2008

Classification and Explanatory Rules of Harmonic Data

Ali Asheibi
University of Wollongong, ali_asheibi@uow.edu.au

David Stirling
University of Wollongong, stirling@uow.edu.au

Danny Soetanto
University of Wollongong, soetanto@uow.edu.au

Publication Details

A. Asheibi, D. A. Stirling & D. Soetanto, "Classification and Explanatory Rules of Harmonic Data," in Australasian Universities Power Engineering Conference, 2008, 2008, pp. 1-5.

Classification and Explanatory Rules of Harmonic Data

Ali Asheibi, David Stirling and Danny Sutanto
Integral Energy Power Quality and Reliability Centre
School of Electrical, Computer and Telecommunications Engineering
University of Wollongong
email: atma64@uow.edu.au

Abstract- Clustering is an important technique in data mining and machine learning in which underlying and meaningful groups of data are discovered. One of the paramount issues in clustering process is to discover the natural groups in the data set. A method based on the Minimum Message Length (MML) has been developed to determine the optimum number of clusters (or mixture model size) in a power quality data set from an actual harmonic monitoring system in a distribution system in Australia. Once the optimum number of clusters is determined, a supervised learning algorithm, C5.0, is used to uncover the fundamental defining factors that differentiate the various clusters from each other. This allows for explanatory rules of each cluster in the harmonic data to be defined. These rules can then be utilised to predict which cluster any new observed data may best described by.

I. INTRODUCTION

Clustering is a process that divides or segments an initial collection of data into a certain number of groups or clusters. Clustering can, in part, be considered as a learning process, and as an analytical method for analysing large volumes of data, by segmenting the large amount of data into clusters and once obtained each cluster can be analysed separately. The premise is that there are several underlying classes that are hidden or embedded within the original data set. The objective of clustering is therefore to identify an optimal model representation of these intrinsic classes, by separating the data into multiple clusters or subgroups.

The Minimum Message Length (MML) technique and mixture modelling was initially developed by Wallace and Boulton in 1968 to classify a large data set into clusters [1]. 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 [2]. Mixture Modelling Methods using MML technique have also been applied to other real world problems such as human behaviour recognition and the diagnosis of complex issues in industrial furnace control [3].

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 truly unique operating conditions, whereas underestimation leads to only small number of clusters each of which may represent a combination of specific events. A novel method which determines the optimum number of clusters, based on the trend of the exponential difference in message length between two consecutive mixture models is proposed in this paper.

In this paper, the proposed technique has been utilised using the MML method to determine the optimum number of clusters (or mixture model size) that can be obtained from a power quality data from an actual harmonic monitoring system in a distribution system in Australia. The clusters obtained are then analysed to understand their relationship to actual operating conditions. A supervised learning algorithm, C5.0, is then employed to identify the essential features of each member cluster and to generate rules for each cluster. These rules can be utilised in predicting which cluster any new observed data may best described by.

II. HARMONIC MONITORING PROGRAM

A harmonic monitoring program was installed in a typical 33/11kV MV zone substation in Australia that supplies ten 11kV radial feeders [4]. The zone substation is supplied at 33kV from the bulk supply point of a transmission network. Figure 1 illustrates the layout of the zone substation and feeder system used in the harmonic monitoring program. The data retrieved from the harmonic monitoring program spans a period from August 1999 to December 2002.

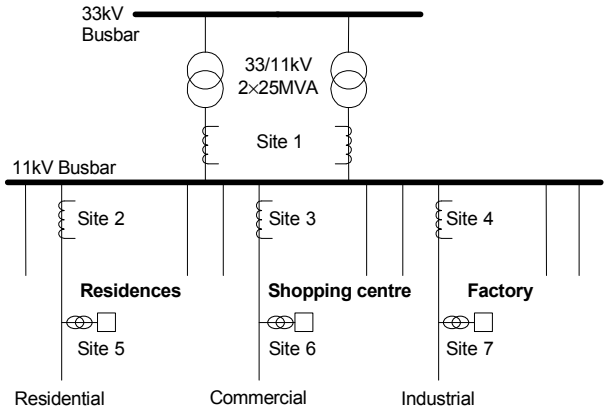

Figure1: Single line diagram illustrating the zone distribution system.

The monitoring equipment used is the EDMI Mk3 Energy Meter from Electronic Design and Manufacturing Pty. Ltd. [5]. Three phase voltages and currents at sites 1-4 were recorded at the 11kV zone substation and at sites 5-7 were recorded at the 430V side of the 11kV/430V distribution transformer, as shown in Fig. 1, The memory capabilities of the above meters at the time of purchase limited recordings to the fundamental current and voltage in each phase, the current and voltage THD in each phase, and three other individual harmonics in each phase. For the harmonic monitoring program, the harmonics chosen to be recorded were the 3^{rd}, 5^{th} and 7^{th} harmonic currents and voltages at each monitoring site, since these are found to be the most significant harmonics [4].

