic voltage attribute was subsequently examined and was found to be the attribute with a lot of noise. A further test was devised to confirm if the 49th harmonic voltage attribute variable could feasibly be used to represent occurrences of system operation events other than the noise in the data. Subsequently, data

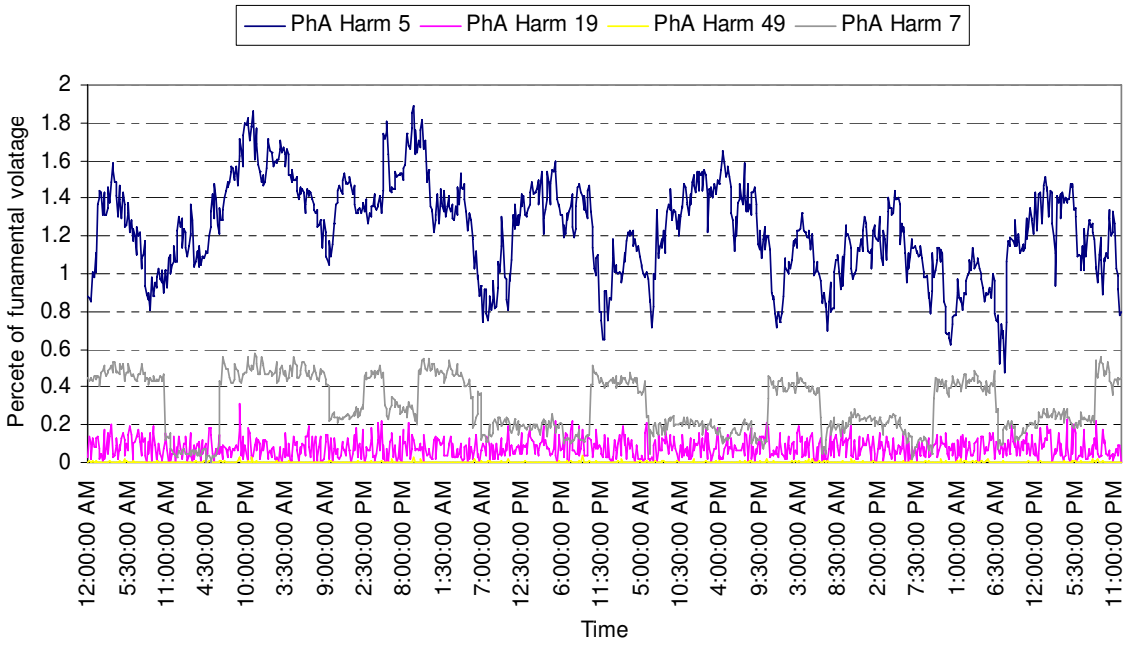

Figure 5.11: Substation site (site 1) low 19th harmonic voltage.

records comprising only the 49th harmonic voltage were isolated and clustered using the ACPro software and the maximum number of the obtained clusters was five. Table 5.3 details the parameter values (μ , σ and π) of these generated clusters in which the most abundance cluster (88 %) has mean value of 0 and a very small standard deviation ($\sigma = 0.000020$).

Table 5.3: Five clusters generated from ACPro for the 49th harmonic voltage.

Cluster	Abundance	Mean	Standard deviation
s0	0.886037	0.000000	0.000020
s1	0.065228	0.010000	0.000020
s2	0.043031	0.020000	0.000020
s3	0.001116	0.040000	0.000020
s4	0.004588	0.030000	0.000020

Indeed all clusters have similar means combined with similarly very small vari-

ance and therefore it is not likely that useful distributions can be obtained from this attribute, but rather it may model some noise source. The source of this noise may conceivably arise from the accuracy of the measuring equipment as the 49th harmonic is a relatively high frequency for the monitoring circuits, or possibly from the nature of 49th harmonic values itself. Unlike the continuous range of the other harmonics attributes (5th, 7th and 19th) the values of the 49th harmonic have a discrete nature (for example, values of 49th voltage are restricted to 0, 0.01, 0.02, 0.04 and 0.05). Accordingly, this variable was removed to reduce any uncertainly, or contributed noise effect, that it may pose. The corresponding 49th harmonic current was found to have similar characteristics and hence the 49th harmonic current and voltage were not expected to carry additional information other than noise.

The total harmonic distortion (THD) current and voltage

It has been found from the monitoring program that the 5th harmonic voltage at all sites is highly correlated with the total harmonic distortion (THD) voltage. Figure 5.12 shows a typical reading at the substation Site 1. This means that the 5th harmonic voltage is the prevailing harmonic variable among other individual harmonics (3rd, 7th, 19th, 49th), which indicates that the THD voltage is a redundant variable and could be removed from the data set. The 5th and the total harmonic distortion (THD) currents also have the same correlation although at a lesser extent and thus the redundancy is still present as shown in Figure 5.12.

5.2.9 Harmonic data selection

From the analysis of the data preparation described in Section 5.2.8, the fundamental voltages and currents are the primary selected attributes. The 5th harmonic voltages and currents are also primary attributes as they have similar trends with the voltage and current THDs. The 7th harmonic voltages is of considerably high magnitudes and also fluctuates significantly a lot, especially in relationship with the 7th harmonic

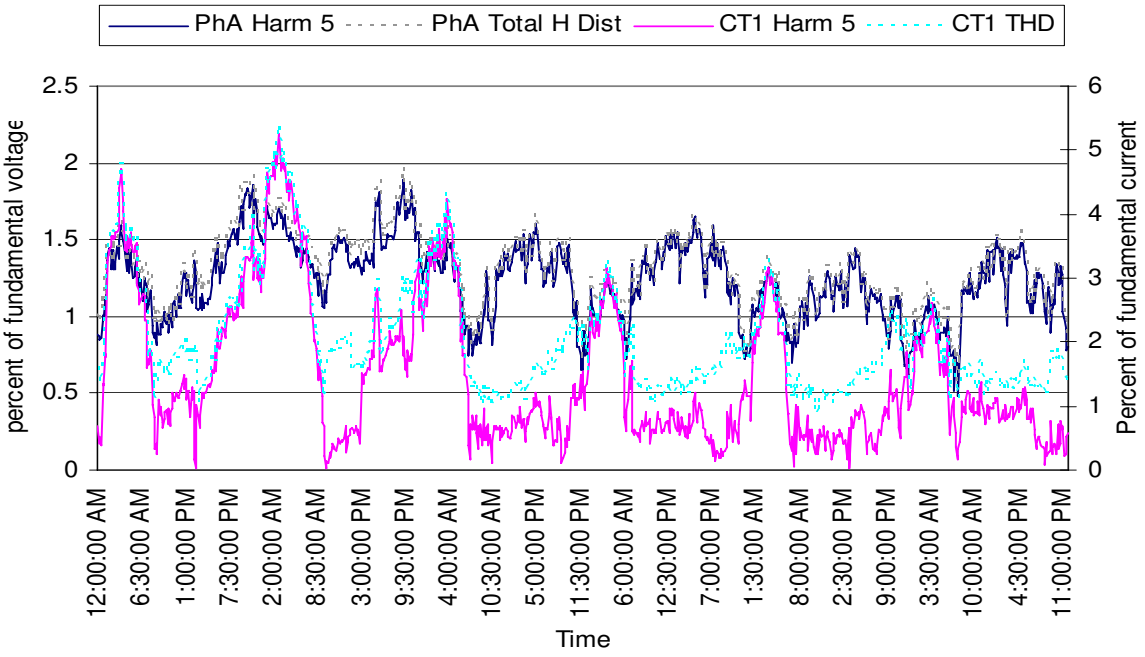

Figure 5.12: Substation site (site 1) total harmonic distortion (THD)and 5th harmonic current and voltage.

current, and therefore they have also been chosen as attributes for analysis. The 3rd harmonic currents and voltages are only included in the data mining process if their values are not truncated by the Δ/Y transformers as in the residential and commercial low voltage sites (sites 5 and 6). The 19th harmonic currents and voltages are not chosen as input attributes because of their low values (compared to the monitoring equipment accuracy) and also the 49th harmonic currents and voltages are also not chosen because of the noisy data.

For the substation site, only the current attributes at each site are used, the voltage is the same at all bus and hence only one voltage value is used. It has also been shown in [15] that extracting information from the current attributes will improve the classification of power quality events. This then gives good reason as why sometimes only the harmonic current attributes are included.

5.2.10 Rescaling of harmonic data

The harmonic monitoring system measures the fundamental voltages and currents from the secondary of the voltage and current transformers at sites 1,2, ..., 7, where the voltage transformers ratio is 100:1 for 11kV and 1:1 for the 415V site. The current transformers at sites 1,2,...7, have the following turns ratio:

- site 1, current transformer turn ratio, 1500:5.
- site 2, current transformer turn ratio, 400:5.
- site 3, current transformer turn ratio, 400:5.
- site 4, current transformer turn ratio, 400:5.
- site 5, current transformer turn ratio, 800:5.
- site 6, current transformer turn ratio, 1000:5.
- site 7, current transformer turn ratio, 3000:5.

The voltage and current values from the measurement therefore need to be re-scaled to take into account the current and voltage transformer ratios. However, the harmonic currents and voltages at each site are measured as a percentage of their corresponding fundamental voltage and current values. For this reason the re-scaled current and voltage harmonic values above need to be returned to its actual values by multiplying it with the re-scaled fundamental currents and voltages. The resulting data values are now given in terms of Volt and Ampere.

As a result of the significant difference between the actual values of fundamental currents and voltages as well as between the actual values of harmonic currents and voltages, a normalisation method needs to be carried out in such a way that all the data lie in the range between 0 and 1.

5.2.11 Normalisation of harmonic data

There are several types of data normalization that are relevant for this data as shown below:

Decimal scaling normalisation

In the decimal scaling a decimal point moves to change integers into decimal numbers. For example if the data is in the range between -100 and 400, then the normalised range will be between -0.1 and 0 .4. This method is impracticable to use because the normalised range is affected by the range of original data. For example a data range between 250 and 350 will be normalised in a very small range between 0.25 and 0.35 in which the entire data set will be distributed.

Standard deviation normalization

In the standard deviation normalisation method, the mean is subtracted from each element of the data values, and then divided by the standard deviation, transforming the distribution of data into a standard normal distribution with mean 0 and standard deviation 1 using the following equation.

$$y = \frac{(x - \mu)}{\sigma} \tag{5.1}$$

where:
x is the data value,
μ the mean value of data,
σ the standard deviation.

In this method the normalised data is different from the original data, as the variability is removed from the data and so the relations between the variables will no longer exist.

Normalisation to average or maximum value

Normalsation to average is to find the average of the data set and then each element of the data is divided by the average. Normalisation to maximum value is to find

the maximum of the data set and then each element of the data is divided by the maximum. This method is used to normalise the data in the range (0 - 1) when the average or the maximum value is dominant in the data. This method is used here in this work in Chapter 6 to normalise harmonic voltage and current data.

(Minimum-maximum) range normalization

In the minimum- maximum normalisation method, the minimum value of the data is subtracted from the data elements and then divided by the range of the data using the following equation:

$$y = \frac{(x - min(x))}{(max(x) - min(x))} \tag{5.2}$$

where:
x is the data set,
min(x) the minimum value of x,
max(x) the maximum value of x.

The advantage of this method is that a better normalised interval (between 0 and 1) can be obtained which is not biased by the data values or its distribution while keeping the variabilities similar as in the original data. This method of normalisation is used in this thesis in Chapters 7 and 8 to normalise harmonic current data.

5.2.12 Other measured data (temperature and reactive power)

To confirm the results from the clustering program, several other data sets are also used as additional information to the harmonic monitoring measurement data, when they are demand to be relevant and important.

Temperature information in the area around the zone substation and the related suburban areas of Sydney the substation under study was seen as important in understanding some of the seasonal variations in the data. This information was obtained from the Bureau of Meteorology, New South Wales regional office [79]. The tempera-

ture values were measured in Celsius every half hour. High temperature always leads to the use of more air conditioning system, and therefore will lead to higher fundamental and harmonic currents. When inspecting clusters obtained from the data mining program, it is usually useful to correlate it with the temperature measurement to isolate those occurring at periods with unusually high temperature.

The local utility distributor also provides other additional data, such as the timing of the switching on and off of shunt capacitor switching details located at the zone substation, as well as, the readings of the var-meter for the reactive power (Q) at that substation during the switching on and off the capacitor. These additional data can be used to confirm any suspected events identified from specific cluster or clusters, such as due to capacitor switching (reactive power data) events. or turning on air conditions in hot days (temperature data) which will be covered in Chapter 7.

5.3 Summary

Large amounts of harmonic data has been monitored in a distribution system in Australia for three years. This data holds much more information than that which has been reported using classical statistical techniques. It is difficult to understand the harmonic events in the data from visual observation. Therefore Data mining techniques are suggested as a means with which to extract further information from the harmonic monitoring data. Preparations of this data in order to successfully applying data mining techniques have been presented. The 19th harmonic currents and voltages have not been measured with enough accuracy and are relatively very small in values and the 49th harmonic currents and voltages were found to represent noise in the data and thus they were excluded from data set in the preparation step. The 3rd harmonic currents and voltages was included for data from low voltage sites (415V) whereas those from the high voltage sites (11kV) were excluded.

Chapter 6

Anomaly detection and pattern recognition

6.1 Introduction

Although researchers have realised that the large amounts of PQ monitoring data hold much more information than that reported using classical statistical techniques [1], few have taken the opportunity to exploit this additional information. Such information could include recognition of disturbance level patterns prior to significant power quality events, relating plant or system events to disturbances, and identifying growth trends of disturbance levels. Anomaly detection is another issue that can provide considerable information about unusual operating conditions that often only occur for short periods of time. In this chapter the Minimum Message Length clustering algorithm is used to identify patterns and detect anomalies in harmonic data from the MV/LV electrical distribution system explained in Chapter 5. The super-group formation from different sites, using the Kullback-Liebler (KL) distance, is explained via link analysis and visualization techniques. The effect of these sites on each other is examined and the causality of harmonic distortion is also discovered using the classification rules which are subsequently generated by the decision tree.

6.2 Data preparation

The dominant harmonic currents and voltages attributes suggested from Chapter 5 (fundamental, 3rd, 5th, 7th and THD) have been selected from four different sites (Substation, Site 1, Residential Site 5, Commercial Site 6 and Industrial Site 7) as shown in Figure 5.1. The THD voltage is included here only to verify the conclusion obtained in Chapter 5 that the THD voltage is a redundant attribute. The harmonic data (fundamental, 3rd, 5th, 7th and THD current and voltage) used in the application is one file of 8064 instances which consists of four combined files (4x2016) from the selected sites. Each of these files is a block of two week data between 19/8/1999 and 1/9/1999. The data was normalised, using the method described in Section 5.2.11, by dividing each data point by the typical values of each corresponding attribute.

The suggested typical value for the harmonic currents is the maximum value whereas for the harmonic voltages is the average value. The voltage attributes are normalised by the average values because the voltage values vary around this average value. On the other hand the fundamental current is only limited by the current carrying capacity of the feeder, and hence the maximum value of the current is selected for normalisation purposes. After normalisation, the maximum value of the harmonic current attributes, and the average value of harmonic voltage attributes will have the value of one. The normalised attributes were used as an input to the MML algorithm with a given accuracy of measurement (Aom) that can be calculated from the variance of the entire data set as following:

$$Aom = \frac{(\alpha * \sigma)}{x} \tag{6.1}$$

where:

Aom: is the Accuracy of measurement.

σ: is the variance of the entire data.

α: confidence interval.

x: range of the data.

The number of clusters obtained was automatically determined based on the significance and confidence placed in the measurements. This significance and confidence can be estimated from the standard deviation (σ) of the data and from confidence interval (α) respectively. Each cluster of data is automatically grouped according to a learned pattern, and the abundance of each group is calculated over the full data range. The abundance value for each cluster represents the proportion of data that is contained in the cluster in relation to the total data set. If for example, only one cluster was formed then the abundance value for that cluster will be 100%.

Each generated cluster can therefore be considered as one of the statistical distributions in the mixture model to represent the thirty attributes (being the fundamental, 3rd, 5th, 7th and THD currents and voltages for each of three phases) of the data set within an acceptable variance. If new data lies beyond the variance, another cluster is created until the full model of the mixture of probability distributions with minimum message length is found. Using a basic spreadsheet tool the clusters are subsequently ordered inversely proportional to the actual abundance, i.e. the most abundant cluster is seen as s0 and those that are progressively rarer have a high value type numbers ($s1, s2, ..., etc$).

6.3 Anomaly detection and pattern recognition from harmonic clusters

A total of 23 ($s0, s1, s2, ..., s22$) clusters, were automatically determined by ACPro based on the Aom value. The generated clusters are then plotted using the graphical features of the spreadsheet tool. The 5th harmonic is the attribute of interest, since it is the most problematic attribute [78]. An example of this is shown in Figure 6.1 where it can be seen that the values of 5th harmonic currents and voltages at Site 7 are changing between low and medium values at high abundance level (to the left hand side) and have high values at low abundance level (to the right hand side).

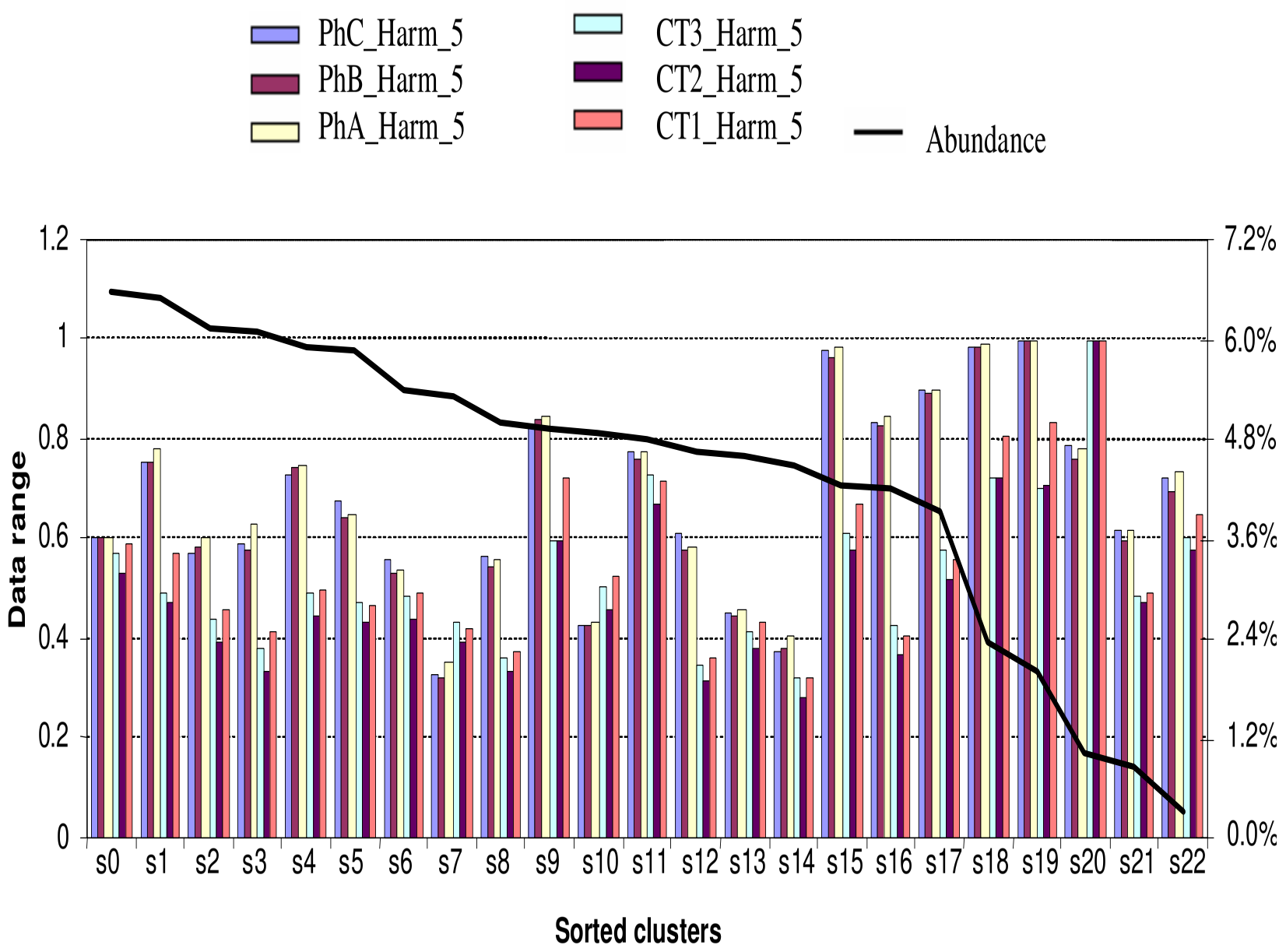

Figure 6.1: Abundance of clusters of 5th harmonic current and voltages over each phase of monitoring results.

This shows that the abundance value (superimposed on the 5th harmonics in Figure 6.1) can provide an indication of the importance of each of these clusters. A small abundance value may mean the cluster represents rare occurrences and this may point out instances when the system needs to be observed more carefully. Also clusters that are of significant abundance (i.e. they happen more often) but contain high magnitudes of the 5th harmonic voltages (for example s1, s4, s9) may need further investigation.

Several other interesting patterns may be extracted from the properties of the obtained clusters. One example of this extracted information is in the form of the interrelationships among harmonic attributes that are illustrated in Figure 6.2. In general, Figure 6.2 shows a strong relationship between the 5th harmonic current and (THD), i.e. THD rises and falls with the value of 5th harmonic. This indicates that for the data set from the harmonic monitoring system the 5th harmonic current is a significant contributor to the overall distortion level and is highly correlated to THD, as previously reported using traditional analysis [74]. This means that THD can be excluded from the data.

In the power quality domain, rare clusters are those that may represent unusual events which consequently need to be highlighted as clusters that may require some further detailed examination. The same data set explained in the last section, for example, is now segmented into only five clusters as shown in Figure 6.3. The cluster s5 has the least abundance value of 6%, however, the mean value of the fifth harmonic in this cluster is the highest.

This cluster (s5) acquires its importance from both the high value of the fifth harmonic current (100% of data range) and its least number of occurrences. This suggests that the event associated with this cluster needs to be observed more carefully in the future and to ensure that these levels do not increase beyond that specified in the harmonic standard (IEC61000-3-6). The concept of rare clusters may also be used to identify the most significant distorting loads at different customer sites.

Figure 6.4 illustrates the pattern of the five clusters (see Figure 6.3) over the

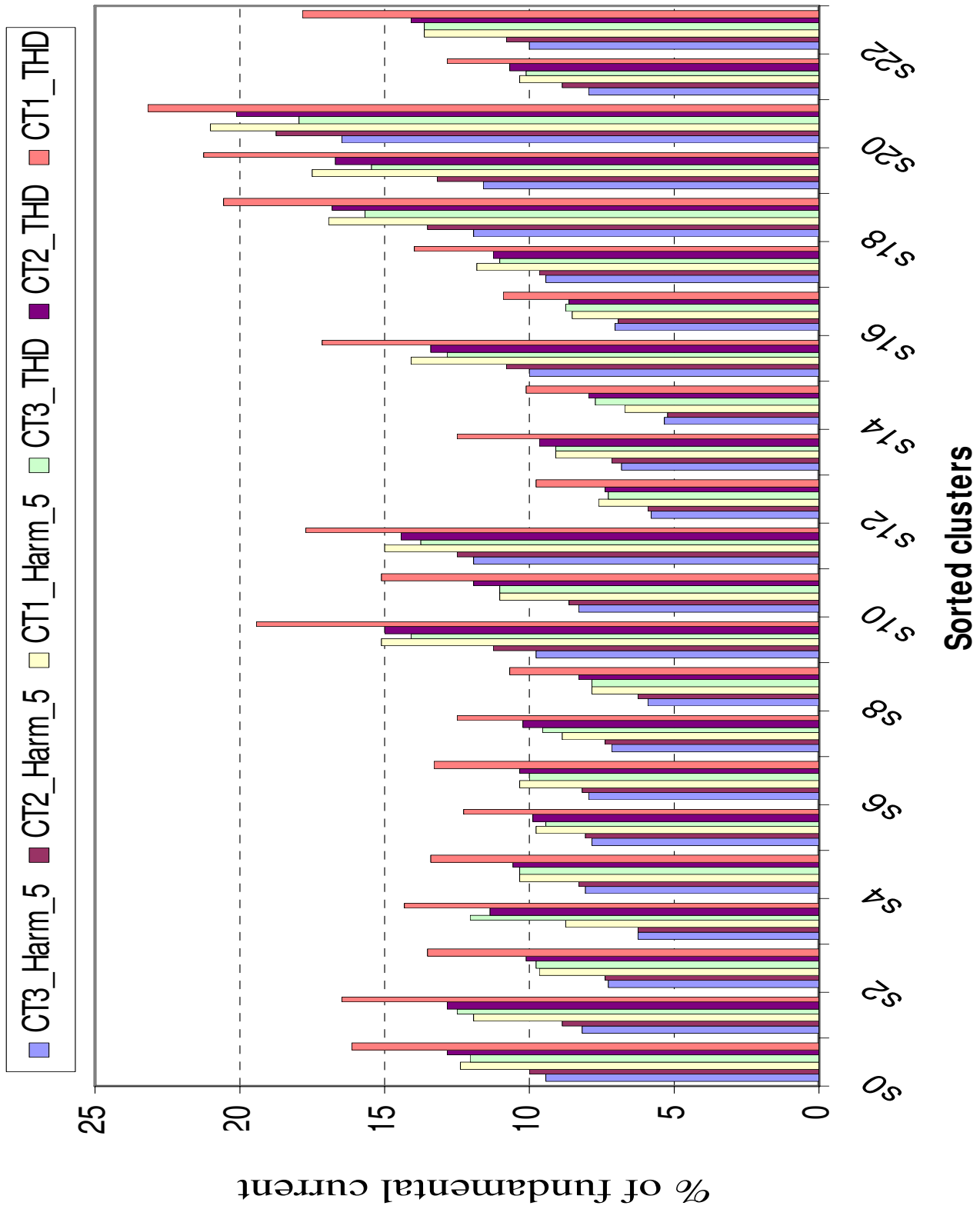

Figure 6.2: Cluster of 5th harmonic current and ITHD over all three phases from Site 7

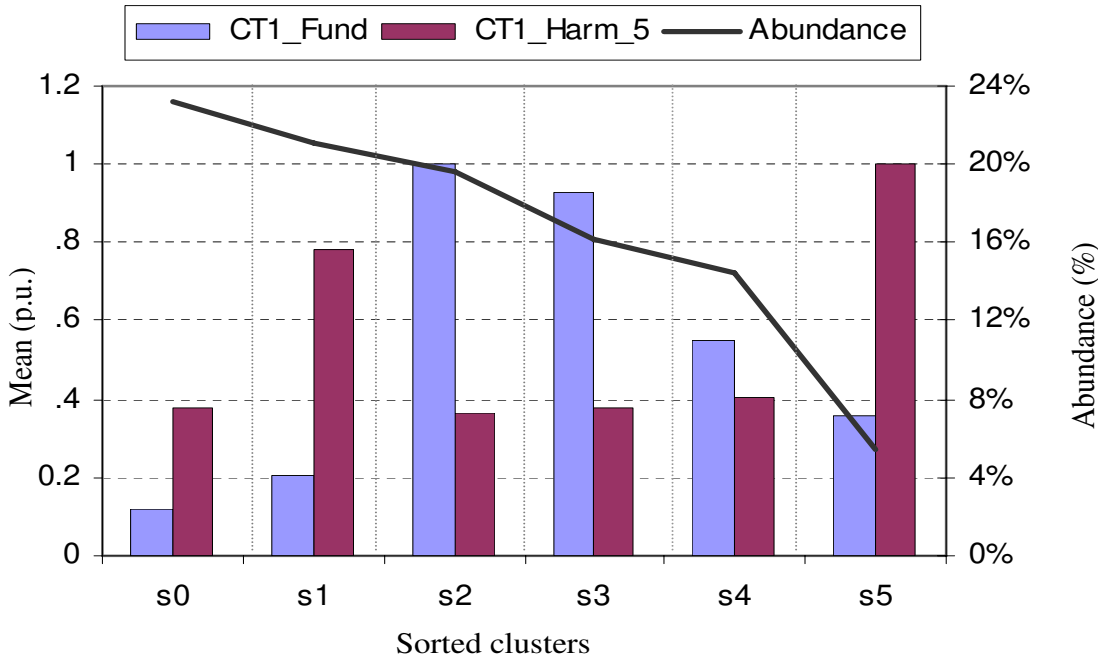

Figure 6.3: Five randomly generated clusters each with its own mean and standard deviation.

period of one week at sites 1, 5, 6 and 7 in Figure 6.4. Each cluster is represented with a colour in grey scale in proportion to the abundance of that cluster, i.e. the least abundant cluster will appear as black and the most abundant cluster will be the lightest shade of grey. Noticeable characteristics from Figure 6.4 include the two distinctive darker patterns towards the left hand side of the MV substation data (Site 1). This indicates that the least abundant occurrences appear during the mornings of the weekend days. Also the commercial site, Site 6, exhibits a recurring pattern of harmonics over each day, noting that the shopping centre is in operation seven days a week. The residential and industrial customer clusters are somewhat more random than the other sites, suggesting that harmonic emission levels follow no clearly distinguishable patterns.

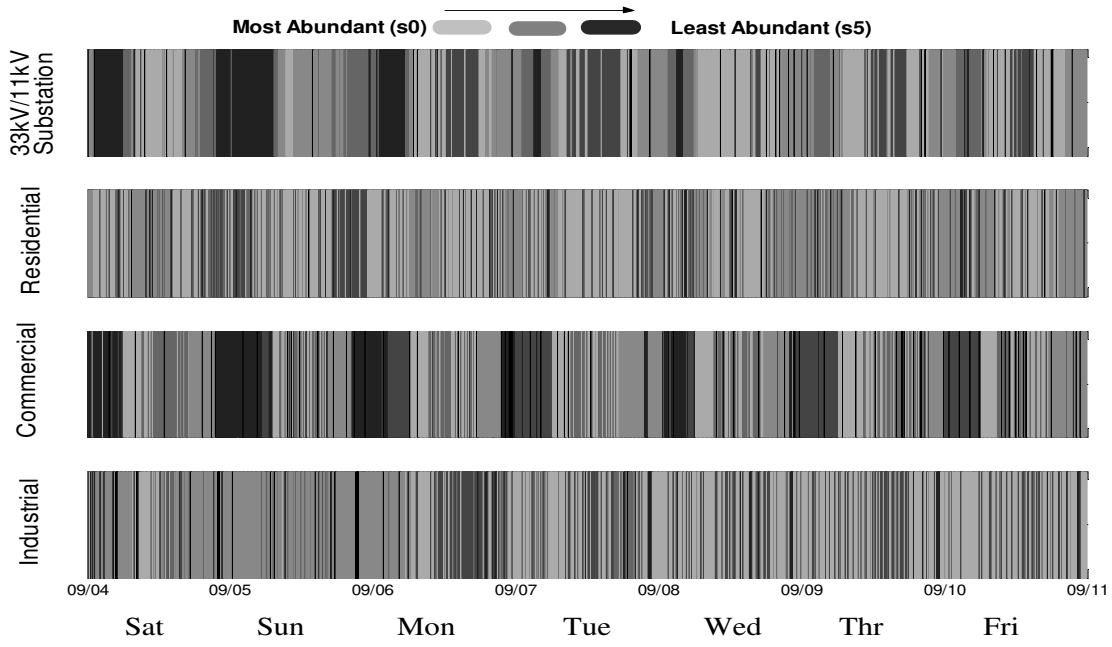

Figure 6.4: Clusters of harmonic emissions from the different customer loads and system overall for a one week period.

6.4 Abstraction of super groups from harmonic data

Further additional information can be retrieved by using the Kullback-Liebler (KL) distance identified in the next section to merge the similar clusters into super-groups.

6.4.1 Kullback Leibler Distance (KL)

The Kullback Leibler Distance (KL), also known as the relative entropy, is a measure of distance between any two distributions.

For two discrete probability distributions, $p = p1, ..., pn$ and $q = q1, ..., qn$, the KL distance is defined to be:

$$D_{KL}(p(x), q(x)) = \sum_{x} q(x) \ln \frac{q(x)}{p(x)} \tag{6.2}$$

where: D_{KL} is the Kullback Leibler between distributions p and q.

For continuous probability distributions, the sum can be replaced by an integral:

$$D_{KL}(p(x), q(x)) = \int_{-\infty}^{\infty} q(x) \ln \frac{q(x)}{p(x)} dx \tag{6.3}$$

From an information theory point of view the KL-distance is interpreted as the expected difference in the number of bits required using a code based on a target distribution, q, compared with a code based on the true distribution, p.

It has been proven [42] that the KL distance is not a true metric distance as the distance between distributions p and q is not necessarily the same distance between distributions q and p, however it can be used to measure similarity between clusters: the smaller the KL distance between two clusters, the more similar these clusters are. KL distance equals 0 if and only if p = q.

KL distances can provide useful information about all the generated cluster components of the segmented mixture model. In this work they are calculated directly from the MML program ACPro, producing a matrix of KL distances between each pair of clusters of the model. These are used along with a multidimensional scaling (MDS) algorithm to formulate various super groups from the harmonic monitoring data. The MDS procedure is explained in the following section.

6.4.2 Multidimensional scaling (MDS)

Multidimensional scaling (MDS) is a technique used to convert proximities (similarity or dissimilarity) between objects in data to distances by which a geometric configuration of these objects can be obtained. These distances can be represented in low dimensional space (one or two) while keeping the same measures. To measure the degree of fitting to the data, an objective function called the stress function is utilised, which yields an index number. The smaller the index number, the better the configuration match to the data. The equation which calculates this index is given in Equation 6.4 :

$$f_{stress} = \sqrt{\frac{\sum_i \sum_j (f(\delta_{ij}) - d_{ij})^2}{\sum_i \sum_j d_{ij}^2}} \tag{6.4}$$

where:

f_{stress}: is an objective function.

δ_{ij}:is the proximity between objects i and j.

d_{ij}: is the distance between objects i and j.

To easily visualise the KL distances between the clusters, the graphical MDS program, Knowledge Network Organising Tools (KNOT) [80] can be used to convert the proximities between the generated clusters to a lower dimensional scale. Here the distances between clusters that are established after calculating KL distances are can be visualised on various 2 dimensional projections. This technique can be useful in identifying certain super group abstractions, by removing links that exceed a selected KL distance threshold.

6.4.3 Segmentation of harmonic data into Super-groups using KL and MDS

To explain the concept of super-groups, a subset of the harmonic data described in Section 5.3 being (3rd, 5th, and 7th) from different sites (1, 5, 6, 7) was used as selected attributes for analysis by the MML segmentation software. Eleven clusters (s0, s1, s2... s10) was selected as an input parameter "k", rather than allowing the clusters to be automatically generated by the software. Details of the abundances, means and standard deviations of the 5th harmonic current across these 11 clusters are illustrated in Figure 6.5. KL-distance tool of ACpro is applied on the model to generate the lower triangular 11 ×11 matrix of KL distances shown in Table 6.1. The highlighted distance values represent the three largest and the three smallest distance values. For example, the distance between s3 and s1 in Table 6.1 is given as 3186, which is the largest distance, which suggests that there is a considerable difference

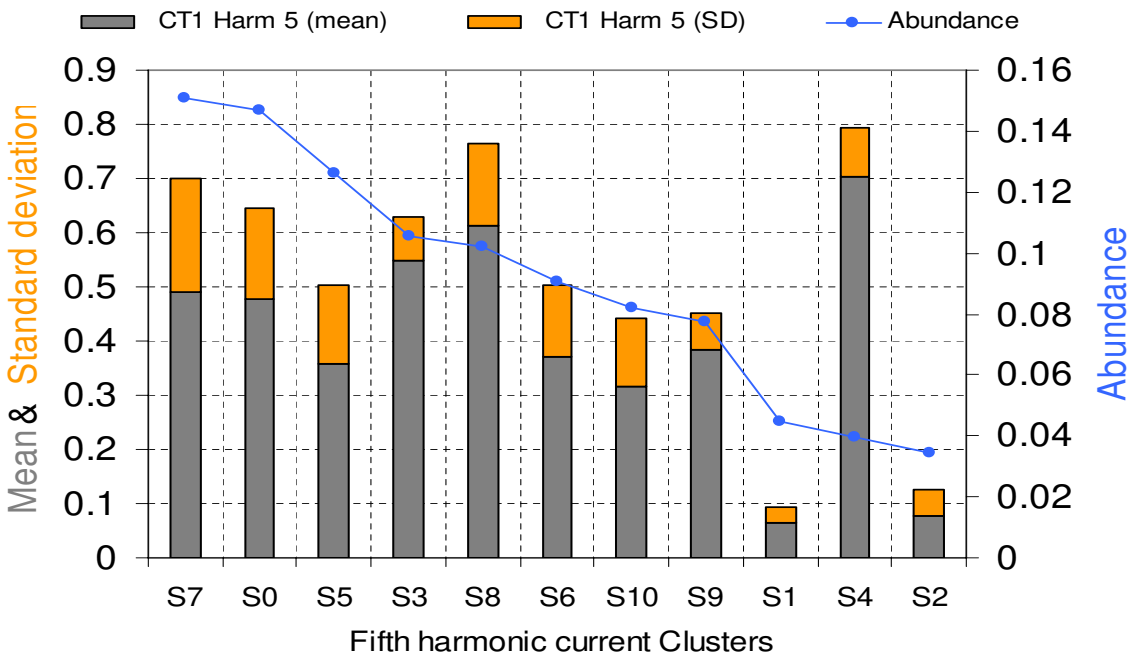

Figure 6.5: Abundance, mean and standard deviation for each clusterof the 5^{th} harmonic current.

Table 6.1: Kullback-Lieber distances between components of the 11 cluster mixture model.

s0	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10
s0										
s1	**2674**									
s2	832	232								
s3	62	**3186**	**1157**							
s4	181	2486	941	178						
s5	59	1077	358	185	127					
s6	51	1277	361	173	169	**37**				
s7	51	2518	871	107	155	58	142			
s8	102	2773	1003	113	169	145	201	39		
s9	450	1486	612	519	649	194	234	471	365	
s10	115	867	332	233	153	**34**	107	**36**	70	116

between these two clusters, while on the other hand the distance between s10 and s5 is only 34, which suggests that there is a lot of similarity between these two clusters.

The links between all clusters, based on the KL distances, were visualized using the graphical MDS program KNOT which effectively reduces an 11 dimensional model into a two dimensional graph. The resulting super-groups were subsequently formed by removing any link whose distance exceeded a certain dissimilarity threshold. The obtained super-groups (A, B, C, D and E) are shown in Figure 6.6. Most of the super-group abstractions are formed based on the site type, for example supergroup A covers the industrial site, supergroup D covers the substation site, super-group C and E covers the commercial sites, with super-group C being separated because the distances between s9 with s2 and s9 with s1 are larger than the distance between s1 with s2. Super-group B is formed from clusters containing data from all sites.

The residential site does not seem to have a particular super-groups which means that the influence of harmonic emission (or participation) from this site is very low. The concurrences or synchronism at different sites of two or more of these super-groups indicates that there is a harmonic mutual effect between these sites at that particular time. For example, a temporal correspondence of super-group A at the industrial site can be observed with both super group D at the substation site and super group E at the commercial site early in the morning of each day as shown in Figure 6.7. The associated pattern of harmonic emission factors that might exist in the formation of these super-groups can also be extracted using the classification techniques of supervised learning.

One of the most commonly used techniques in symbolic (transparent or understandable) supervised learning is the automatic induction of decision trees which will be explained in the next section.

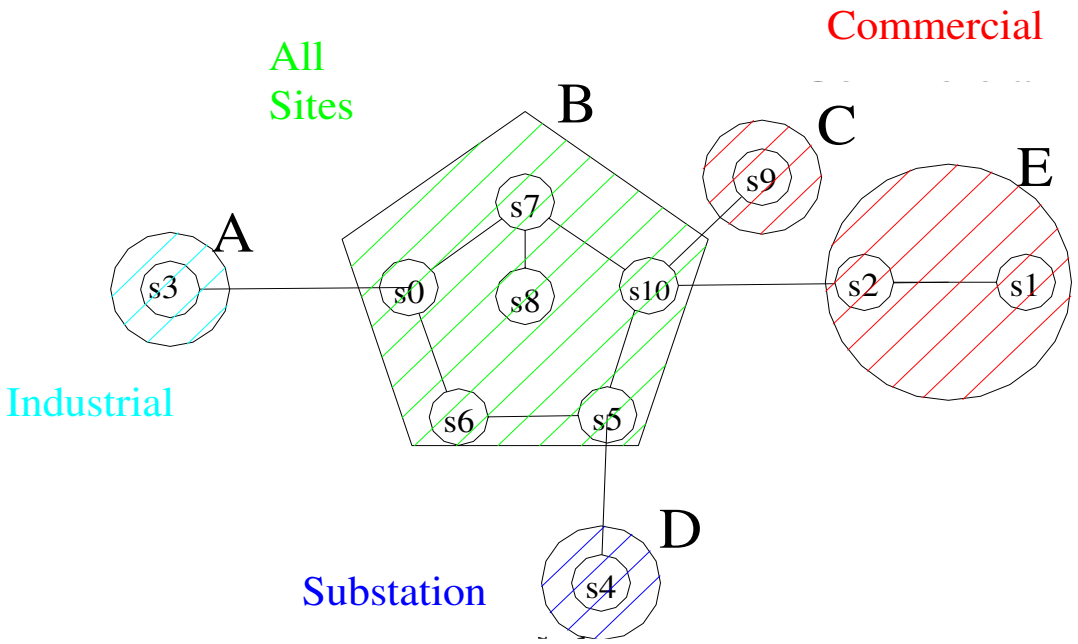

Figure 6.6: Super-group abstraction by MDS.

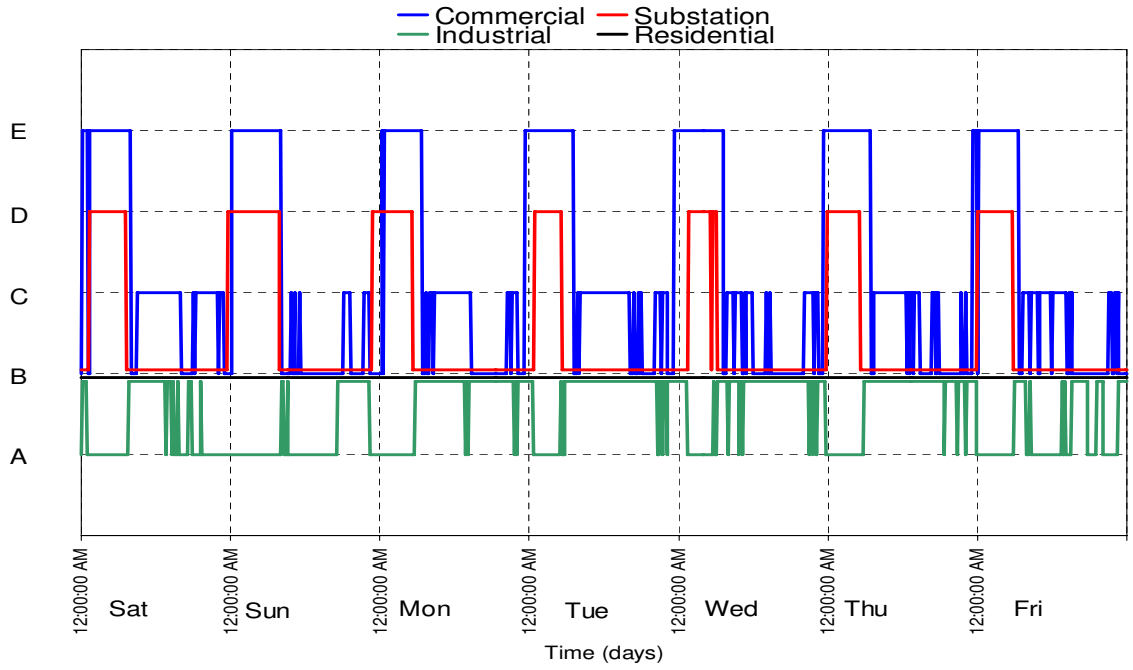

Figure 6.7: Super-groups in all sites over one week.

6.5 Decision tree of supervised learning

Decision trees are one example of the supervised learning techniques such as neural network or Bayes classifiers. With decision trees, a model is generally automatically induced or built, based on some guiding information theoretical metric such as entropy gain [22] proposing plausible relationships between the input data (training set) and the classes, here being the super group labels, or indeed cluster labels obtained from the unsupervised learning of MML. Once the model is trained with sufficiently good accuracy, it can then be applied to another data set (test data) often having unknown classes. In doing so, one is able to predict which class each data point in the test data set best belongs to. An optimum model is the one that has low error rates in each of the above two steps. In order to obtain high accuracy in the first step (training step), a large decision tree might be generated; however, this level of accuracy might not be sustained in the second step (test step). In addition, a large (bushy) tree might be difficult to interpret. Pruning is considered to be one solution to reduce the size of a tree.

In this thesis, the supervised learning C5.0 algorithm [81] is used to carry out the supervised learning process. The C5.0 algorithm is an advanced supervised learning tool with many features that can efficiently induce plausible decision trees and also facilitate the pruning process. The resulting models can either be represented as tree-like structures, or as rule sets, both of which are symbolic and can be easily interpreted. The usefulness of decision trees, unlike neural networks, is that it performs classification without requiring significant training, and its ability to generate a visualized tree, or subsequently expressible and understandable rules. Once trained, the decision tree or rule set obtained can then be used to infer or classify which cluster any new data belongs to. Depending on the type and amount of noise (uncertainty) within the data set, various problems may arise wherein a large number of attributes are utilised in the classification of several classes. The resulting decision tree may often be very large for humans to easily comprehend as a whole. The solution to this problem is to transform the class attribute, of several possible alternative values, into

a binary set including the class to be characterised as first class and all other classes as second class. In this research work, Clementine [82], an integrated data mining work bench is used to carry out various data processing and management tasks, and the supervised learning C5.0 algorithm.

6.6 Rules discovered from the super-groups using decision tree

In order to gain a closer insight into the abstracted super-groups obtained in Section 6.4.3, in particular, how they differ from each other, the supervised learning C5.0 tool was applied to the measured data set which had been augmented with the super-group labels derived from the MML and MDS processes of Section 6.4.3. The benefit of most supervised symbolic learning techniques, and in particular the C5.0 algorithm, is that it can be used to both describe and predict classes. The symbolic outcomes are represented either as decision trees or sets of textual "if ... then" rules, that require little computation.

The generated rules describing each super group are shown in Table 6.2. By examining the nine generated rules that identify various levels of harmonic disturbance, the power utility engineer can more readily deduce the type of power quality event that may be associated with these rules. The conditions of Rule A1 for example, which is the first rule for super-group A, occur at the industrial site late at night and early in the morning. In particular, this rule identifies conditions involving a high level of third harmonic current and seventh harmonic voltage for phase A. This is an understandable consequence of the operation of single phase harmonic producing equipment, such as small AC motors during the night time shift. The third harmonic current is usually blocked from flowing upstream due to the presence of Δ/Y transformers downstream, whereas the high seventh harmonic current is assumed to flow in the substation site and causes a high seven harmonic voltage. To confirm this assumption, the seven harmonic current at industrial site of phase A is plotted along

Table 6.2: Generated rules from super groups (A to E) using the C5.0 algorithm.