III. Minimum Message Length (MML) Algorithm

A method based on the successful Minimum Message Length (MML) technique has been chosen for clustering the harmonic monitoring data obtained from the harmonic monitoring program. The MML technique has been used extensively in AutoClass [6] and the Snob research programs [7].

The Minimum Message Length (MML) technique is an inductive inference methodology that treats any data set as a hypothetical encoded message. The MML technique then seeks to identify efficient models by evaluating the length of the encoded message that describes each model together with any data which does not fit to the supposed model (exceptions). By evaluating this message length, the algorithm is able to identify, from a sequence of plausible models, those that yield an incrementally improving efficiency, or reducing length. The general concept here is that the most efficient model, describing the data will also be the most compact. Compression methods generally attain high densities by formulating efficient models of the data to be encoded.

The encoded message here 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 for each different model is then calculated and the best model (shortest total message length) is selected. The MML expression is given as [8]:

$$L(D,K) = L(K) + L(D/K) \quad (1)$$

where:
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

An example of how the Mixture Modeling Method using MML technique works, can be illustrated by applying the method to a small data set that contains five distinct distributions of data points (D's) each of which are randomly generated (D1, D2, …, D5), with its own mean and standard deviation. The generated clusters from the model that has the minimum message length correctly identify the five parameters (means and standard deviations) of the five randomly generated distributions using the MML algorithm as shown in Table I. Further the algorithm provides the abundance of each distribution. The abundance value for each cluster represents the proportion of data that is contained in the cluster in relation to the total data set.

Table I. The parameters (π, μ and σ) of the five generated clusters.

Cluster	*Abundance* (π)	*Mean* (μ)	*SD* (σ)
s0	0.198	1.02189	0.27816
s1	0.2	4.00873	0.61683
s2	0.19821	7.91065	0.98041
s3	0.20054	11.8643	1.14631
s4	0.20316	16.0582	1.44659

IV. Proposed Method of Determining Optimal Number of Clusters using MML

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 truly unique operating conditions, whereas underestimation leads to only small number of clusters each of which may represent a combination of specific events. To determine the optimum number of clusters, we propose a method based on the trend of the exponential difference in message length when using the MML algorithm.

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. The total message length in (1) declines as more clusters are generated and hence the difference between the message lengths of two consecutive mixture models is close to zero as it approaches its optimum value and stays close to zero. A series of very small values of the difference of the message length of two consecutive mixture models can then be used as an indicator that an optimum number of clusters has been found. 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

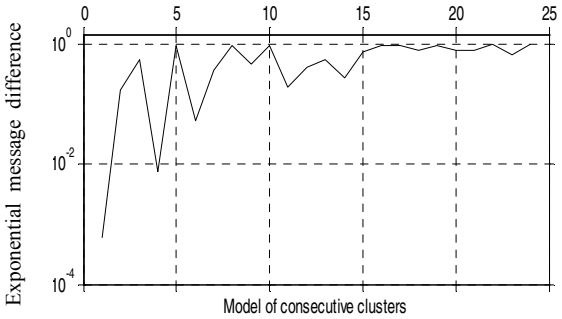

Figure 2: Exponential message length difference

reaches this value can be considered to be the optimum number of clusters.

To illustrate the use of the exponential message length difference curve on determining the optimal number of clusters for the harmonic monitoring system described in Section II, the measured fundamental, 5th and 7th harmonic currents (CT1 Fund, CT1 Harm 5, CT1 Harm 7) from sites 1, 2, 3 and 4 in Fig.1 (taken on 12 -19 January 2002) were used as the input attributes to the MML algorithm. The trend in the exponential message length difference for consecutive pairs of mixture models is shown in Fig. 2.

The optimum number of clusters is taken as when the exponential difference in message length shown in Fig. 2 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 16 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 profiles of the sixteen clusters detected by this exponential method are shown in Fig. 3. With the help of the operation engineers, the sixteen clusters detected by this exponential method were interpreted as given in Table I. It is virtually impossible to obtain these 16 unique events by visual observation of the waveforms shown in Fig. 4.

Table II: the 16 clusters by exponential method..