Rules for A contains - 2 rules	Rules for B contains - 4 rules	Rules for C, D and E contain - 1 rule each
Rule 1 for A: if CT1 Harm 3 > 0.808 and CT1 Harm 5 <= 0.694 and CT1 Harm 5 <= 0.694 then A Rule 2 for A: if CT1 Harm 3 > 0.808 and CT1 THD > 0.608 and CT3 Harm 7 <= 0.414 and CT3 THD <= 0.604 and PhA THD > 0.768 then A	Rule 1 for B: if PhA Harm 5 > 1.487 then B Rule 2 for B: if CT1 Harm 3 <= 0.808 and CT1 THD > 0.237 and PhC Harm 7 <= 0.851 then B Rule 3 for B: if CT3 Harm 5 > 0.138 PhC Harm 3 <= 0.510 then B Rule 4 for B: if PhC Harm 3 > 0.510 then B	Rule 1 for C: if CT1 THD <= 0.237 and CT3 Harm 7 <= 0.318 and CT3 THD > 0.180 and CT3 THD <= 0.281 and PhA Harm 5 <= 1.487 and PhA THD > 0.849 and PhC Harm 7 <= 0.851 then C Rule 1 for D: if CT1 Harm 3 <= 0.808 and CT3 Harm 5 > 0.564 and CT3 Harm 7 <= 0.430 and PhA Harm 7 > 1.635 then D Rule 1 for E: if CT3 Harm 5 <= 0.138 and PhC Harm 3 <= 0.510 then E

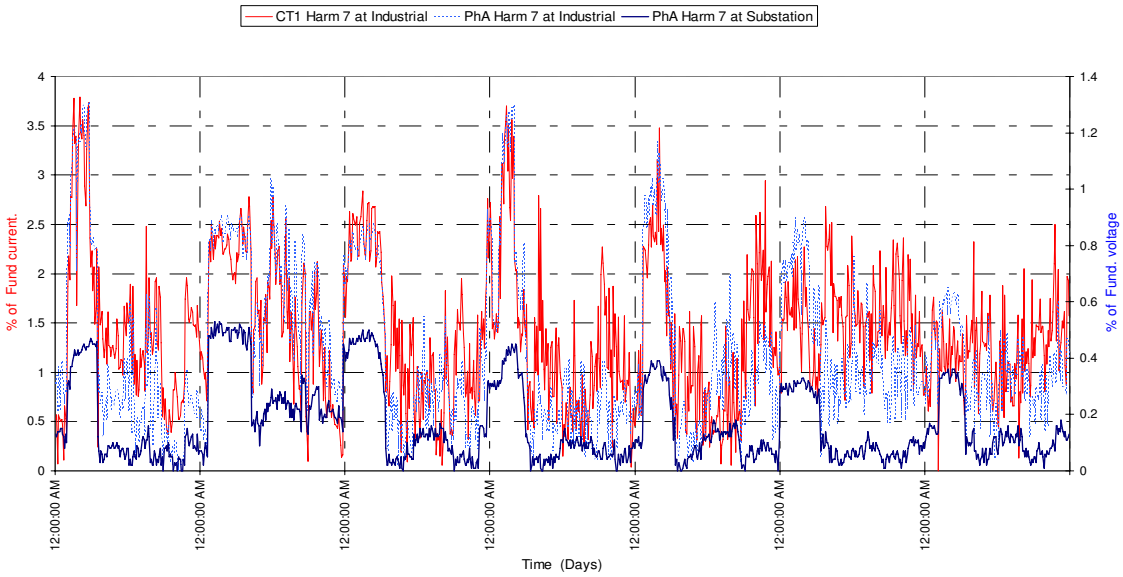

Figure 6.8: High 7th harmonic current at industrial site causing high 7th harmonic voltage at substation.

with the seventh harmonic voltage of the same phase at substation site in Figure 6.8 which shows that the seventh harmonic current is well correlated with the seventh harmonic voltage at that site. Further, rule A1 at the industrial site is partly synchronized with the rule D1 of substation site which has high seven harmonic voltage as part of its rule. There are other conditions in rule D1 that need to be fulfilled from phase C, such as an average or median value (50%) to high fifth and low seventh harmonic current. When these exist then the super group state D, (rule D1) is seen to be synchronised with the super group state A (rule A1) on Thursday, Friday and Saturday which is shown in Figure 6.9. Another application of these generated rules is that of rule B1 of super-group B, which is ultimately formed from clusters containing data from all sites. Rule B1 is characterized by high fifth harmonic voltage which occurs simultaneously at all sites except for the residential site, as illustrated in Figure 6.10. This means that there is an interaction between these sites, i.e. one or more sites causing the high fifth harmonic voltage to occur at the other sites. The suggested likely source is the commercial site where the conditions for Rule B1 are be-

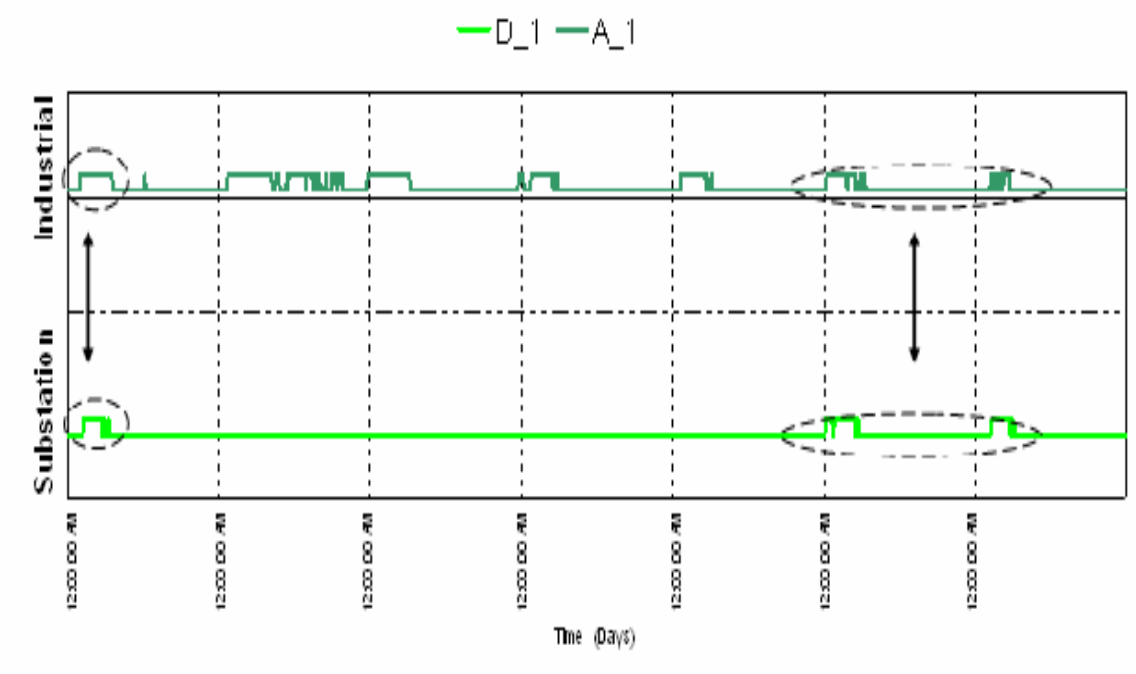

Figure 6.9: Rules A1 and D1 are synchronised on Thursday, Friday and Saturday at the industrial and the substation sites in one week time frame.

ing fulfilled, or that it is "firing", more frequently. This would be partially explained by the large number of fluorescent lamps generally found in commercial areas. Evidence of the B super-group is also noted in the industrial site as the trend of B1 in Figure 6.10, however, this appears to occur as a comparatively reduced frequency.

6.6.1 Visualisation of the the super-groups generated rules

With the aid of visualisation techniques, some of the above mentioned rules can be visually clarified. Rule A1 at industrial site which is characterized by high third and fifth harmonic currents together with a high harmonic voltage in the following rule:

Rule 1 for A:

if CT1 Harm 3 $>$ 0.808, and

CT1 Harm 5 $<=$ 0.694, and

PhA Harm 7 $>$ 1.293

then A

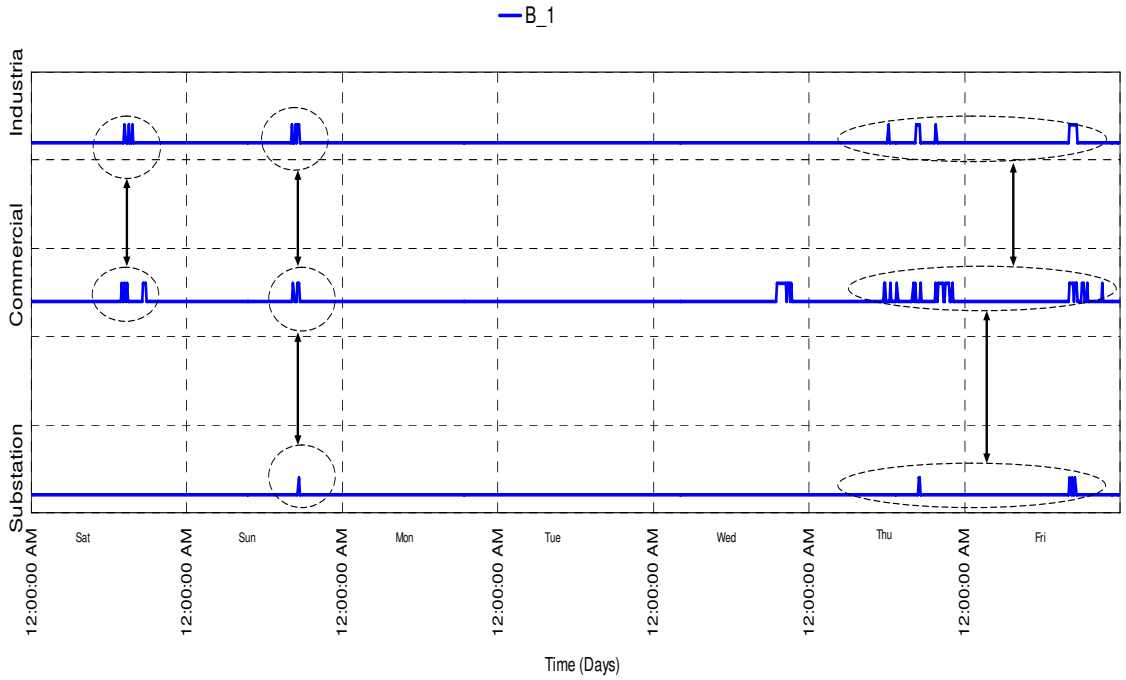

Figure 6.10: Evidence of Fifth harmonic producing loads at phase C due to commercial site.

This is visualized by plotting the three attributes together in 3-D graph for one week data from the industrial site as shown in Figure 6.11. A proportion of the instances associated with the super group A can be visually identified as red dots defined by the conditions of Rule A1. Indeed as super group A is defined by two parallel hypotheses (or two rules) the alternative Rule A2, would also cover a number of the remaining data instances (blue dots) shown in Figure 6.11. However, these would be circumscribed by additional or differing dimensions such as CT1 THD, CT3 Harm 7, CT3 THD and PhA THD, and can not be readily represented in the same three-dimensional space. Alternative views of rules can also be useful. For example, Rule B1 which is one of the four parallel hypotheses defining states of super group B, identifies conditions where the fifth harmonic voltage is high at all sites in the rule:

Rule 1 for B:

if PhA Harm 5 $>$ 1.487

then B

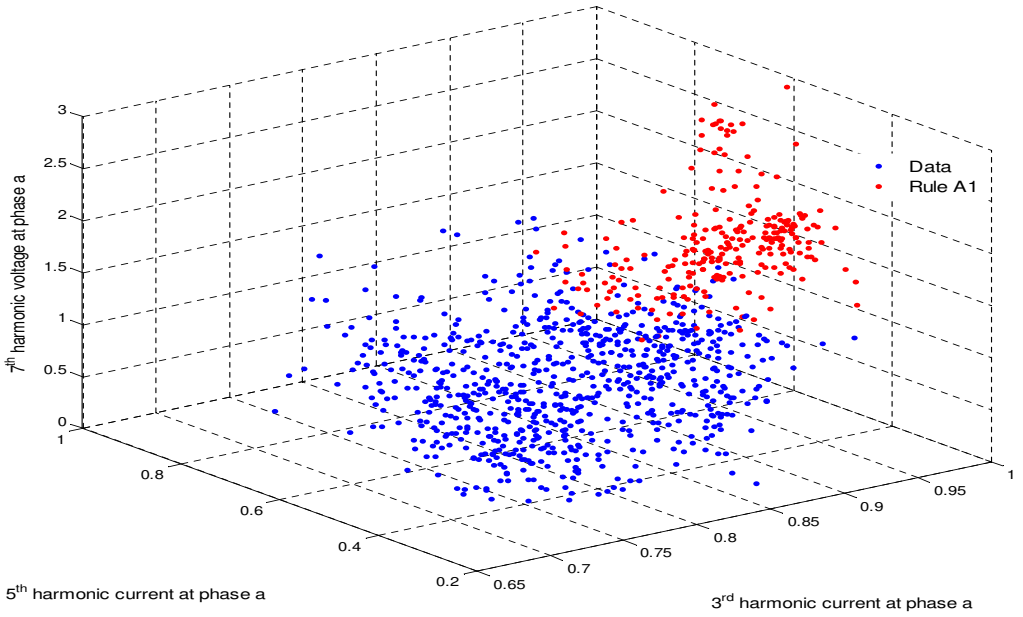

Figure 6.11: Visualization of Rule A1 at the industrial site for one week data.

This is visualised at one typical site (commercial site) by plotting the fifth harmonic current against the time for one week as can be seen from Figure 6.12. As can be seen from Figure 6.12 the red dots (rule firing of B1) are at the highest level of the fifth harmonic voltage which indicates that rule B1 is associated with the worst 5th harmonic voltage. Although the limits of the harmonic standards (IEC61000-3-6) have not been exceeded by this rule, investigation on future data is still needed to verify that the harmonic level of this rule is not increased due to seasonal change of data or other reasons. As a further visualization example, the E super group is completely defined by a single hypothesis Rule E1. E super-group was formed by joining the elements of the s1 and s2 MML clusters, as shown in Figure 6.6. The harmonic instances that are explained here are different from the three rules associated with A1 and B1 in supeer-groups in that they are described by only two rules, a low fifth harmonic current and a low third harmonic voltage at phase C. An example of this concept is illustrated from the commercial site by the rule:

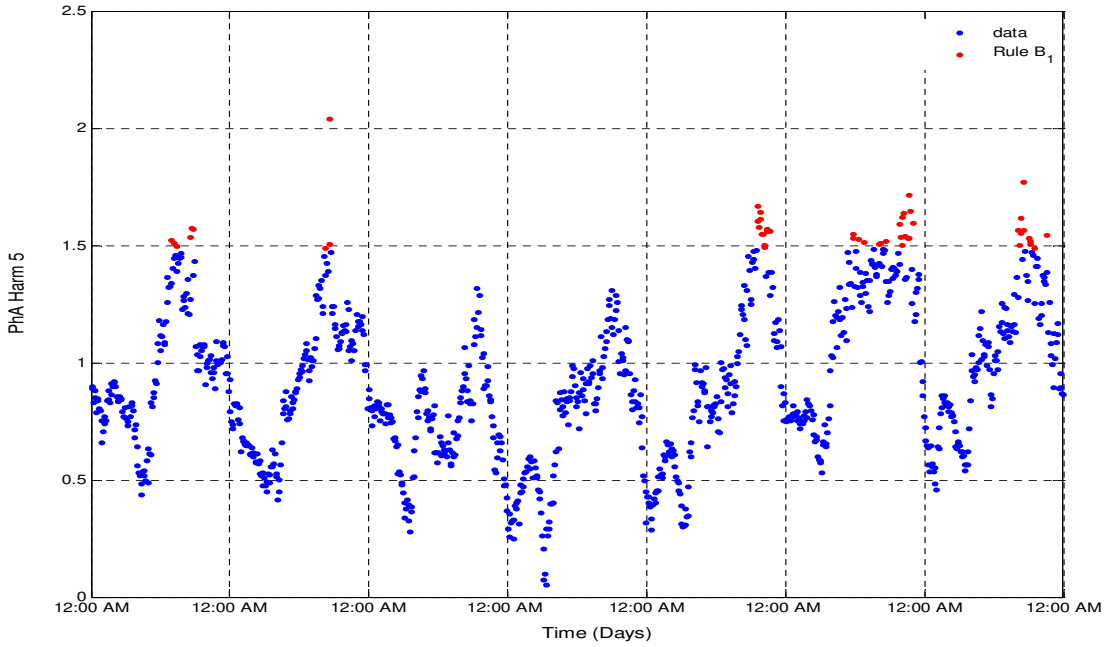

Figure 6.12: Visualisation of Rule B1 at commercial site for a one week period.

Rules for E1:

if CT3 Harm 5 <= 0.138, and

PhC Harm 3 <= 0.510

then E

This rule can be visualized using the scatter plot of the two variables involved in the rule as shown in Figure 6.13. It can be seen from Figure 6.13 that the red dots lie in the lower left corner of the graph, and hence this rule is associated with events with low 3rd and 5th harmonics.

6.7 Summary

Harmonic data from an MV/LV distribution system containing residential, commercial, and industrial customers has been analysed using data mining techniques. Both unsupervised learning (clustering) and supervised learning (classification) techniques

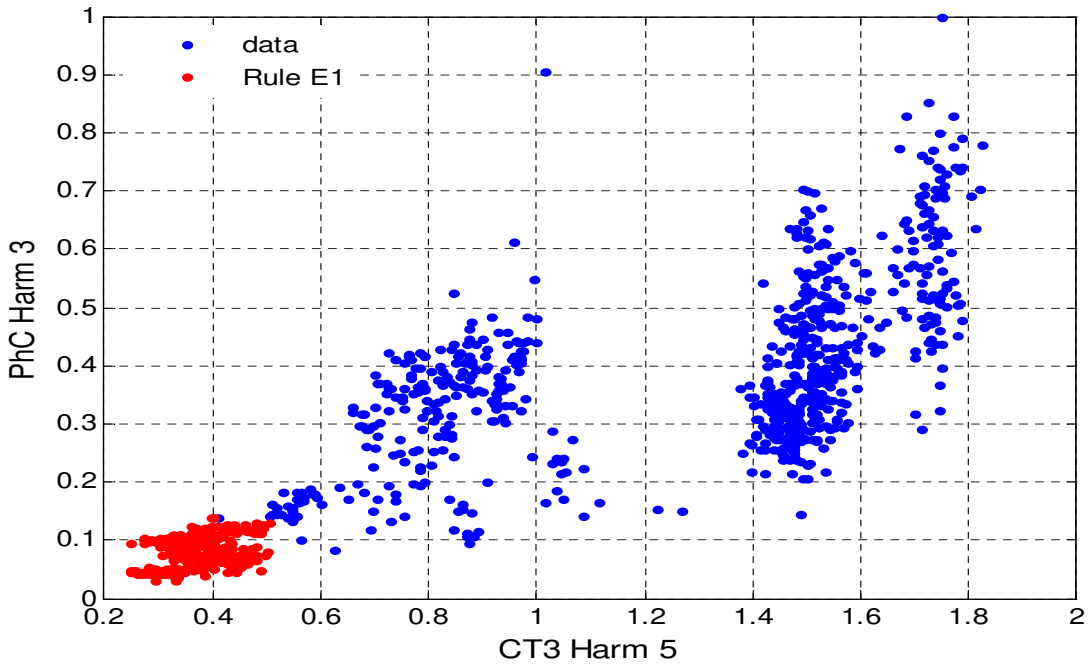

Figure 6.13: Visualization of Rule E1 at commercial site for one week period.

of data mining have been shown to be able to identify useful patterns within the harmonic data set. From the clustering process of the harmonic data, significant results have been obtained including:

1. Verification of a strong relationship between 5th harmonic and THD levels for current at customer sites.

2. Significantly high harmonic disturbance levels usually only occur for short periods of time (low abundance).

3. Identification of "footprints" of the overall system harmonic distortion, as well as residential, commercial and industrial customer harmonic emissions.

4. Abstraction, and detection of super-groups from the harmonic data, each of which comprise similar clusters based on the Kullback Leibler Distance (KL) between the clusters.

Using symbolic decision tree supervised learning techniques, expressible and understandable rules have been generated by which the characteristics of each super group can be identified. The effects of the different sites (Substation, Residential, Commercial and Industrial) on each other as well as the causality of unwanted distortion were discovered using these classification rules. Link analysis and visualisation techniques are also data mining tools that can assist in discovering useful patterns and relationships that are present in harmonic data sets.

Chapter 7

Harmonic event detection using supervised and unsupervised learning

7.1 Introduction

In this chapter, the harmonic measurement data from the power system in Australia is classified using unsupervised learning (clustering) based on mixture modelling using Minimum Message Length (MML) technique. By observing how the data has been classified, the engineers have at their disposal a visually oriented method of evaluating the underlying operational information contained within the clusters. Once the data has been classified as clusters, a supervised learning tool based on C5.0 algorithm, is then used to describe the essential influences/factors that form the clusters and to predict the occurrences of unusual clusters in future measurement data.

7.2 Results from unsupervised learning using MML

ACPro was repeatedly applied to the measured harmonic data from the monitoring program in order to explore a range of plausible models. The program was controlled to produce a series of models each with an incrementally increasing number of clusters for the same fixed values of Aom, and the message lengths of these models have been plotted against the number of clusters as shown in Figure 7.1. This was undertaken so that the message length criterion of the MML could be directly utilised to select the model (number of clusters) that best represent the measured harmonic data. The smaller the encoded message length the better the model fits the data.

The three attributes (fundamental, 5th and 7th harmonic currents) were selected from different sites (sites 1, 2, 3 and 4). The 3rd harmonic current was excluded as its level was low due to the presence of Δ/Y transformers downstream, which block most of the 3rd harmonic current from flowing up stream. The chosen data was initially normalised to the range (0 1) and then used as the input to the ACPro software together with an appropriate accuracy of measurement (Aom) that is calculated from Equation 5.1. It can be observed from Figure 7.1 that the best model to represent the measured haramonic data was identified as that with six clusters. The reasoning behind selecting this number of clusters is that the decline in the message length significantly decreases until the model size reaches 6 clusters, and the message length is comparatively constant afterward as shown in Figure 7.1. In other words, this can be considered to represent the first point of minimum sufficiency for the model.

Using a basic spreadsheet tool the clusters are subsequently sorted in ascending order (s0, s1, s2, s3, s4 and s5) based on the mean value of the fundamental current, such that cluster s0 is associated with the off peak loads period whilst cluster s5 related to the on-peak load periods as shown in Figure 7.2. The mean value (μ) of the fundamental, 5th and 7th currents along with the standard deviation (σ) and the abundance (π) of each model cluster is given in Table 7.1. Each generated cluster can therefore be considered as a profile of the three variables (fundamental, 5th and 7th harmonic currents) within an acceptable variance. If new data lies beyond this

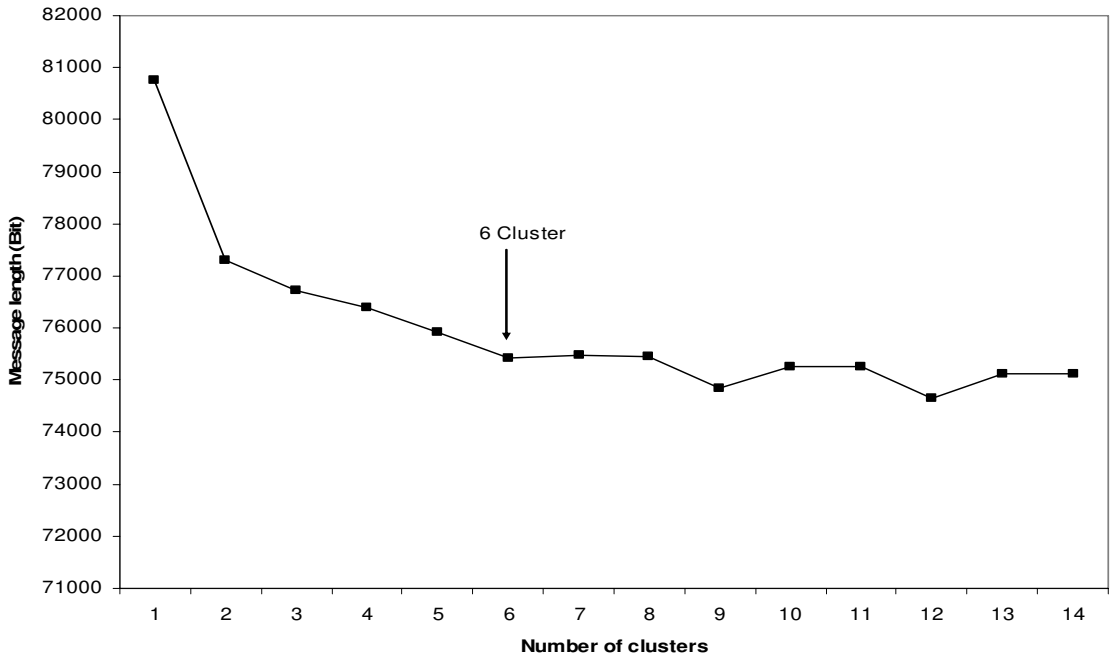

Figure 7.1: Message length vs. increasing mixture model size (number of clusters).

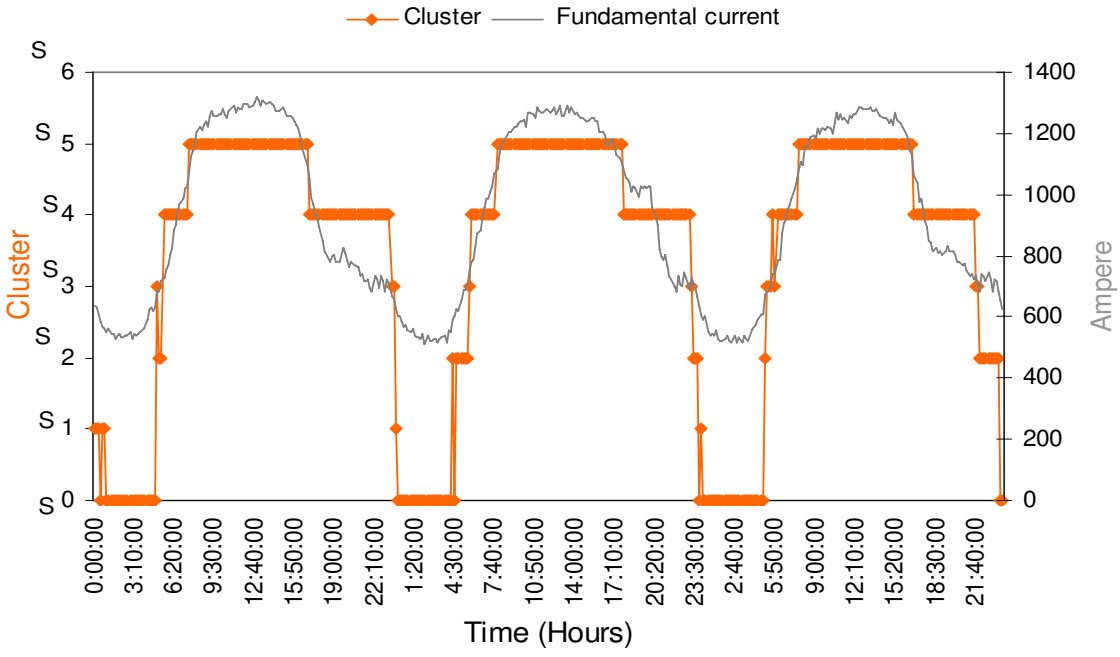

Figure 7.2: Abundance, mean and standard deviation for each cluster of 5th harmonic current per phase.

Table 7.1: Generated model detailing the abundance value (π) of the six cluster a long with the mean (μ) and standard deviation (σ).

Cluster	Abundance	Fund. current		5th Harm. current		7th Harm. current	
(N)	(π)	(μ)	(σ)	(μ)	(σ)	(μ)	(σ)
s0	0.06838	0.09657	0.04194	0.16586	0.13098	0.06293	0.02288
s1	0.15561	0.10610	0.06153	0.44505	0.12335	0.25080	0.12777
s2	0.05677	0.16940	0.09343	0.30038	0.14996	0.11521	0.02859
s3	0.09099	0.35053	0.13280	0.30837	0.12079	0.33083	0.14232
s4	0.34265	0.38735	0.12375	0.52437	0.19318	0.60431	0.18195
s5	0.28555	0.72860	0.09522	0.52180	0.19172	0.51690	0.14954

variance, additional clusters are created until all of the data is enclosed within the generated model as shown in Figure 7.3. Despite the cluster labels having no specific meaning when initially generated, one can appreciate the benefit of their visual profiles in conjunction with previous sorting process as shown in Figure 7.3, in particular one can see that cluster s5 not only has the highest fundamental current, but also has high 5th harmonic current as in cluster s4. From Figure 7.2, one can infer that the high 5th harmonic currents are due to the following events: (1) ramping load (decreasing or increasing) associated with cluster s4, and (2) on-peak load associated with cluster s5.

The occurrence of these six clusters, at sites 1-4 for a period of two weeks, are visualised as a scatter plot in Figure 7.4(b). Here, each data instance is portrayed as a coloured marker associated with its cluster type, and plotted according to the values of its 3 attributes. Superimposed in this same projection are the corresponding representation of the statistical distribution of the mixture model shown as ellipsoids with centres ($\mu5$, $\mu7$, $\mu1$) and radii ($\sigma5$, $\sigma7$, $\sigma1$) as given in Table 7.1. For clarity, these ellipsoids are shown separately in Figure 7.4(a). Here, each ellipsoid is also represented with a variable shade of grey scale in proportion to the abundance of

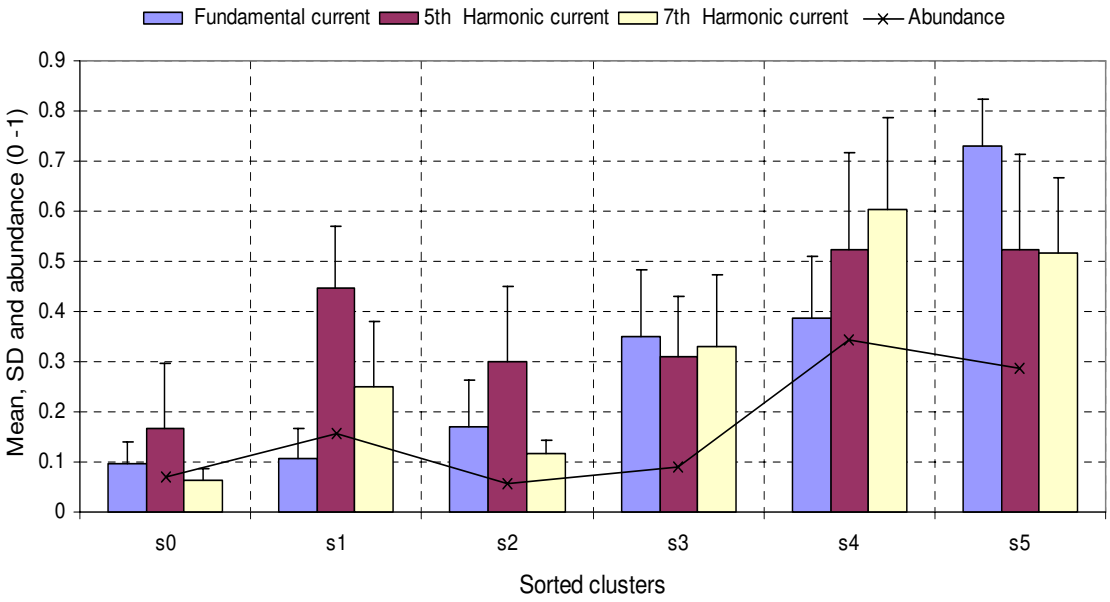

Figure 7.3: Graphical profile view of model clusters indicating the statistical parameters mean (μ), standard deviation (σ) and abundance (π).

that cluster, i.e. the least abundant cluster will appear the darkest, and the most abundant cluster will have the lightest shade of grey. Essentially the volumes of each ellipsoid approximately represent the 3-dimensional spread the cluster data for one standard deviation of its membership (68%). This will only be true of course for purely Gaussian distributions for each attribute. From Figure 7.4, it can be seen for this particular data, that the mixture modelling using MML has generated distinct clusters since each cluster is distributed around its ellipsoid with minimal overlapping with other clusters. The abundance of each cluster is identified by the grey scale colour of the associated ellipsoid. For example, the ellipsoid of cluster s2 is the darkest one and hence it is the least abundance cluster compared to cluster s0 being similar but slightly larger in abundance as seen in Table 6.1. Qualitative values for the harmonics (low, medium, high) can be used to characterise the location of the various ellipsoids as well as the data set around it. Cluster s5 (on-peak), for example, has the highest fundamental, high 5th and 7th harmonic current as opposed to cluster s0 (off-peak) which has the lowest fundamental, 5th and 7th harmonic current.

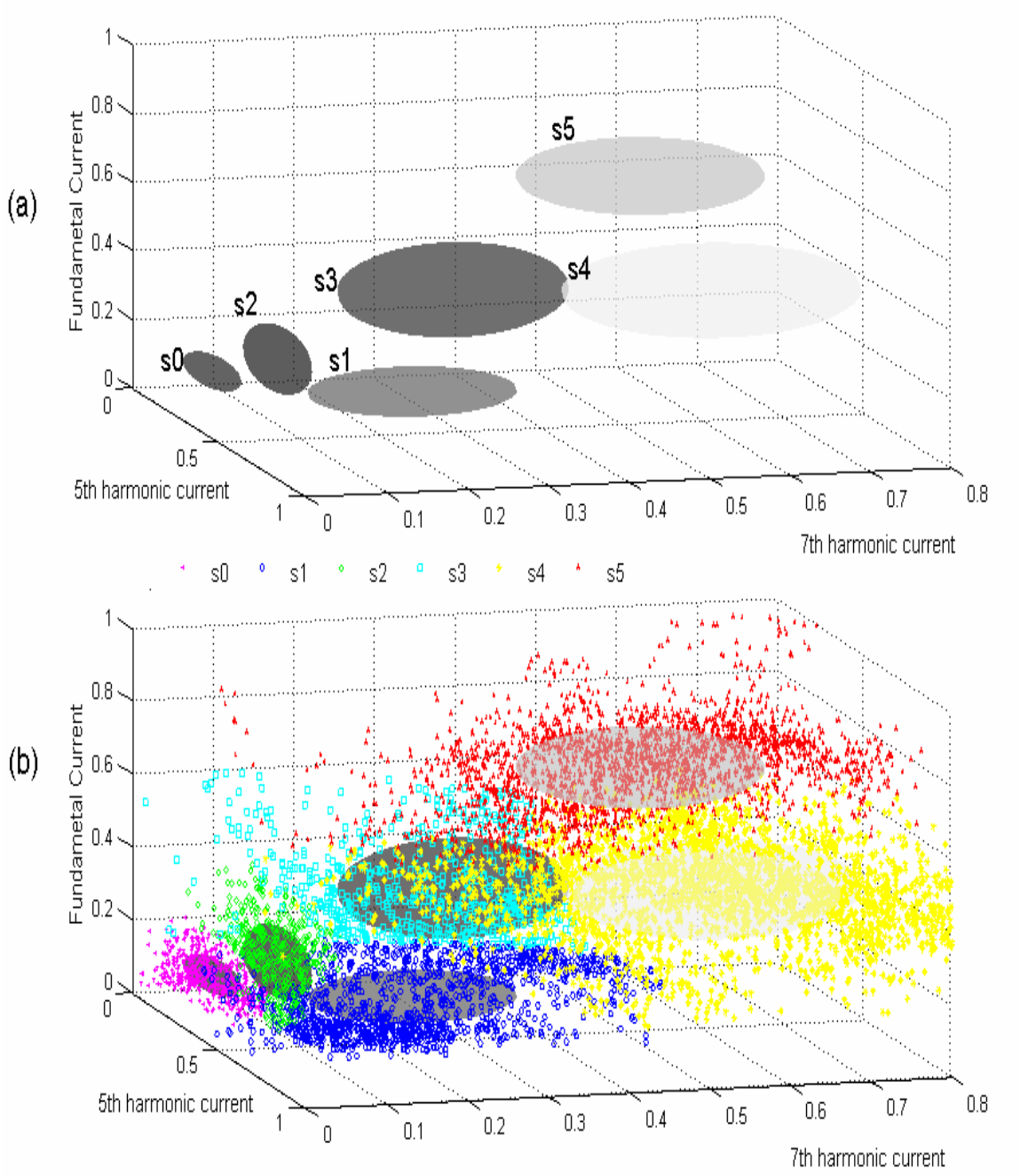

Figure 7.4: (a) Model of six Gaussian distribution clusters obtained at sites(1-4) and (b) The data fitted to the model.

7.2.1 Interpretations of the generated clusters

By observing how the measured data are classified into various clusters, the power utility engineer can more readily deduce the power quality event that may have triggered a change from one cluster to another cluster. To confirm the observation, other available data available to the engineers can be used, such as temperature and reactive power measurements or by discussion with the system engineers or operators. Further, visualising these data along with the obtained clusters at each individual site (site 1-4) can provide useful information that enables the utility engineer to understand the causes and effects of the harmonics obtained and to predict future events. Generally by examining the behaviour of MML model classifications (of the recorded data) one is able to attribute further meaning to each of its cluster components. For example, it is noted that there are several sudden changes to cluster s2 at particular time instances during the day. Figure 7.5(a), illustrates the trend of clusters obtained from substation site (Site 1) superimposed on the fundamental current measurement data for two days. For the same period Figure 7.5(b), shows the 7th harmonic current and 7th harmonic voltage at the substation.

By observation, it appears that the sudden changes of cluster s2 is due to causal changes in the 7th harmonic current. After further investigation of the reactive power MVAr measurement at the 33kV side of the power system shown in Figure 7.5(c), it can be deduced that the second cluster (s2) is related to a capacitor switching event. Early in the morning, when the system MVAr demand is high as shown in Figure 7.5(c), the capacitor is switched on in the 33kV side to reduce bus voltage and late at night when the system MVAr demand is low, the capacitor is switched off to avoid excessive voltage rise. By just observing the fundamental current, it is difficult to understand why the second cluster has been generated. The 7th harmonic current and voltage plots as shown in Figure 7.5(b) provide a further clue that something is happening during cluster s2, in that the 7th harmonic current increases rapidly and 7th harmonic voltage decreases, although by observing just two plots the reason is still unknown, only by having expert opinions of the operation engineers and Figure 7.5(c)

can one understand why such cluster is generated. In this case, the clustering process

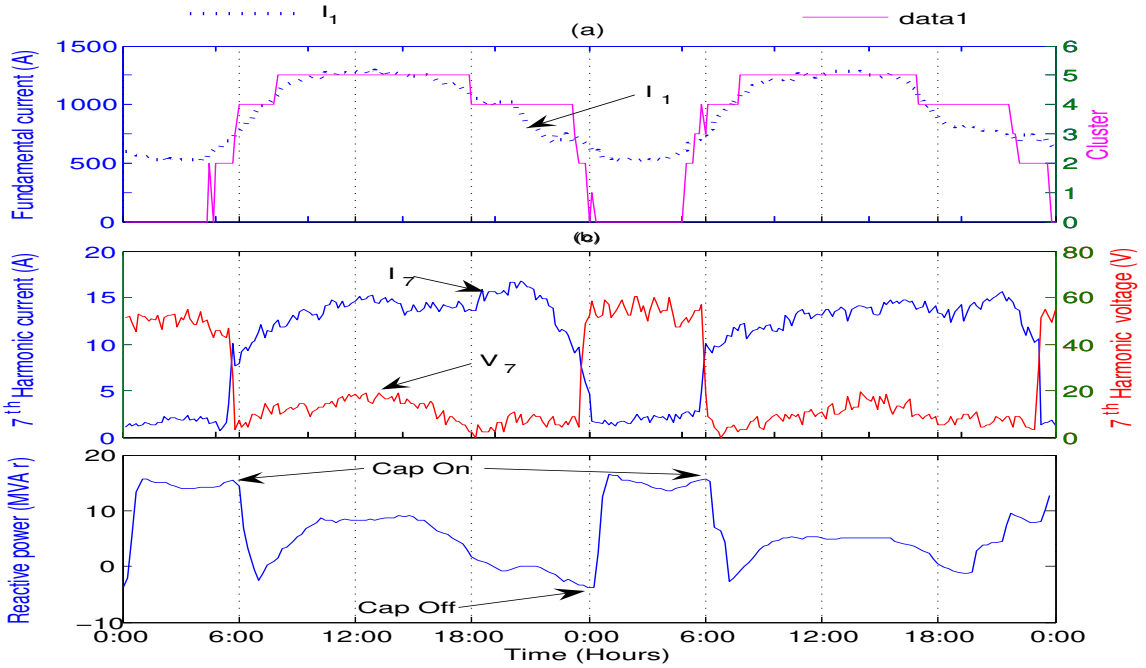

Figure 7.5: Clusters at substation site in two working days (a) Clusters superimposed on the fundamental current waveform, (b) 7th harmonic current and voltage data. (c) MVAr load at the 33kV.

correctly identified this period as a separate cluster compared to other events, and this can be used to alert the power system operator of the need to understand the reasoning for the generation of such a cluster, particularly when considering the fact that the abundance value for s2 is quite low (5%). When contacted, the operator identified this period as a capacitor switching event which can be verified from the MVAr plot of the system (which was not used in the clustering algorithm). The capacitor switching operation in the 33kV side can also be detected at the other sites (sites 2, 3 and 4) at the 11kV side.

Although in this case the cause can be easily uncovered, there may be other cases where the clustering process could identify a cluster that may be associated with detrimental effects to the power system. Subsequent monitoring of such cases, would provide an early warning to the power system operator to its impending occurrence.

This is one of the main advantages of the MML clustering algorithm, in that new clusters can potentially identify unique or different operating conditions based on the different data attributes that are provided to the program (fundamental, 5th and 7th harmonic currents). Once identified, more information can be gathered to deduce the major factors as to why the cluster is formed. The deduction can then be confirmed by discussion with system engineers or system operators. In this way, anomalous cases can be quickly identified and analysed.

The same method of observation can be applied to the other clusters, for example cluster s4 at the residential site is associated with a peak period where high fundamental currents and high 5th harmonic voltage are the characteristics of this cluster. This can be attributed to the use of air conditioners. To demonstrate this, Figure 7.6shows a period of three days during the same time of the year when the temperature is normal for the time of the year. Figure 7.7, shows the results for a period of three days when the weather is very hot, resulting in significant use of air conditioners. Figure 7.7 shows that s4 is generated during the day when it is expected that more air conditioners will be used. There is usually a lag (human response) between the peak temperature and the onset of the peak use of air conditioners, thereby causing a noticeable lag between the peak temperature and the sudden increase in the fifth harmonic as shown in Figures 7.7(b) and (c).

Besides using chronological time domain plots, other types of plots can be used to help the engineer to have better visual understanding of what operating conditions each cluster signifies For example, in the following polar coordinate plots of Figures 7.8and 7.10, the magnitude of the variable of interest is represented by the radius vector of the circle whereas the angle from the x-axis to the radius vector represents the time of the day. In Figure 7.8, the polar plots are drawn for the residential site (Site 2) between the normal weather days and the hot days. It is evident that Cluster s5 (red dots) occurs more often at daytime during the hot period compared to the days when the temperature is relatively mild. From Figures 7.7 and 7.8, it can also be observed that there is a period when cluster s5 (peak load) occurs around

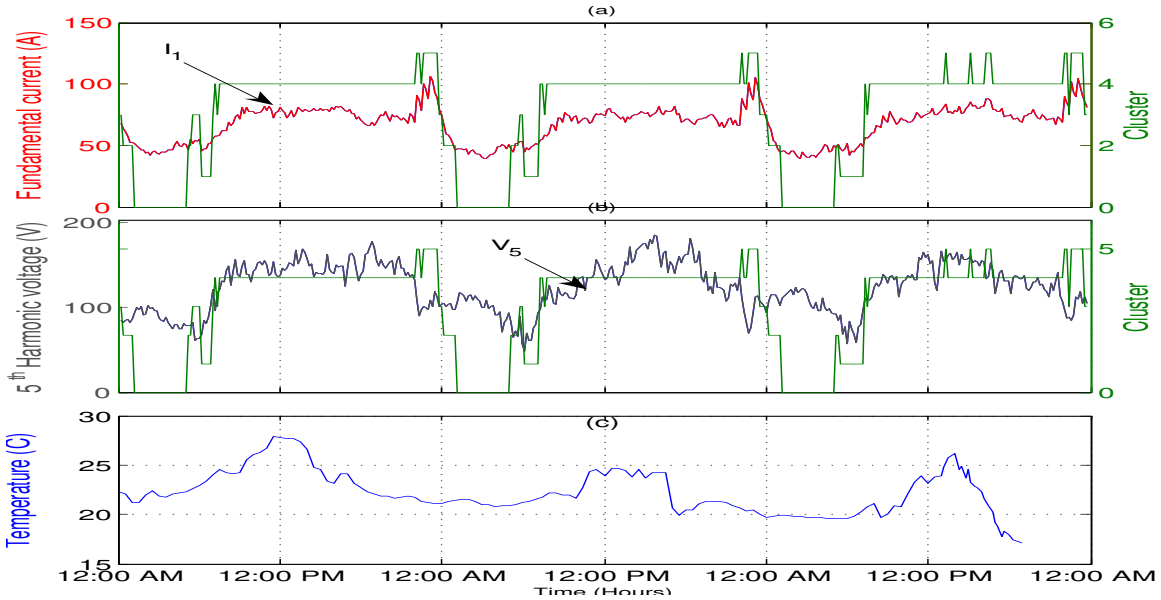

Figure 7.6: Three normal temperature days at the residential site (Site 2), (a) Fundamental current and generated clusters, (b) 5th harmonic voltage and generated clusters, (c) temperature near Site 2.

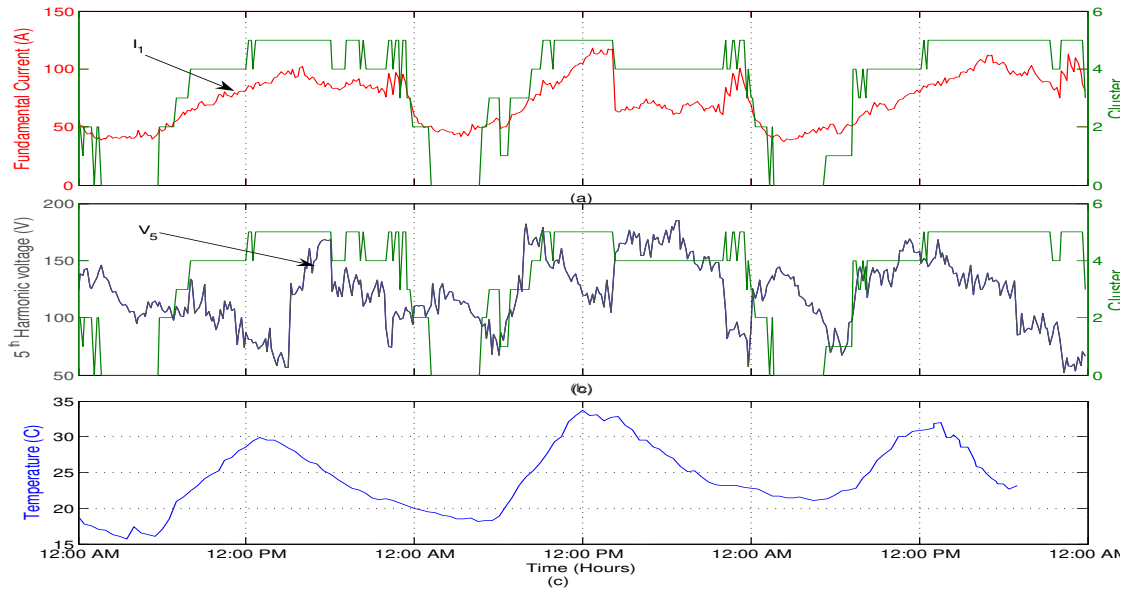

Figure 7.7: Three hot days at the residential site (site 2), (a) fundamental current and generated clusters, (b) 5th harmonic voltage and generated clusters, (c) Temperature near site 2. (c) MVAr load at the 33kV.