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 at load Commercial due to Industrial Site
s4	Off peak at Industrial Site
s5	Off peak at Substation Site
s6 & s7	Switching on and off of capacitor at Substation Site
s8	Ramping load at industrial Site
s9	Switch on 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	Switching on harmonic loads at industrial and residential Sites
s14	Ramping load at substation due to commercial Site
s15	On peak load at substation due to commercial Site

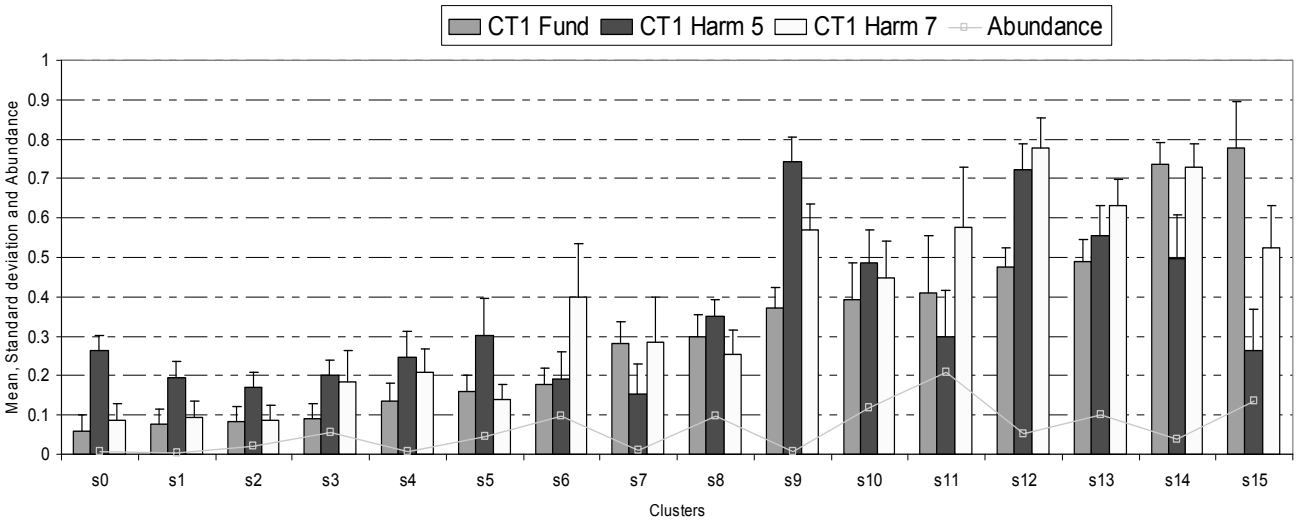

Figure 3: The statistical parameters mean (μ), standard deviation (σ) and abundance (π) of the sixteen generated clusters.

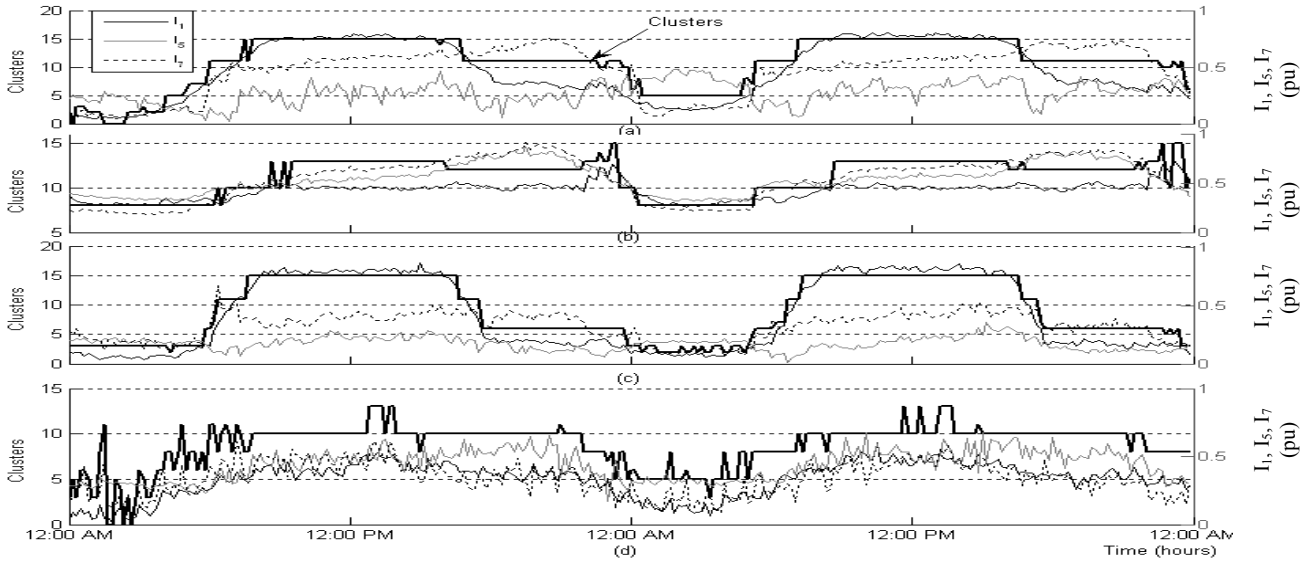

Figure 4: Sixteen clusters superimposed on four sites (a) Substation, (b) Residential, (c) Commercial and (d) Industrial.