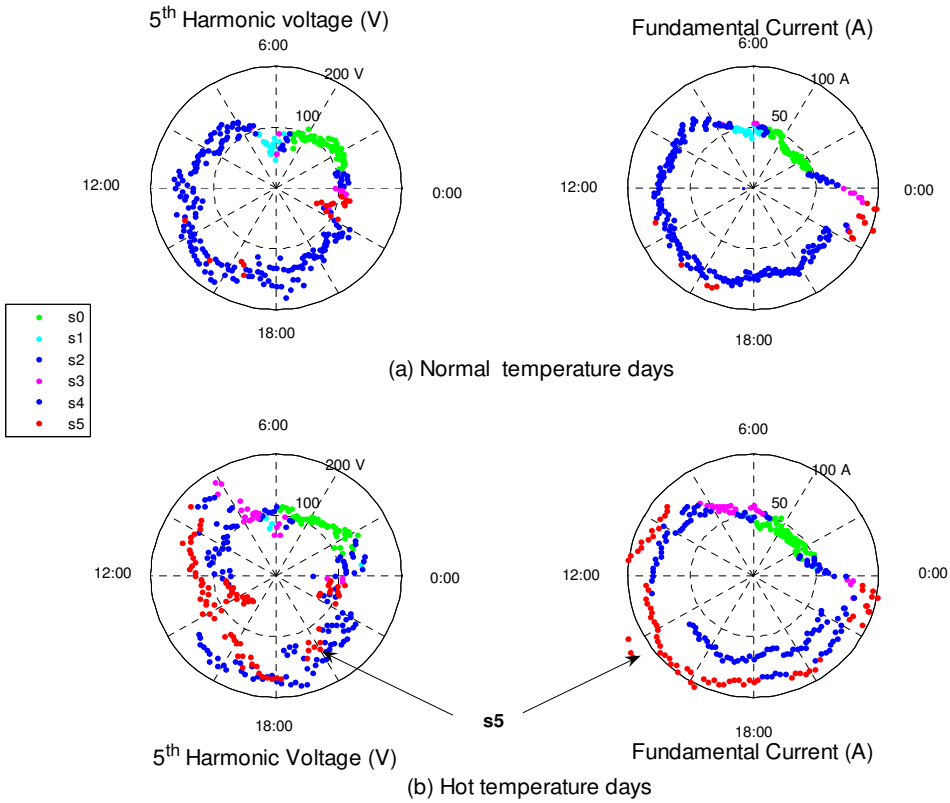

Figure 7.8: Normal and hot days at Residential site (site 2).

midnight, and following discussion with the utility engineer, this is found to be related to off-peak water heaters being automatically switched on at that time.

Another observation of harmonic events can be extracted from Figure 7.9, which shows the 5th harmonic current at industrial site (Site 4) at different days of the week. On Saturday, for example, cluster s5 (represented by red dots) is only present from early morning to early in the afternoon which may indicate an industrial process running during this period. On Sunday however, cluster s5 has disappeared inferring that these loads were off. These loads were on again during the weekday at day and night time showing the long working hours in this small factory at the weekdays. Similar results of the 5th harmonic current can be seen at the commercial site, as shown in Figure 7.10.

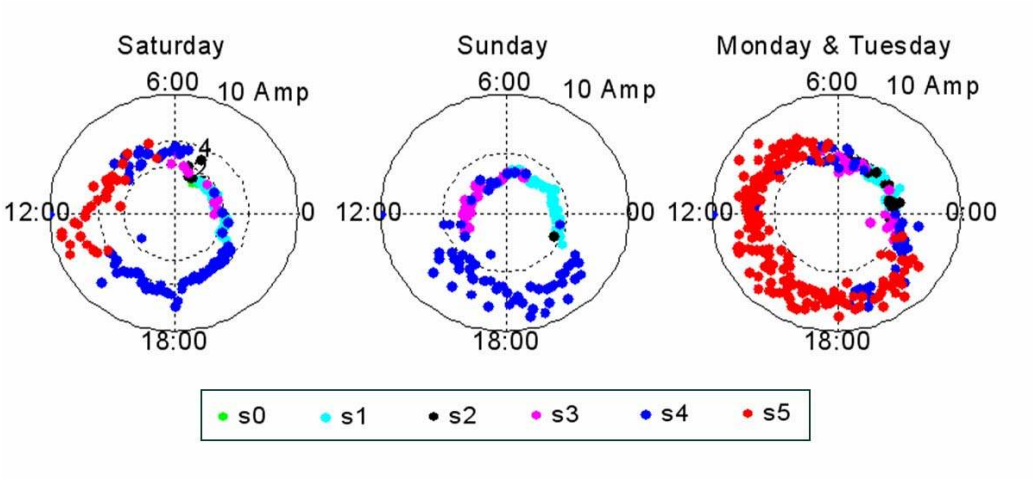

Figure 7.9: 5th harmonic current clusters at industrial site (site 4) for different week days.

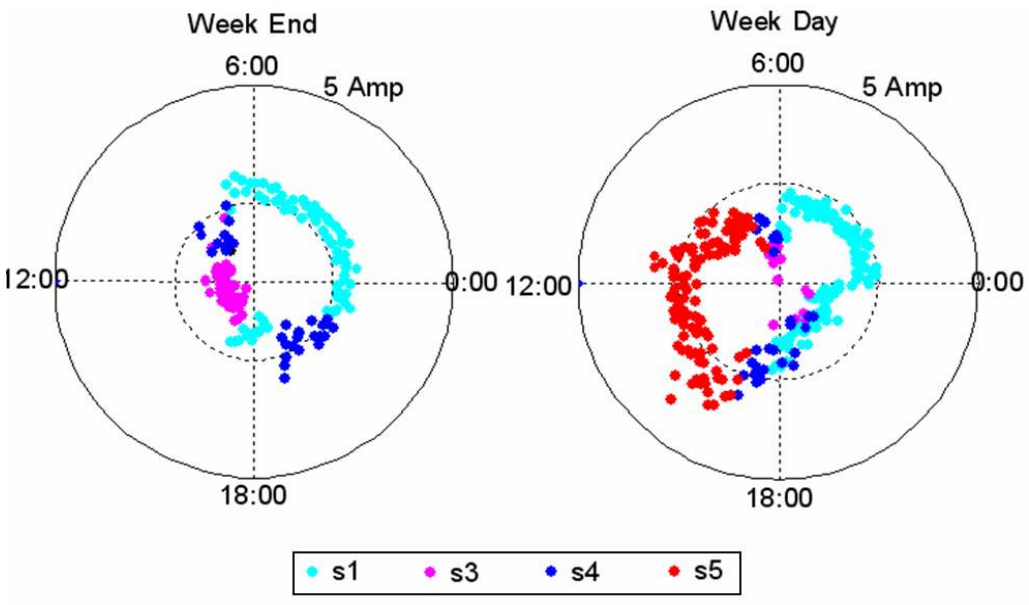

Figure 7.10: 5th harmonic current clusters at commercial site (site 3) for two different week days.

7.3 Results from supervised learning using C5.0

To gain a closer insight into the obtained clusters, and to understand what makes specific clusters differ from each other, the C5.0 classification algorithm was applied to the clusters generated from the MML model from the measured harmonic data. These generated clusters were used as class labels to the input data (fundamental, 5th and 7th currents). Table 7.3 shows a sample of the raw data labelled with the six clusters generated from the MML. The supervised learning algorithm C5.0, explained

Table 7.2: Labelling the data with the clusters produced by the MML.

CT1 Fund.(A)	CT1Harm 5(%)	CT1Harm7(%)	Cluster	Interpretation
63.06992	3.16611	1.829028	s4	On Peak ' evening'
86.07376	2.737146	1.644009	s5	Hot water system
84.398	2.793573	1.628881	s5	Hot water system
79.21832	2.733033	1.505148	s4	On Peak ' evening'
100.1655	2.544205	1.382284	s5	Hot water system
95.976	2.533765	1.334066	s5	Hot water system
86.9116	2.511744	1.329747	s5	Hot water system
104.7357	2.073766	1.246355	s5	Hot water system
99.02288	2.178504	1.257591	s5	Hot water system
93.61472	2.078246	1.20763	s5	Hot water system
83.78864	2.069579	1.131147	s3	Transient
76.01912	1.748439	1.064268	s3	Transient
72.05816	1.729396	0.626906	s3	Transient
64.89808	1.732779	0.648981	s2	Capacitor switching
59.7184	1.666144	0.567325	s2	Capacitor switching
57.28096	1.706973	0.584266	s2	Capacitor switching
49.73992	1.785663	0.586931	s0	Off peak ' evening'
46.3884	1.813786	0.603049	s0	Off peak ' evening'
46.46456	1.826057	0.585453	s0	Off peak ' evening'

in section 6.5 and applied in section 6.6 to extract rules from the super-groups, is reapplied here with the addition of a lagging time window of different ranges of time (30, 60, 90 and 120 minutes) in order to predict the occurrence of the clusters. This results in sets of rules describing each cluster in terms of a minimal (sufficient) set of attributes values from the range of: the fundamental, 5th and 7th harmonic currents, together with the four time lagging windows. This prediction method using the generated rules from C5.0 algorithm is described in the next section.

7.3.1 Prediction of capacitor switching with C5.0 and lagging window

The most important rules are usually the ones associated with the least abundant clusters, as these clusters are considered to be anomalies among other clusters. Clusters s2 and s0, with proportions of 5.6% and 6.8%, respectively, are the least abundant clusters, as shown in Figure 7.3. Cluster s2, as mentioned in Section 7.2.1, is associated with the capacitor switching phenomena (see Figure 7.5) which is worthy of further investigation to determine what would be the previous range of the attributes that could trigger this event to occur.

In order to pursue this, Clementine [82], an integrated data mining workbench which includes an implementation of the C5.0 algorithm, was used in this section to produce the rule set related to this cluster. The discovered rules for cluster s2 with a window size of 60 minutes are listed in Table 7.3. The range of attributes values is (0-1), as explained in Section 5.2.11. The accuracy of the model used to generate these rules was rated high at 98.8%. The quality measure of each rule is described by two numbers (n, m) shown in Table 7.3, in brackets, preceding the description of each rules, where, n, is the number of instances assigned to the rule and, m is the proportion of correctly classified instances. In Table 7.3, the number of data instances, that were predicted by Rule 1 is 286, with 0.934 or 267 being correctly classified. Rule 1 means that if the fundamental current in phase a (C1a) was within the range (13.8% - 37.4%) 50 minutes ago (see Figure 7.11(a)), and if the fifth harmonic current in the same phase (C5a) was in the range (9.5% - 47.4%) 50

Table 7.3: Rules describing cluster s2 generated by C5.0.

Rule Set for s2 - contains 3 rule(s)		
Rule 1 for s2 (286, 0.934)	Rule 2 for s2 (254, 0.898)	Rule 3 for s2 (399, 0.726)
if C1a[-50 min] > 0.138 and C1a[-50 min] <= 0.374 and C5a[-50 min] > 0.095 and C5a[-50 min] <= 0.474 and C7a[-50 min] > 0.069 and C7a[-50 min] <= 0.153 then s2	if C1a[-50 min] > 0.195 and C1a[-50 min] <= 0.374 and C7a[-50 min] > 0.047 and C7a[-50 min] <= 0.174 then s2	if C5a[-50 min] <= 0.382 and C7a[-50 min] > 0.095 and C7a[-50 min] <= 0.174 then s2

minutes before (Figure 7.11(b)), and if the 7th harmonic current (C7a) was between (6.9%-15.3%) 50 minutes ago (Figure 7.11(c)), then s2 will occur.

For example, if we consider a 3 hour period between 3 am and 6 am on 14/01/02 at the substation site, it can be observed that s2 occurs between 4:10 am and 5:40 am (Figure 7.11). At time 4:50 am, we can see that the value of I1 some 50 minutes previously is between 0.138 and 0.374, and I5 is between 0.095 and 0.474 and I7 is between 0.069 and 0.153 and hence we can observe that at time 4:50 am, C5.0 will predict that s2 will be generated as can be observed in Figure 7.11. On the other hand at time 5:50 am, although both I1 and I5 meet the rules associated with these two currents at the previous 50 minutes (i.e. at 5:00 am), I7 does not meet the rules condition because it is not between 0.069 and 0.153 and therefore at time 5:50 am, s2 is not predicted by C5.0, but rather s3 is predicted. The three rules that characterise the capacitor switching (s2) can be visually clarified using the visualisation techniques explained in Section 6.6.1. Figure 7.12 shows the three rules visualised in a three dimensional space.

In each of the views of Figure 7.12 one can observe the (red, black or blue) coloured instances of s2 residing within the rectilinear grey volumes defining the three rules.

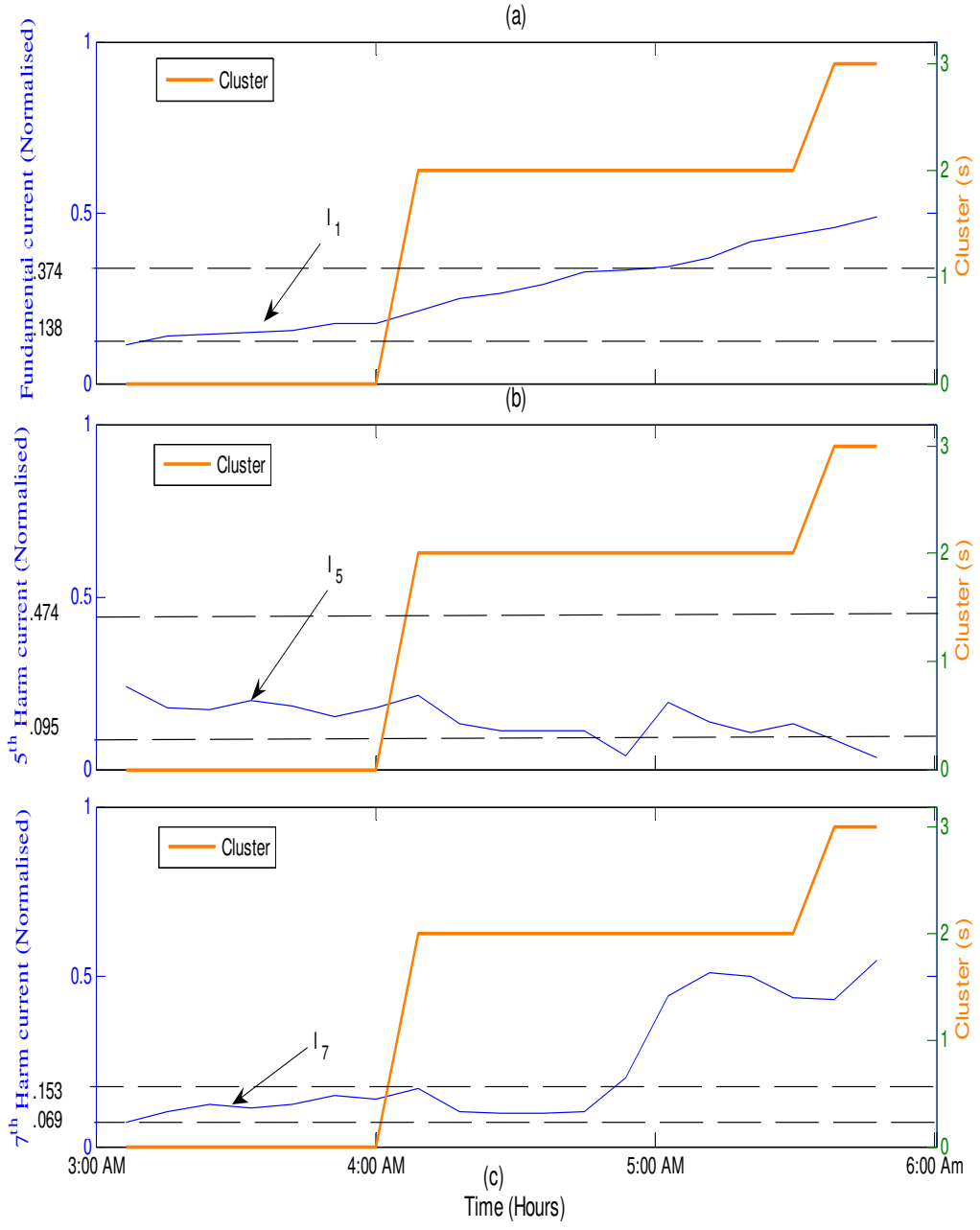

Figure 7.11: Rule-1 of predicting Cluster (s2) of capacitor switching explained in Table 7.3.

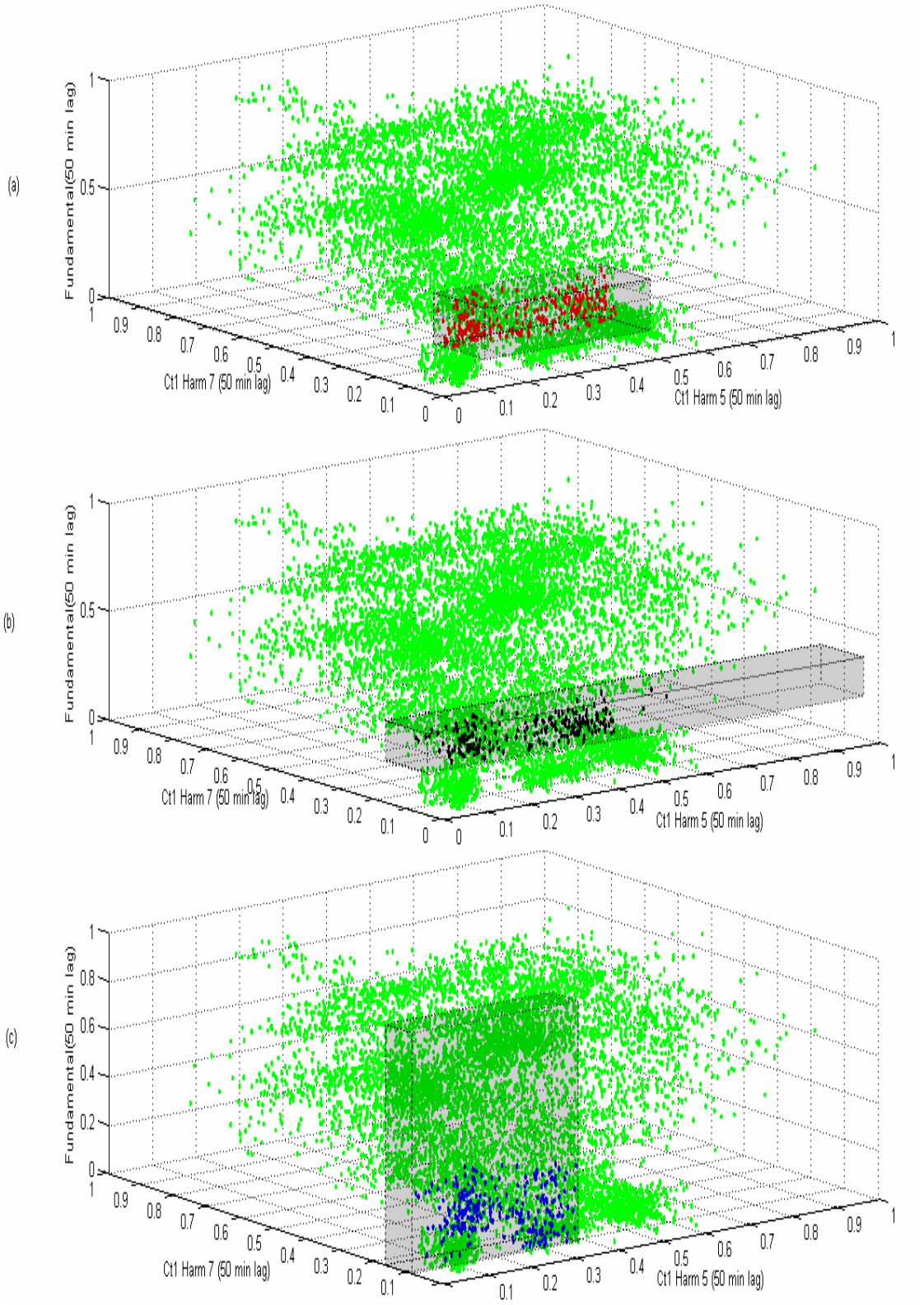

Figure 7.12: Three rules predicting Cluster (s2) associated with capacitor switching events: (a) Rule-1 (b) Rule-2 (c) Rule 3.

The surrounding cloud of other instances (coloured green) resided either behind or in front of these rectilinear spaces.

There are several instances where more than one rule may need to be applied at the same time as can be seen from Figure 7.13. The C5.0 algorithm then applies the rules in the rule set, and makes a majority decision based on whether the cluster is generated or not. This means that one rule is not sufficient to characterise or predict the cluster (or here, the class) and so all of these rules should be considered. From Figure 7.13, there are some instances where only one rule is occurring which means that this rule is able to predict the occurrence of cluster s2, however some instances more than one rule is needed to predict this cluster.

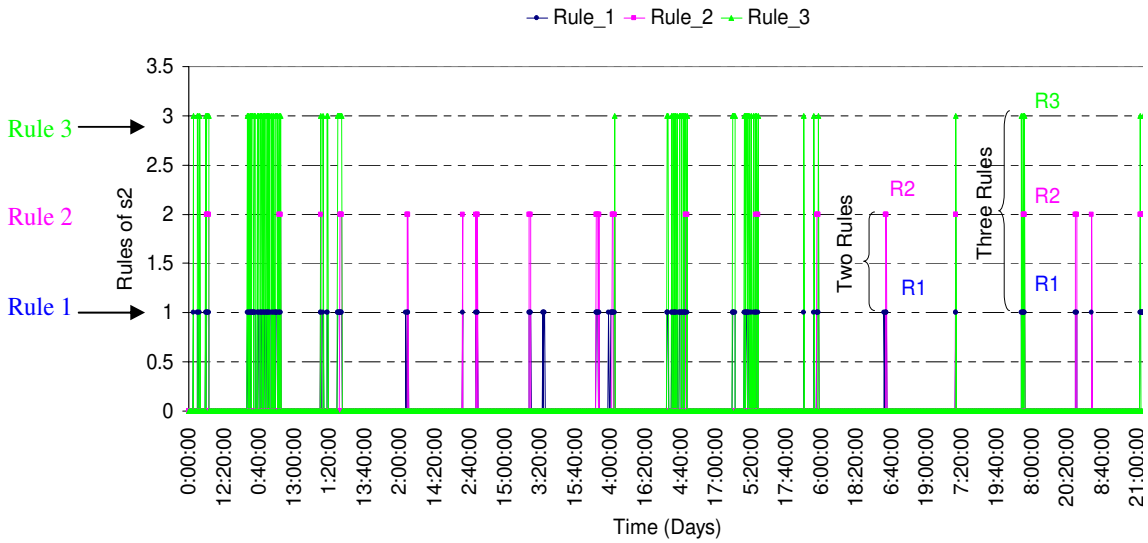

Figure 7.13: Prediction of s2; more than one Rule can occur at the same time instant.

7.4 Summary

Harmonic data from a harmonic monitoring program in an Australian MV distribution system containing residential, commercial and industrial customers has been

analysed using data mining techniques. The technique presented in this chapter allows utility engineers to detect unusual PQ events from monitored sites, using the clusters generated by unsupervised learning, and in particular, cluster analysis using MML, which searches for the best mixture model describing the data using a metric of an encoded message. The obtained clusters can then be characterised by the supervised learning techinque using C5.0 algorithm to infer information about future PQ performance at the monitored sites.

The usefulness of supervised learning using C5.0 is that it can performs the supervised learning without requiring significant training and it has the ability to generate expressible and understandable rules. The main disadvantage of this method is that there are instances where more than one rule needs to be applied at the same time. This may be a consequence of the description language for the domain, in that the combinations of the available attributes cannot efficiently partition the concept space to isolate the concepts of interest. In the case where more than one rule occurs in the same instance when using the C5.0 algorithm then all these rules should be applied entirely as they form a multiple hypothesis set, although the final decision will be made based on the majority decision.

Chapter 8

Determination of the Optimal Number of Clusters in Harmonic Data Classification

8.1 Introduction

In this chapter, the results from applying the novel method to determine the optimum number of clusters using MML technique described in Chapter 4 to the measured harmonic data from the harmonic monitoring system will be presented and discussed. The method is based on the trend of the exponential difference in message length from the MML algorithm. The results confirm the effectiveness of the proposed method.

The results are then compared with the results from the fitness function method [72] described in Section 4.2.2. The results show that the fitness function does not perform as well with the measured harmonic data compared to the proposed method using the trend of the exponential of message length difference. This is because the harmonic data was measured at several points in the network where the attributes at one point are correlated to the same attributes at the other part of the network.

Super-groups formation using link analysis is also used to verify the optimum

number of clusters obtained.

The C5.0 algorithm is subsequently used to uncover the fundamental defining factors that differentiate the clusters obtained from the optimum number of clusters from each other

8.2 Optimal number of clusters in harmonic data

In the previous chapter, the mixture models derived through the MML technique were used to classify the harmonic monitoring data obtained from the Australian harmonic monitoring system into clusters. Each cluster can be considered as a group data in the data set that represents a unique operating condition during the period under study, such as peak loads, off-peak loads, capacitor switching operation etc. Once clustered, the operating conditions associated with each of these clusters can be analysed and deduced from observation. The correctness of the deduction can be confirmed by the operation engineers [67]. In this way, clusters due to power quality issues can be identified quickly and further can be used to identify future occurrence of similar power quality problems. Repeated occurrences of known power quality issues may require countermeasures to be designed to reduce or eliminate the identified power quality problems. If in the analysis of future data, new clusters are formed, this suggests that new and unknown operating conditions have occurred and this can trigger an alarm for the engineers to investigate further.

It is obvious that in the above process, determining the correct optimum number of clusters becomes important since overestimating the number of clusters will produce a large number of clusters each of which may not necessarily represent truly unique operating conditions, whereas underestimation leads to only small number of clusters each of which may represent a combination of specific events. A proposed method which determines the optimum number of clusters, based on the trend of the exponential of message length difference between two consecutive mixture models, has been formulated in Section 4.2.3 along with explanatory example on random

data. This method is now tested in the next section using data from a simulation of a power system, and later with the actual harmonic data obtained from the Australian harmonic monitoring system used in this thesis.

8.3 Results from the study system harmonic monitoring data

The proposed method to determine the optimum number of clusters is applied to the harmonic data obtained from the Australian harmonic monitoring system. The measured fundamental, 5th and 7th harmonic currents from sites 1, 2, 3 and 4 in Figure 5.1 (taken on 12 -19 January 2002) were used as the input attributes to the MML algorithm (here ACPro). The trend in the exponential of message length difference for consecutive pairs of mixture models is shown in Figure 8.1. Here, the exponential of the message length difference does not remain at 1 after it initially approaches it, but rather oscillates close to 1. This is because the algorithm applies various heuristics in order to avoid any local minima that may prevent it from further improving the message length. Once the algorithm appears to be trapped at the local minima, ACPro tries to split, merge, reclassify and swap the data in the clusters found so far to determine if doing so it may result in a better (lower) message length. This leads to sudden changes to the message length and more often than not, the software can generate large number of clusters which are generally not optimum.

This results in the exponential of message length difference deviating away from 1 to a lower value, after which it gradually returns back to 1. To cater for this, the optimum number of clusters is taken as optimum when the exponential of message length difference first reaches its highest value. Using this method, it can be concluded that the optimum number of cluster is 16, because this is the first time it reaches its highest value close to 1 at 0.9779.

The clusters are subsequently sorted in ascending order based on the mean value of the fundamental current, such that cluster s0 is associated with the off-peak load period and cluster s15 related to the on-peak load period. The statistical parameters

of the 16 clusters (mean (μ), standard deviation (σ) and abundance (π)) are shown in Figure 8.2. However, once the data is segmented into clusters, the operation engineer can quickly look at each cluster and based on his experience interpret the operation conditions associated with the cluster.

With the help of the operation engineers, the sixteen clusters detected by this exponential method were interpreted as given in Table 8.1. It is virtually impossible to obtain these 16 unique events by visual observation of the waveforms from the harmonic monitoring system datashown in Figure 8.3.

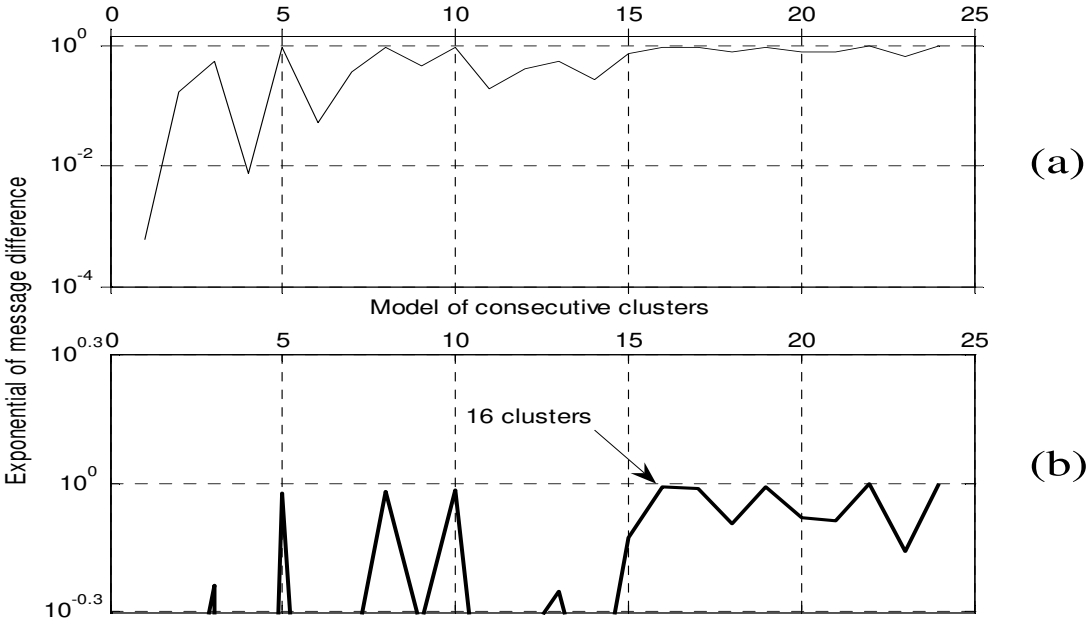

Figure 8.1: (a) Detection of sixteen clusters of harmonic data, (b) Enlargement of (a).

8.4 Using Fitness Function to determine the optimal number of clusters

The fitness function method is then applied to the same data as shown in Figure 8.4 from the harmonic monitoring data. The highest fitness function is found to be 5,

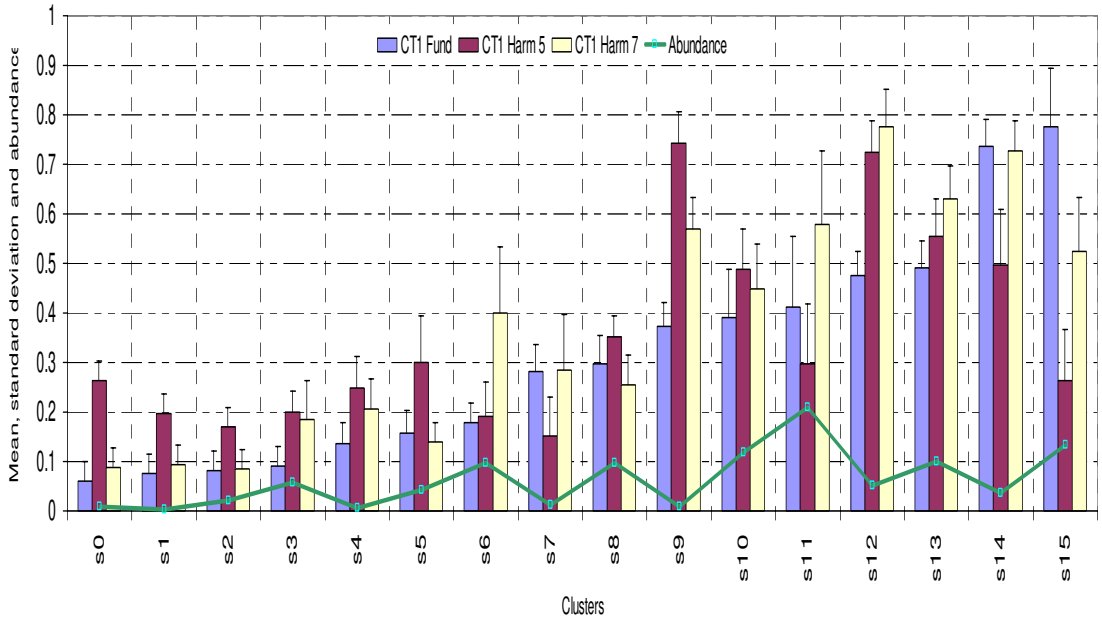

Figure 8.2: The statistical parameters mean(μ), standard deviation (σ) and abundance (π) of the 16 clusters.

which suggests that the optimum number of clusters should be 5. The erroneous results from the fitness function in producing much smaller number of optimum clusters is attributed to the correlation effects between attributes in the measurement data especially between the 5th and 7th harmonic currents. It is not unusual that the fitness function underestimates the number of clusters in correlated data since the theoretical maximum entropy Equation 4.1 in Section 4.2.2 assumes that the attributes are all independent [70].

8.5 Verification of the optimum model using Super-groups

Figure 8.1 shows that when the difference between the message lengths of two consecutive mixture models is close to zero (or its exponential is close to 1) and stays close to zero (or its exponential stays close to 1), then it can be inferred that the two consecutive models are similar. The later model has been formed by splitting one or

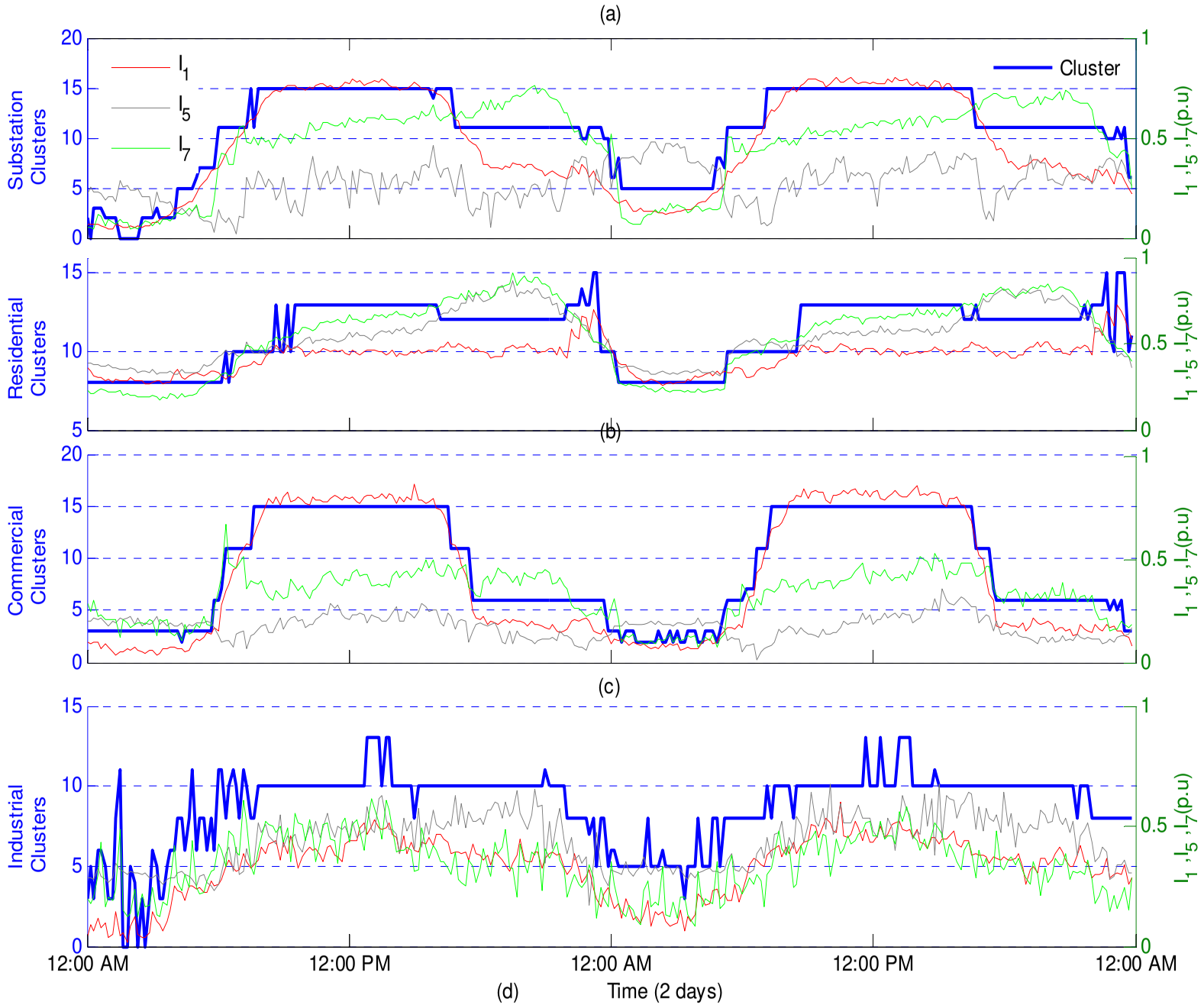

Figure 8.3: Sixteen clusters superimposed on four sites (a) Substation, (b) Residential, (c) Commercial and (d) Industrial.

Table 8.1: The 16 clusters by the method of exponential difference in message length.

Cluster	Event
s0	5th harmonic loads at Substation due to Industrial Site
s1	Off peak load at Substation Site
s2	Off peak load at Commercial Site
s3	Off peak load at Commercial Site due to Industrial load
s4	Off peak load at at Industrial Site
s5	Off peak load at Substation Site
s6 and s7	Switching **on** and **off** of capacitor at Substation Site
s8	Off peak load at Residential
s9	Harmonic load at Industrial Site
s10	Ramping load at Residential Site
s11	Ramping load at Commercial Site
s12	Switching on TV's at Residential Site
s13	Harmonic loads at Industrial and Residential Sites
s14	Ramping load at Substation Site due to Commercial loads
s15	On-peak load at Substation Site due to Commercial loads

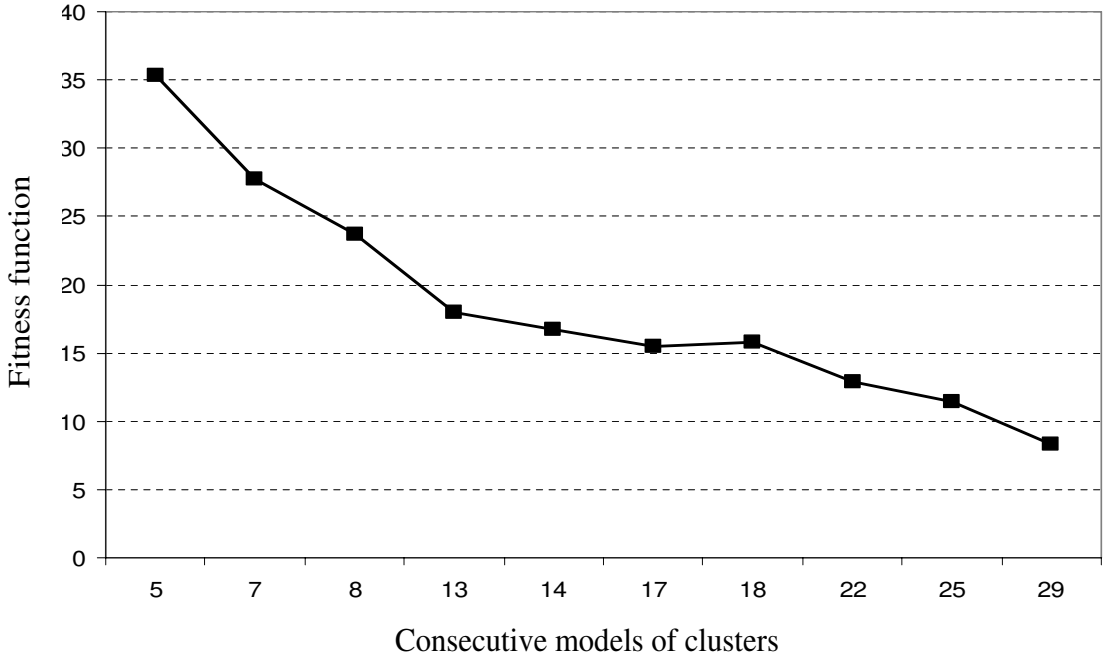

Figure 8.4: Fitness function showing only five clusters as optimum number.

more of the clusters in the previous model into two or more similar clusters. These similar clusters can be re-merged into super-groups (see section 6.4.3) in order to return to the optimum model. This suggests that the super-group techniques can be applied to the MML algorithm to reduce the total number of clusters to the optimum value. This suggests that the super-group technique is a good technique to verify the proposed method of using the trend of the exponential of message length difference to obtain the optimum number of clusters.

To verify this, the same data from section 8.3 was used as an input to ACPro, but now allowing ACPro to produce the maximum number of clusters (30 clusters). The trend in the exponential of message length difference for consecutive pairs of mixture models of the 30 clusters clusters is shown in Figure 8.5. The clusters are subsequently sorted in ascending order based on the mean value of the fundamental current, such that cluster s0 is associated with the off-peak load period and cluster s29 related to the on-peak load period. The profiles of the 30 generated clusters are given in Figure 8.6. The KL distances between the 30 clusters are sorted from the lowest to the highest (the most similar clusters to the most difference ones) as shown in Figure 8.7. A multidimensional scaling algorithm (MDS) [83], which is a dimension reduction technique is then used to form a network from the KL distances shown in Figure 8.7. The MDS Knowledge Network Organising Tools (KNOT)software [80] is used to form the super-group abstractions by removing the links whose distances exceed a dissimilarity threshold (in this case when KL distance is less than 13.5 Bits).

Using this technique and the defined dissimilarity threshold, it was found that sixteen super-groups are obtained as visualised in Figure 8.8. This is the same number of clusters obtained from the proposed method of determining optimum number of clusters usig the trend of the exponential of message length difference. Table 8.2 shows the relationship between the sixteen super-groups obtained using the MDS method and the optimum sixteen clusters obtained from the proposed method based on the trend of message length difference. The profile of the super-groups is shown in Figure 8.9 which are very similar to the profile of the 16 clusters of the optimum model

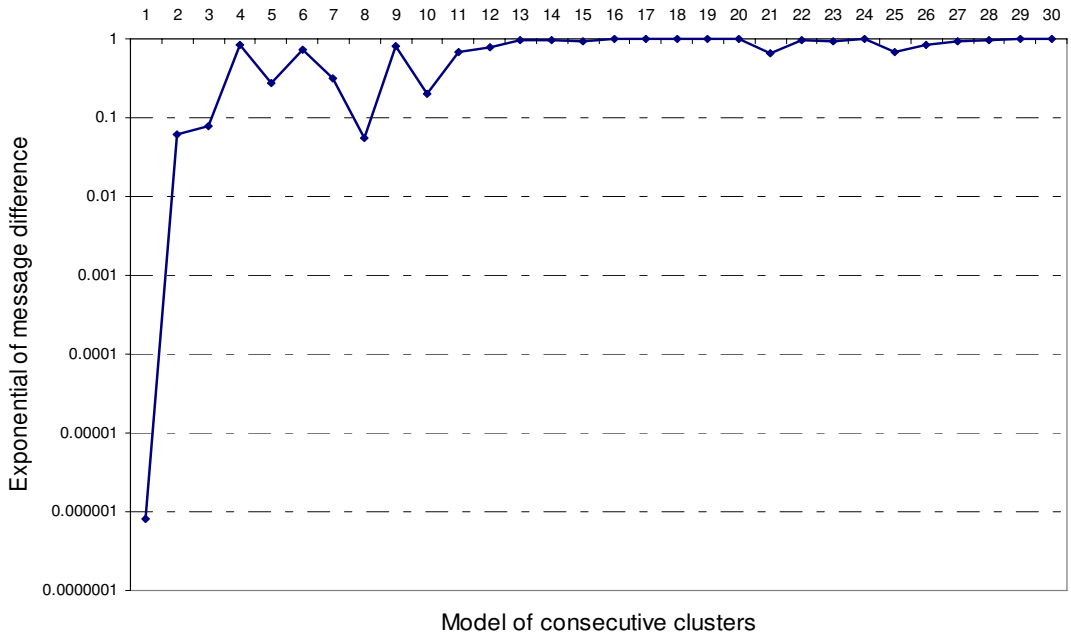

Figure 8.5: Exponential curve for the maximum number of the generated clusters.

shown in Figure 8.2. The KL distances of the clusters that form each super-group is shown in Table 8.3. Figure 8.10 shows the time series data of both the 16 clusters (optimum model) and the 16 super-groups from the 30 clusters. In this way, the MDS method can be used to explain how the clusters obtained from an overestimation of the number of clusters can be merged to form the optimum number of clusters.