V. CLASSIFICATION OF THE OPTIMUM CLUSTERS USING C5.0

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.

A. Categorisation of harmonic monitoring data into ranges

Two main problems may arise when applying the C5.0 algorithm on continuous attributes with discrete symbolic output classes. Firstly, 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 combined as the second class. Secondly, too many rules might be generated as a result of classifying each data point in the training data set to belong to which recognized cluster. To overcome this problem, the data is split into ranges instead of continuous data. These ranges can be built from the average parameters (mean (μ), standard deviation (σ)) of data distributions as listed in Table III and visualised in Fig. 5.

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

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

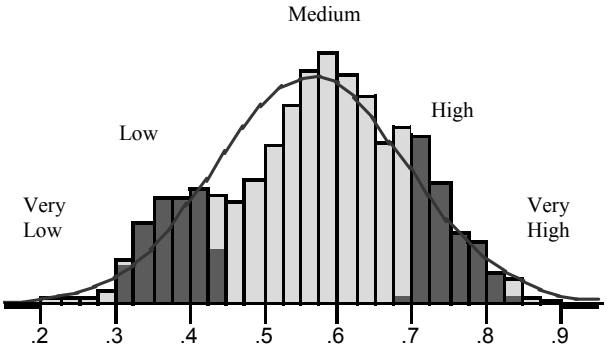

Figure 5: The five regions of Gaussian distribution used to convert the numeric values.

Table IV: The generated rules by C5.0 for clusters s12 and s13.

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

B. Explanatory rules for classifying harmonic monitoring data

The C5.0 algorithm classification tool was applied to the measured data set and the sixteen generated clusters, obtained from the previous section, as class labels. Using the symbolic values (VL, L, M, H and VH) of input attributes (fundamental, 5^{th} and 7^{th} harmonic current) and the binary sets of classes {(s0, other), (s1, other)….(s15, other)} the C5.0 algorithm has been applied as much as the number of clusters (16 times) to uncover and define the minimal expressible and understandable rules behind each of the harmonic-level contexts associated with each of the sixteen cluster listed in Table II. Samples of these rules is shown in Table IV for s12 which has been identified as the cluster associated with switching on TV's at the residential site and s13 which is a cluster encompassing the engagement of other harmonic loads at both Industrial and Residential sites. The quality measure of each rule is described by two numbers (n, m) shown in Table III, 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%) was used as test set, as generally the larger proportion of data used in training the better, however care needs to be exercised to avoid overtraining

C. Rules for predicting harmonic future data

Once generated, the rules from C5.0 can be used for predicting which cluster each future data should belong to. Several available harmonic data from different dates were used for this purpose. Data of the same period from another year (Jan-Apr 2001) and data from different time of the year (May-Aug 2002) were used to test the applicability of the generated rules. The model accuracy for the data from a similar period was considerably higher compared to the accuracy obtained from different period. 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 poor.

CONCLUSION

The paper has used the MML technique to classify a large database of harmonic monitoring data from a distribution system in Australia. A technique is proposed to find the optimum number of clusters when using the MML technique. 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. Correct determination of the number of system unique operating conditions is important in the diagnosis of power quality disturbances as well for prediction of these events in the future. Generated rules of the C5.0 algorithm were used for classification and the provision of a minimal explanatory basis for the optimum clusters.

REFERENCES

[1] C.Wallace, and D.M. Boulton An information measure for classification *The Computer Journal*, Vol 11, No 2, August 1968, pp185-194.

[2] Y. Agusta, Minimum Message Length Mixture Modelling for Uncorrelated and Correlated Continuous Data Applied to Mutual Funds Classification, *PhD Thesis*, Monash University, Clayton, Victoria, Australia, 2004.

[3] P.Zulli and D. Stirling, "Data Mining Applied to Identifying Factors Affecting Blast Furnace Stave Heat Loads," *Proceedings of the 5th European Coke and Ironmaking Congress*, 2005.

[4] D. Robinson, "Harmonic Management in MV Distribution System" *PhD Thesis*, University of Wollongong, 2003.

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

[6] P. Cheeseman, and J. Stutz, Bayesian Classification (AUTOCLASS): Theory and Results, In *Advances in Knowledge Discovery and Data Mining*, Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. ; Uthurusanny, R., eds, pp. 153-180, AAAI press, Menlo Park, California, 1996.

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

[8] J. J. Oliver, and D. J. Hand, (1994) Introduction to Minimum Encoding Inference, [TR 4-94] Dept. Statistics. Open University. Walton Hall, Milton Keynes, UK.