This validates that the proposed method based on the trend of the exponential of message length difference can be used to obtain distinct clusters whose KL distances exceed 13.5 Bits as shown in Figure 8.3. It is to be noted that to obtain the sixteen super-groups, the dissimilarity threshold needs to be defined. Increasing or decreasing this threshold will provide smaller or larger number of super-groups and unlike the proposed method based on the trend of message length difference, the MDS method does not in any way imply that the super-groups obtained are optimum. However, the proposed method based on the trend of the exponential message length difference can guarantee that the optimum number of clusters are obtained, since the candidate for the optimal clusters to be found is based on the fact that the difference in message

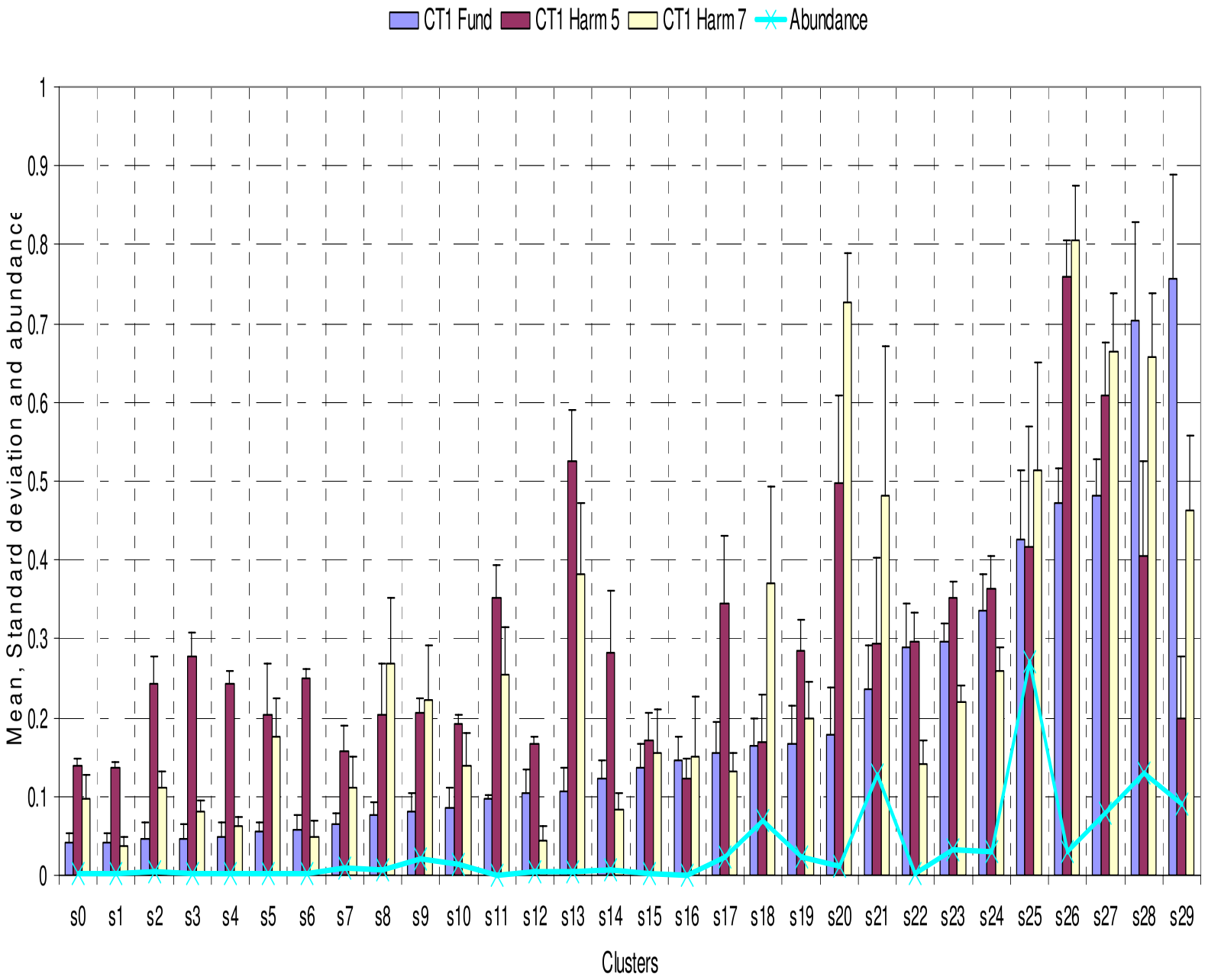

Figure 8.6: The statistical parameters mean(μ), standard deviation (σ) and abundance (π)of large model with 30 clusters.

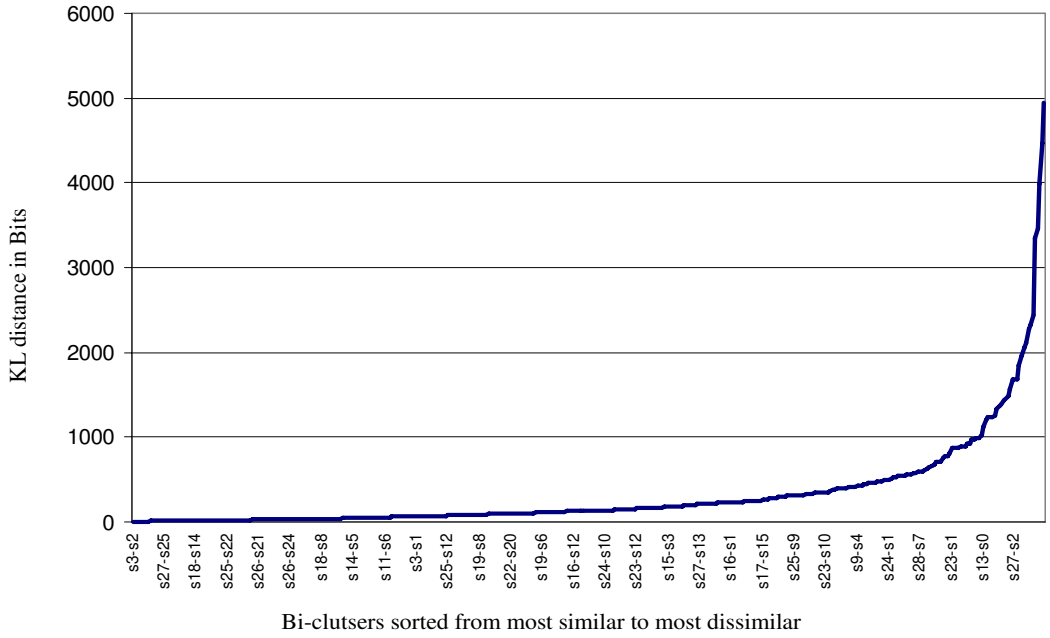

Figure 8.7: The KL distances between the 30 clusters sorted in ascending order.

length between the two models approaches zero or the exponential of message length difference approaches one.

8.6 Interpretation of the Optimal Number of Clusters in Harmonic Data using supervised learning

The C5.0 supervised learning algorithm was applied to the measured harmonic data set but now the cluster names obtained from the optimum cluster method are used as class label to each of the measured harmonic data. As described in Section 6.5 and 6.6, C5.0 algorithm can be used to induce decision trees from labelled data set as well as explanatory rules for each of the cluster or label.

Two main problems may arise when applying the C5.0 algorithm on continuous attributes with discrete symbolic output labels. First, the resulting decision tree may often be very large for humans to easily comprehend as a whole. The solution to this problem is to transform the class attributes, of several possible alternative

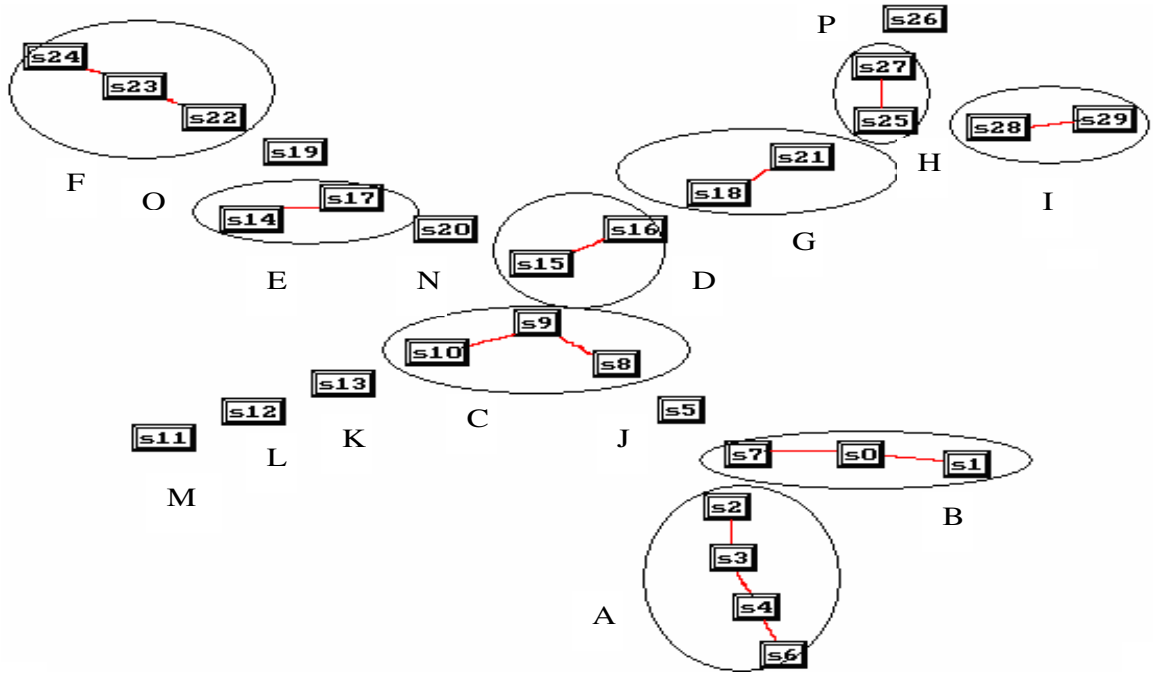

Figure 8.8: Multidimensional scaling: KL-distances are mapped as cumulative link lengths in the graph between any pair of clusters; Super group abstractions are formed through removal of links whose KL-distances exceed a pre-determined dissimilarity threshold.

Table 8.2: Alignment between optimum 16 clusters and super-groups.

Index	optimum 16 clusters	Super Groups
1	s0	A
2	s1	J
3	s2	B
4	s3	C
5	s4	D
6	s5	E
7	s6	G
8	s7	L
9	s8	F
10	s9	K
11	s10	O
12	s11	M
13	s12	P
14	s13	H
15	s14	N
16	s15	I

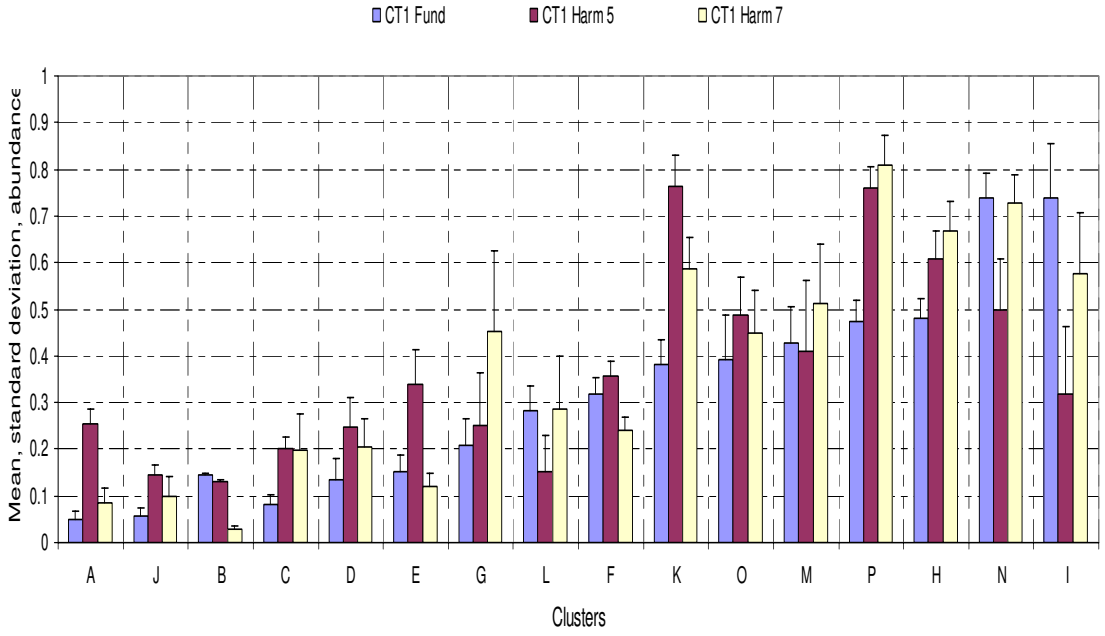

Figure 8.9: The statistical parameters mean(μ), standard deviation (σ) and abundance (π) of the super-Groups (A, B, C, ..., P).

Table 8.3: KL distances (below the threshold value) of the similar clusters.

Cluster sequences	Kl distances
s3-s2	2.202743
s16-s15	3.289948
s10-s9	5.126309
s6-s4	5.126612
s20-s15	5.420664
s9-s8	5.748453
s4-s3	5.994574
s24-s23	7.339619
s17-s14	7.657729
s21-s18	8.123979
s7-s0	10.43736
s29-s28	10.5346
s6-s3	10.60908
s27-s25	11.74551
s6-s2	12.54827
s4-s2	12.70432
s23-s22	12.94932
s1-s0	13.46465

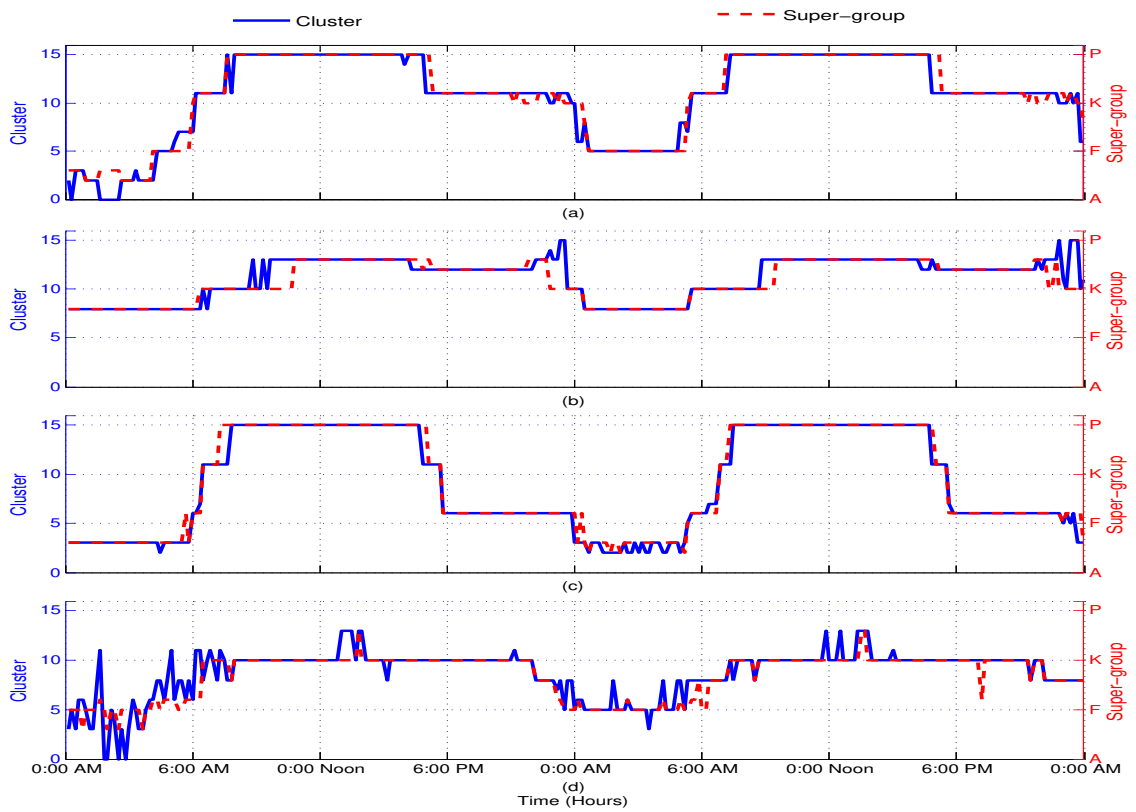

Figure 8.10: The 16 clusters(s0, s1, ..., s16) of the optimum model superimposed on the super-Groups (A, B, C, ..., P) on four sites (a) Substation, (b) Residential, (c) Commercial and (d) Industrial for two days.

values, into a binary set, where the target class is maintained as the first class and all other classes combined as the second class. Second, too many rules might be generated as a result of interpreting each data point in the training data set to belong to the relevant recognized cluster. To overcome this problem, the data is split into ranges instead of considering them as continuous data. These ranges can be built from the average parameters (mean and standard deviation) of each of the mixture of statistical distributions as listed in Table 8.4 and visualised in Figure 8.11 for one of the statistical distribution.

Table 8.4: The continuous data is grouped into five ranges.

Range	Range Name
(0 , μ-2σ)	Very Low (VL)
(μ-2σ , μ-σ)	Low (L)
(μ-σ , μ+σ)	Medium (M)
(μ+σ , μ+2σ)	High (H)
(μ+2σ , 1)	Very High (VH)

8.6.1 Rules discovered from the optimum clusters using decision tree

Using the symbolic values (VL, L, M, H and VH) of input attributes (fundamental, 5th and 7th harmonic current) and the binary sets of classes (s0; as the first class, other; as the second class), (s1; as the first class, other; as the second class), ..., (s15; as the first class, other; as the second class), the C5.0 algorithm is used to uncover and define the minimal expressible and understandable rules behind each of the harmonic-level contexts associated with each of the sixteen cluster described in Section 8.3. Samples of these rules are shown in Table 8.5 for both s12 which has been identified in Table 8.1 as the cluster associated with switching on TV's at the Residential site and s13 which is a cluster encompassing the engagement of other

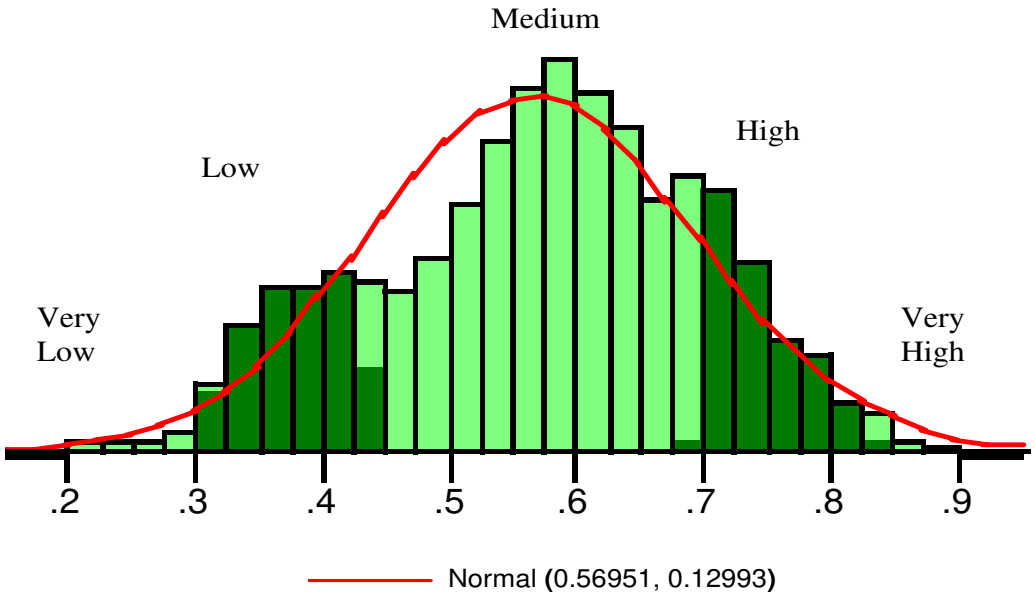

Figure 8.11: The five regions of Gaussian distribution used to convert the numeric values.

harmonic loads at both Industrial and Residential sites. The quality measure of each rule is described by two numbers (n, m) shown in Table 8.5, in brackets, preceding the description of each rules, where:

n: the number of instances assigned to the rule and.

m: the proportion of correctly classified instances.

For this process some 66% of the data has been used as the training set and the rest (33%) are used as the test set, because generally the larger proportion of data used in training the better the outcome will be, however care needs to be exercised to avoid overtraining. The accuracy of the rule using the test data was found to be reasonably close to that of the training data for most of the clusters. The three month data set was also tested and resulted in the same accuracy level as sample data. Table 8.6 shows the accuracy levels for cluster s7, s8, s9 and s10. The utilisation of these rules on new data sets is explained in the next section.

Table 8.5: The generated Rules by C 5.0 for clusters s12 and s13.

Rules for s12 - contains 3 rule(s)		
Rule 1 for s12 (513, 0.891) if FundI = M, and 5thI = VH, then s12	Rule 2 for s12 (523, 0.874) if 5thI = VH, then s12	Rule 3 for s12(10, 0.583) if 5thI = H, and 7thI = VH, then s12
Rules for s13 - contains 1 rule(s)		
Rule 1 for s13 (1,572, 0.622) if FundI = M, and 5thI = H, then s13		

Table 8.6: The accuracy the obtained rules using three months (Jan-Apr 2002) of training and testing data for clusters s7-s10.

Cluster ID	Training (66%)	Testing (33%)	Full data
s7	92.52	91.67	90.91
s8	92.11	91.67	91.46
s9	79.04	80.22	79.50
s10	94.55	95.36	94.04

8.6.2 Rules for prediction of harmonic future data

The generated rules of the C5.0 algorithm used for classifying the optimum clusters have also been used for prediction. Several available harmonic data from different dates were used for this purpose. Data of the same period from another year and data from different time of the year are used to test the applicability of the generated rules. The model accuracy (see Figure 8.12) for data of the same period from another year is considerably high whereas the model accuracy for data from different time of the year is relatively low. This is due to fact that the algorithm performs well when the range of training data and test data are the same, but when these ranges are mismatched then the model will perform poorly and hence the accuracy of the future data (unseen data during training) will be low. Another reason why the accuracy of

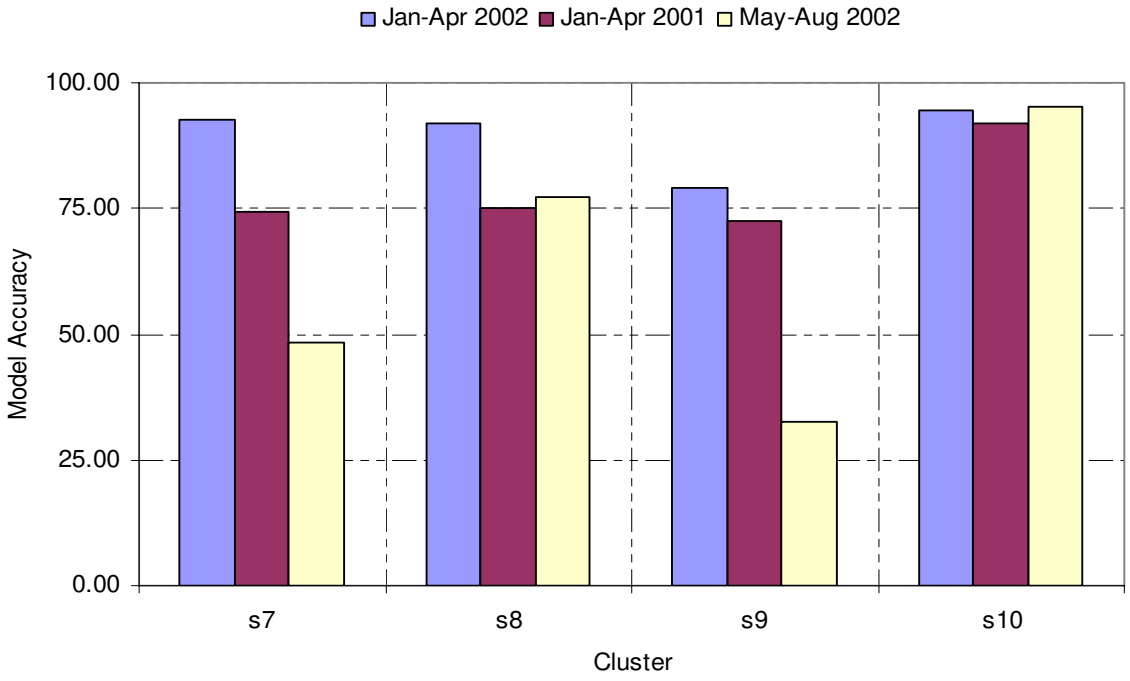

Figure 8.12: Prediction Model accuracy levels for the clusters s7-s10 on training and future data.

the model is low for data set from different period is because the features of the data can change slightly over time and hence the profiles of the generated clusters need to be adjusted accordingly. To explain this idea, the three months harmonic data

used for training has been clustered twice, one when the hot days were included and another when these days were excluded. The detected optimum number of clusters in the first case was larger when applying the exponential of message length difference method as shown in Figure 8.13.

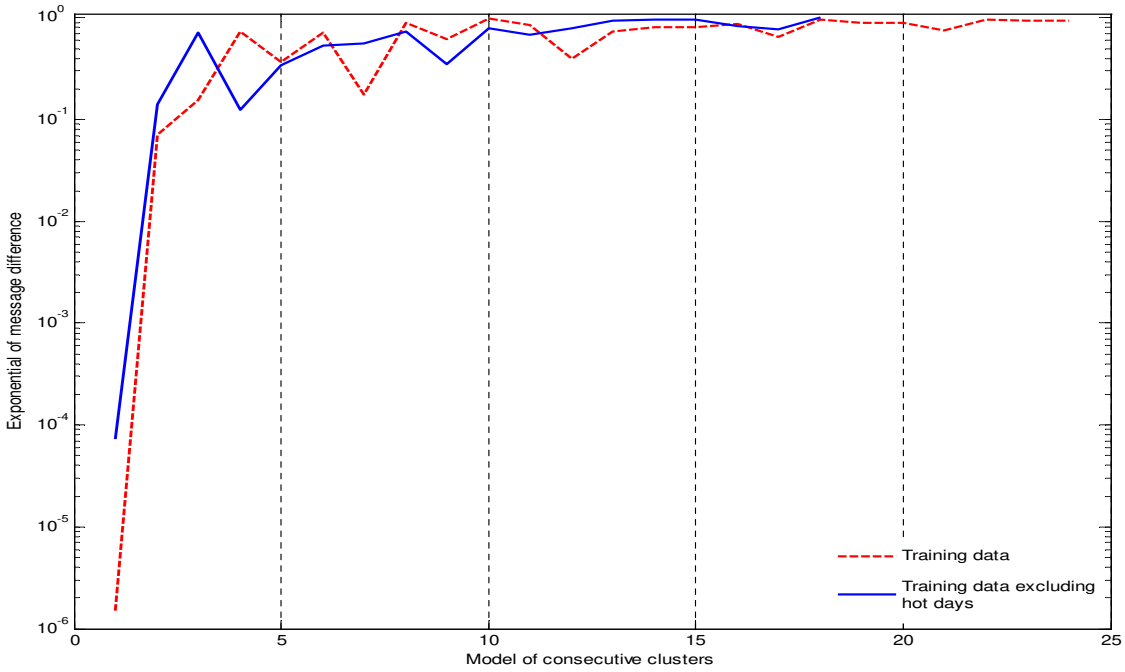

Figure 8.13: Exponential of message length difference for data with and without hot days.

8.7 Summary

The optimal numbers of clusters in two different types of data sets were investigated using the proposed method based on the trend of the exponential difference in message length between two consecutive mixture models. The results of many tests using various two-weekly data sets from the harmonic monitoring data over three year period show that the suggested method is effective in determining the optimum number of clusters in harmonic monitoring data. A commonly used fitness function technique is found to produce underestimation because of the correlated dependent natures of

the attributes presented to the MML program. To validate the optimum number of clusters obtained using the proposed method based on the trend of the exponential difference in message length between two consecutive mixture models, the MDS method was used to form super-groups with a defined dissimilarity threshold. When the threshold is set to 13.5, it was found that similar number of super-groups are obtained as the optimum number of clusters obtained from the proposed method based on the trend of the exponential difference in message length between two consecutive mixture models. The MDS method shows how the clusters obtained from an overestimation of the number of clusters can be merged to form the optimum number of clusters. Correct determination of the number of clusters is important in the diagnosis of power quality disturbances as well for prediction of these events in the future. C5.0 supervised learning tool was applied in the data set from the harmonic monitoring system but now the cluster names obtained from the optimum cluster method are used as class label to each of the measured data. Generated rules from the C5.0 algorithm were used for classification and prediction of future events to determine which cluster any new data should belong to.

Chapter 9

Conclusions

9.1 Conclusions and recommendations

With the increasing amount of data available from the harmonic monitoring system, it is becoming more difficult for power engineers to obtain meaningful information from the harmonic monitoring data. The ability to be able to mine the information from the large amount of measured harmonic data by classifying the data into clusters using data mining technique becomes increasingly important.

This thesis has illustrated that the application of data mining, in particular mixture modelling based on the MML method, to power quality data can identify useful patterns within the harmonic data of a large monitoring program in an Australian medium voltage (MV) distribution system containing residential, commercial and industrial customers.

Mixture modelling based on the Minimum Message Length (MML) algorithm essentially searches for a model which best describes the data using a metric of an encoded message. This method of unsupervised learning, or clustering, has been shown to be able to detect anomalies and identify useful patterns for a given data set. Anomaly detection and pattern recognition in harmonic data can provide engineers with a rapid, visually oriented method for evaluating the underlying operational

information contained within the data set.

The MML method has also identified other clusters within the harmonic data set. Each cluster can represent a specific operating condition such as peak load, off-peak load, capacitor switching operation that can be analysed and confirmed by the operation engineers. By observing how the measured data is classified into various clusters, the power quality event that may have triggered a change from one cluster to another cluster can be more readily deduced. Harmonic producing loads such as TV, air conditioning usage and switching-on of off-peak water heaters have been identified from the generated clusters at residential sites. The associated times of switching, on and off, for other harmonic loads which produced high 5th harmonic levels at commercial and industrial were also identified. The causes of other harmonic events, such as capacitor switching have also been identified using the generated clusters. Other available data (which were not used in the clustering algorithm) such as temperature and reactive power measurements, have been used together with discussions with the system engineers or system operators to confirm these observations.

Once the clusters are generated using the MML method, classification techniques of supervised learning are generally subsequently employed. The C5.0 algorithm of supervised learning techniques has been used to identify the fundamental factors in each of the clusters from which the various clusters can be differentiated from each other. This facilitates the generation of explanatory symbolic rules defining each cluster. The generated rules have been used for interpretation as well as for the provision of a minimal explanatory basis for these clusters. These rules can then be utilised to predict which cluster any new observed data may be best described by.

The main difficulty in applying the MML technique is to determine the optimum number of clusters associated with the global minimum message length, due to its inadequate stopping criterion that may lead to a local minimum within the solution space without finding the optimum number of clusters. Determining the optimum number of clusters becomes important since overestimating the number of clusters will produce a large number of clusters, each of which may not necessarily represent

unique operating conditions, whereas underestimation leads to a small number of clusters, each of which may represent a combination of distinct events.

In this thesis a novel method based on the trend of the exponential of message length difference from the MML algorithm has been developed to determine the optimum number of clusters (or mixture model size) using an actual measured data set from a harmonic monitoring in Australia. Consequently, each of the generated clusters represents a unique operating condition.

The proposed method has been tested by three types of data and proved to be an efficient method. These are data from a known number of clusters with randomly generated data points, with data from a simulation of a power system using the PSCAD®/EMTDC™ and power quality data from an actual harmonic monitoring system using various two-weekly data sets from the harmonic monitoring data in a distribution system in Australia over a three year period.

This approach was also benchmarked against another method based on a fitness function that is also used for the determination of the best (or optimum) number of clusters. However, this later approach failed to find the truly optimum number of cluster in correlated continuous multivariate data such as the harmonic data mentioned in this thesis.

The proposed method has also been validated using a link analysis method to join or merge clusters into super-groups based on a defined dissimilarity threshold. Subsequently it was found that the number of super-groups obtained is similar to the optimum number of clusters obtained from the proposed method. The correct determination of the number of unique operating conditions within the system is important in the diagnosis of power quality disturbances as well in the prediction of these events in the future. If new clusters are found in the analysis of future data, this suggests that new and unknown operating conditions have arisen and this can be used to trigger an alarm for the engineers to investigate further. The C5.0 algorithm has been extensively used in this study to identify the fundamental factors defining and differentiating the various clusters from each other, and also to generate rules used

for both classification and prediction. The cluster names are used as class label to each of the measured harmonic monitoring data. The accuracy level of using the rules obtained from the C5.0 supervised learning algorithm is generally very good both for classifying the training and test data. The accuracy level is also relatively good when applied for prediction of future data as long as the data is from the same period in another year. The accuracy level becomes worse when the test data is from a period coming from a different time of the year. This is due to the altered contextual (or environmental) issues associated with the additional data, such as seasonal change. This data drift, which occurs in many applications, can be dealt with by retraining the test data in order to update the model, after which the prediction stage can be engaged to increase the model accuracy

9.2 Future work

The test system presented in this work was not originally selected for the purpose of a data mining application. Thus there are some important points that need to be taken into consideration when selecting a future study system for the application data mining techniques such as sampling time, network topology and load categorization.

1. Sampling time

   The sampling time used in the harmonic monitoring program was a 10 minute interval and was dictated by memory restrictions of the monitoring equipment. This follows the suggested measurement time interval by the International Electrotechnical Commission (IEC) standard as given in IEC61000-4-30 for measurements of harmonic, inter-harmonic and unbalanced waveforms. The 10 minute interval was however not appropriate in regard to the application of data mining, as this will lead to loss of significant details (such as transient behaviours of some power system component like capacitor switching). The next fastest sampling rate mentioned in the IEC standard is to sample every 3 seconds, which will result in capturing more noise and demand more memory in the

monitoring equipment. An alternative solution to this problem is to use a sampling period of 1 minute, which will produce a trade-off between the need for more information and the need for having very large memory in the monitoring equipment.

2. Network topology

   The distribution system selected for this study is a radial system typical of most distribution systems in Australia. However there are some mesh interconnected distribution systems that are used in the inner city areas to enhance the reliability of the supply to these areas. Further work is required to extend the application of data mining techniques to meshed distribution networks, as the different topology of a distribution system affects the interaction between sites differently and this has the effect of changing the type of information that can be extracted.

3. Load categorization

   In this study, three types of distribution loads - Residential, Commercial and Substation - were deliberately selected to identify the footprints of harmonic emissions from these loads. However, attention should also be given to actual and more diverse loads which are usually represented by a part or full composite of these three load categories.

References

[1] W. W. Dabbs and T.E. Sabin. Probing power quality data. *IEEE Transactions on Computer Applications*, 7, no 2:8–14, 1994.

[2] P. Cheeseman and J. Stutz. Bayesian classsification (autoclass) theory and results. In U. M. Faayad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in knoweldge discovery and data mining*. AAAI Press / The MIT Press, Menlo Park, California, 1996.

[3] C. Wallace and D. Dowe. Intrinsic classification by MML - the snob program. In *proceeding of 7th Australian Joint Conf. on Artificial Intelligence*, Armidale, Australia, 1994. World Scientific Publishing Co., pp. 37–44.

[4] R. Oliver, J.and Baxter and C. Wallace. Unsupervised learning using MML. In *Proceedings of the 13th Int. Conf in Machine Learning:(ICML-96),pp. 364–372.*

[5] D. Robinson. *Harmonic Management in MV Distribution System.* PhD thesis, School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, Wollongong, Australia, 2003.

[6] B.W. Kennedy. *Power Quality Primer.* McGraw-Hill, 2000.

[7] R. Lamedica, G. Esposito, E. Tironi, D. Zaninelli, and A. Prudenzi. A survey on power quality cost in industrial customers. In *IEEE Power Engineering Society Winter Meeting,*, volume 2, no 3, pages 938 – 943, 2001.

[8] R.C. Dugan, S. McGranahan, M.F.and Santoso, and H.W. Beaty. *Electrical Power Systems Quality.* McGraw-Hill, 2002.

[9] D. O. Koval and M. B. Hughes. Canadian national power quality survey: frequency of industrial and commercial voltage sags. *IEEE Transactions on Industry Applications*, 33, no 3:622–627, May/June 1997.

[10] P. P. Barker, T. A. Short, and Burns C. W. Power quality monitoring of distribution system. *IEEE Transactions on Power Delivery*, 19, no 2:1136–1142, April 1994.

[11] V. J. Gosbell, A. Baitch, and M. Bollen. The reporting of distribution power quality surveys. Montreal, Quebec, Oct 2003. CIGRE/IEEE-PES International Symposium "quality of electric power delivery system", pp. 1–6.

[12] M. H. J. Bollen and I.Y.H. Gu. *Signal Processing of Power Quality Disturbances*. John Wiley and Sons, INC, 2006.

[13] X. Mamo and J.L. Javerzac. Power quality indicators. In *IEEE Porto Power Tech, pp. 112–118*, Portugal, 2001.

[14] V. Gosbell, B.S.P. Perera, and H.M.S.C. Herath. Unified power quality index (UPQI)for continuous disturbances. Rio de Janeiro, Brazil. International Conference on Harmonics and Quality of Power ICHQP, 2000, Paper ID: 9.

[15] R. Flores. *Signal processing tools for power quality event classification*. PhD thesis, Chalmers University of Technology, Gotenberg, Sweden, 2003.

[16] C. K. Chui. *Wavelets: A tutorial in theory and applicatios*. Academic, Boston, 1992.

[17] H. Ma and A. A. Girgis. Identification and tracking of harmonic sources in a power system using kalman filter. *IEEE Transaction on Power Delivery*, 11:1659–1665, 1996.

[18] A. Grossmann and J. Morlet. Decomposition of hardy functions into square integrable wavelets of constant shape. *SIMAT*, 15:723–736, 1984.

[19] A.W. Galli, G.T. Heydt, and P.F. Ribeiro. Exploring the power of wavelet analysis. *IEEE Transaction on Computer Application in Power*, 9:37–41, 1996.

[20] M. Wang and Alex. V. Mamishev. Classification of power quality events using optimal time-frequency representations-part 1: theory. *IEEE Transaction on Power Delivery*, 19:1488–1495, 2004.

[21] P.K. Dash, B.K Panigrahi, and G. Panda. Power quality analysis using s-transform. *IEEE Transaction on Power Delivery*, 18:406–411, 2003.

[22] I. H. Witten and E. Frank. *Data mining practical machine learning tools and techniques*. Morgan Kafman, San Francisco, 2005.

[23] H. Mannila. Data mining: machine learning, statistics, and databases. In *8th Inter. Conf. on Scientific and Statistical Database Systems, pp. 2–9*, 1996.

[24] R. Groth. *Data Mining: Building Competitive Advantage*. Prentice Hall, USA, 2000.

[25] C. Westphal and T. Balxton. *Data Mining Solutions: Method and tools for solving real-world problems*. Wiley, USA, 1998.

[26] R. Groth. *Data mining: A hands-on approach for business professional*. Prentice Hall, 1998.

[27] P. Cabena, P. Hadjinian, R. Stadler, J. Verhees, and A. Zanasi. *Discovering data mining from concept to implementation*. Prentice Hall, 1998.

[28] M. Kantardzic. *Data mining concepts, models, methods, and algorithms*. John Wiley and Sons, Inc., 2003.

[29] L. A. Wehenkel. *Automatic learning techniques in power systems*. Kluwer Academic, 1998.

[30] S. Rahman and R. Bhatnagar. An expert system based algorithm for short term load forecast. *IEEE Transaction on Power Systems*, 3:392–398, 1988.

[31] FB. D. Pitt. *Application of data mining techniques to electric load profiling*. PhD thesis, Electrical and electronic Engineering: Manchester Institute of Science and Technology, Manchester, UK, 2000.

[32] R. L. Grossman, C. Kamath, P Kegelmeyer, V. Kumar, and R. Namburu. *Data Mining for Scientific and Engineering Applications*. Kluwer Academic, USA, 2001.

[33] U.M. Fayyad, G. Piatestsky-Shaprio, P. Smyth, and R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining*. MIT Press, 1996.

[34] A.K. Ghosh and D. Lubkeman. The classification of power system disturbance waveforms using a neural network approach. *IEEE Transaction on Power Delivery*, 10:109–115, 1995.

[35] S. Santoso, E.J. Powers, W.M. Grady, and A.C. Parsons. Power quality disturance waveform recognition using wavelet-based neural classifier-part 2. *IEEE Transaction on Power Delivery*, 15:229–235, 2000.

[36] D. Dancey, D. A. McLean, and Z. A. Bandar. Decision tree extraction from trained neural networks. pages isbn: 1–57735–201–7, Florida, United States, 2004. the 17th International FLAIRS Conference, isbn: 1-57735-201-7.

[37] A. Torres, M. T. Rueda, and D. Reyes. Baysian networks for power quality analysis in industrial sector. 9th International Conference on Probabilistic Method Applied to Power Systems, pp. 1–7, June 11–15 2006.

[38] K. Vivek, M. Gopal, and B.K. Panigrahi. Knowledg discovery in power quality data using support vector machine and s-transform. In *IEEE Third International Conference on Information Technology*, pages 32 – 40, 2006.

[39] E. Styvaktakis, M.H.J. Bollen, and I.Y.H. Gu. Expert system for classification and analysis of power system events. *IEEE Transaction on Power Delivery*, 17:423–428, 2002.

[40] T. Pang, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Pearson Education, Boston, 2006.

[41] D. MacKay. *Information theory, inference and learning algorithm*. Cambridge University Press, New York, 2003.

[42] O. Duda, E. Hart, and G. Stork. *Pattern classification*. Wiley Interscience, New York, 2003.

[43] B.S. Everitt, S. Landau, and M. Leese. *Cluster Analysis*. Arnold, New York, 2001.

[44] R. Xu and D. Wunsch. Survey of clustering algorithms. *IEEE Transaction on Neural Networks*, 16, no 3:645–678, May 2005.

[45] M. Negnevitsky. *Artificial intelligence: a guide to intellegent systems*. Addison-Wesley, second edition, 2005.

[46] K. Teknomo. K-means tutorial. www.people.revoledu.com/kardi/tutorial/kMean, Accessed, 02 May 2007.

[47] M.H. Dunham. *Data Mining introductory and advanced topics*. Prentice Hall, USA, 2003.

[48] J.S.R. Jang, C.T. Sun, and E. Mizutani. *Neuro-Fuzzy and soft computing a computational approach to learning and machine intelligence*. Prentice Hall, USA, 2003.

[49] Geofferey McLachlan and Thriyambken Krishnan. *The EM Algorithm and Extensions*. John Wiley & Sons, Inc., 1997.

[50] D. McLachlan, G.J.and Peel, K.E. Basford, and P. Adams. The EMMIX software for the fitting of mixtures of normal and t-components. *Journal of Statistical Software 4*, no. 11:15–21, 1999.

[51] D.M. Boulton and Wallace C. S. A program for numerical classification. *The Computer Journal*, 13, no 1:63–69, 1970.

[52] B. L. Bowerman and O'Connell R. T. *Forecasting and time series: an applied approach*. Duxbury, Belmont, CA, 2006.

[53] Wallace C. S. and D.M. Boulton. An information measure for classification. *The Computer Journal*, 11, no 2:185–194, 1968.

[54] Y. Agusta. *Minimum message length mixture modelling for uncorrelated and correlated continuous data applied to mutual funds classification*. PhD thesis, School of Computer Science and Software Engineering, Monash University, Mellbourne, Australia, 2004.

[55] P. Zulli and D. Stirling. Data mining applied to identifying factors affecting blast furnace stave heat loads. In *5th European Coke and Ironmaking Congress, pp.Tu7:2-1-15.*, 2005.

[56] D. W. Kissane, S. Bloch, P. Onghene, D. P. McKenzie, R. D. Snyder, and D. L. Dowe. The melbourne family grief study, ii: Psychosocial morbidity and grief in bereaved families. *The American Journal of Psychiatry*, 153, no 5:659–666, May 1996.

[57] M. Prior, S. Leekam, B. Ong, R. Eisenmajer, L. Wing, J. Gould, and D. L. DOW. Are there subgroups within the autistic spectrum? a cluster analysis of a group of children with autistic spectrum disorders. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 39, no 6:893–902, 1998.

[58] C. Sun. Human behavioural skills modelling and recognition. Master's thesis, University of Wollongong, Wollongong, Australia, 2007.

[59] M. Field, D. Stirling, F. Naghdy, and Pan Z. Empirical modelling of human gaits for biped robots. Australasian Conference on Robotics and Automation 2007, Paper ID 143, 2007.

[60] T. Edgoose, L. Allison, and D.L. Dowe. An MML classification of protein structure that knows about angles and sequence. Pacific Symposium on Biocomputing (PSB98), pp. 585-96, 1998.

[61] J.J. Oliver and D.J. Hand. Introduction to minimum encoding inference. Technical report, Dept. Statistics. Open University, Walton Hall, Milton Keynes,UK., 1994.

[62] C. M. Stow, A. C. T. Kenington, Milona C., and W. FitzgeraldAsheibi. Experimental issues of functional merging on probability density estimation. Fifth International Conference on Artificial Neural Networks (CP440), Cambridge, UK, July 1997 pp. 123–128.

[63] R. Hanson, J. Stutz, and P. Cheeseman. Baysian classification with correlation and inheritance. San Mateo Clifornia, 1991. The 12^{th} International Joint Conference on Artificial Intellegence, pp. 692–698.

[64] B. Kanefsky, J. Stutz, P. Cheeseman, and W. Taylor. An improved automatic classification of a landsat/tm image from kansas (fife). Technical report, NASA Ames Research Center, Artificial Intelligence Branch, Moffet Field, California, May 1994.

[65] J. Oliver, T. Roush, P. Gazis, W. Buntine, and R. Baxter. Analysis rock samples for the mars lander. *American Association for Artificial Intelligence*, 1998.

[66] P. Cheeseman, J. Stutz, M Self, W. Taylor, J. Goebel, K. Volk, and H. Walker. Automatic classification of spectra from the infrared astronomical satellite (iras). Technical report, NASA Reference no 127 National Technical Information Service, SpringField, Virginia, 1989.

[67] A. Asheibi, D. Stirling, and D. Sutanto. Analyzing harmonic monitoring data using data mining. In *Proceedings of the fifth Australasian Conference on Data Mining and Analystics*, volume 61. pp. 63–68, 2006.

[68] Wallace C. S. Intrinsic classification of spatially correlated data. *The Computer Journal*, 41, no 8:602–611, 1998.

[69] M. A.T. Figueiredo and A. K. Jain. Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Learning*, 24, no 3:381–396, March 2002.

[70] T. M. Cover and Thomas J. A. *Elements of information theory.* Wiley-Interscience, Hoboken, N.J., 2006.

[71] W. Lu and I. Traore. An unsupervised anomaly detection framework for network intrusions. Technical report, Department of Electrical and computer engineering, University of Victoria, SpringField, Virginia, October 2005.

[72] W. Lu and I. Traore. Determining the optimal number of clusters using a new evolutionary algorithm. In *Proceedings of the 17th IEEE International Conference on Tools with Artificial Intelligence (ICTAI05), pp. 712-713.*

[73] A. Asheibi, D. Stirling, and D. Robinson. Identification of load power quality characteristics using data mining. Ottawa, Canada, May 2006. Canadian Conference on Electrical and Computer Engineering (CCECE06), pp. 157–162.

[74] V. Gosbell, D. Mannix, D. Robinson, and Barr. Harmonic survey of an MV distribution system. In *Proceedings of Australasian Universities Power Engineering Conference, pp. 338–342*, Perth, Australia, 23- 26 September 2001.

[75] EDMI. *Users Manual - EDMI 2000-04XX Energy Meter.* Electronic Design and Manufacturing International., 2000.

[76] IEC standard for electromagnetic compatibility (EMC) - part 4-30: Testing and Measurement Techniques - Power Quality Measurement Methods, IEC61000-4-30, 2001.

[77] S. Elphick, V. Gosbell, and S. Perera. The effect of data aggregation interval on voltage resultss. In *Proceeding of Australasian Universities Power Engineering Conference AUPEC07, Paper 15–02*, Perth, Australia, 2007.

[78] I.M. Nejdawi, A.E. Emanuel, D.J. Pileggi, M.J. Corridori, and R.D. Archambeault. Harmonics trend in NE USA: a preliminary survey. *IEEE Transactions on Power Delivery*, 14 no.4:1488–1494, 1999.

[79] Bureau of Meteorology, New South Wales Regional Office Climate Services NSW. www.bom.gov.au/climate/ahead/temps_ahead.shtml, Accessed 24 January 2005.

[80] Knowledge Network Organisation tool KNOT. www.interlinkinc.net/KNOT.html, Accessed, 24 August 2007.

[81] J. R. Quinlan. C4.5: Programs for machine learning. Morgan Kaufmann Publishers, Inc., 1993.

[82] SPSS Inc. *Clementine 8.0 User's Guide*. SPSS Inc., 2003.

[83] J. B. Kruskal and M. Wish. Multidimentional scaling. *Sage University Papers*, 7, no 11, 1978.