2009

Analyzing harmonic monitoring data using supervised and unsupervised learning

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

Analyzing Harmonic Monitoring Data Using Supervised and Unsupervised Learning

Ali Asheibi, David Stirling, *Member, IEEE*, and Danny Sutanto, *Senior Member, IEEE*

***Abstract*—Harmonic monitoring has become an important tool for harmonic management in distribution system. A comprehensive harmonic monitoring program has been designed and implemented on a typical electrical medium-voltage distribution system in Australia. The monitoring program involved measurements of the three-phase harmonic currents and voltages from the residential, commercial, and industrial load sectors. Data over a three year period have been downloaded and available for analysis. The large amount of acquired data makes it difficult to identify operational events that significantly impact the harmonics generated on the system. More sophisticated analysis methods are required to automatically determine which part of the measurement data are of importance. Based on this information, a closer inspection of smaller data sets can then be carried out to determine the reasons for its detection. In this paper, we classify the measurement data using unsupervised learning based on clustering techniques using the minimum message length technique, which can provide the engineers with a rapid, visually oriented method of evaluating the underlying operational information contained within the clusters. Supervised learning is then used to describe the generated clusters and to predict the occurrences of unusual clusters in future measurement data.**

***Index Terms*—Classification, clustering, data mining, harmonics, monitoring system, power quality (PQ), segmentation.**

I. Introduction

With the increased use of power electronics in residential, commercial, and industrial distribution systems, combined with the proliferations of highly sensitive microprocessor-controlled equipment, more distribution customers are sensitive to excessive harmonics in the supply system [1], some even leading to the failure of equipment. An increasing number of electric distribution network service providers are installing harmonic monitoring equipment to measure the three-phase harmonic voltage and current waveforms in their power system to detect and mitigate the harmonic distortion problems [2]–[7].

Recently, a harmonic monitoring program was designed and implemented in a medium-voltage (MV) distribution system in Australia [8], [9]. The monitoring involved simultaneous measurements of the three-phase harmonic current and voltage from the residential, commercial, and industrial load sectors. The simultaneous measurements of three-phase harmonic currents and voltages from the different load sectors allow for the effect on the net distribution system harmonic voltage and current to be determined. The coordinated approach in obtaining the results has overcome some of the problems with synchronizing and reporting data [10], [11].

An enormous amount of data over a three-year period has been downloaded and available for analysis. However, it is difficult to analyze the data using visual inspection of the acquired voltage and current waveforms. It is also difficult to identify operational issues that generate the harmonics produced at varying operation times. A more sophisticated analysis method is required to automatically segment the data into a manageable data set for analysis to understand the causes and effects of the harmonics obtained and to predict future events.

In this paper, a data-mining tool (ACPRO) is used for the automatic clustering of the harmonic database. Clustering is the discovery of similar groups of multidimensional records in a database. ACPRO is based on the successful AutoClass [12] and Snob programs [13] and uses mixture models [14] to represent clusters. ACPRO allows for the automated selection of the number of clusters and for the calculation of means, variances, and relative abundance of the clusters in the data set.

This paper first describes the design and implementation of the harmonic monitoring program and the data obtained. These data are then clustered using the data-mining tool ACPRO. This paper discusses the significance of the clusters obtained and how the associated operational conditions can be deduced from these clusters. The use of the supervised learning C5.0 algorithm to explain and predict unusual operational conditions is then presented.

II. Harmonic Monitoring Program

A harmonic monitoring program [8], [9] was installed in a typical 33/11-kV MV zone substation in Australia that supplies ten 11-kV radial feeders. The zone substation is supplied at 33 kV from the bulk supply point of a transmission network. Fig. 1 gives the layout of the zone substation and feeder system for the harmonic monitoring program.

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 Fig. 1 are all within the substation at the sending end of the feeders identified as being a predominant load type. Site 5 was along the feeder route approximately 2 km from the zone substation, which feeds a residential area. Site 6 supplies a shopping center with a number of large supermarkets and many small shops. Site 7 supplies a factory manufacturing paper product, such as paper towels, toilet paper, and tissues.

Manuscript received June 06, 2007; revised February 19, 2008. Current version published December 24, 2008. This work was supported by the higher education administration of the Libyan government under Grant HEA 614. Paper no. TPWRD-00394-2007.

The authors are with the School of Electrical Engineering, University of Wollongong, Wollongong 2522, Australia, and also with the Integral Energy Power Quality and Reliability Centre, Wollongong NSW 2522, Australia (e-mail: atma64@uow.edu.au).

Digital Object Identifier 10.1109/TPWRD.2008.2002654

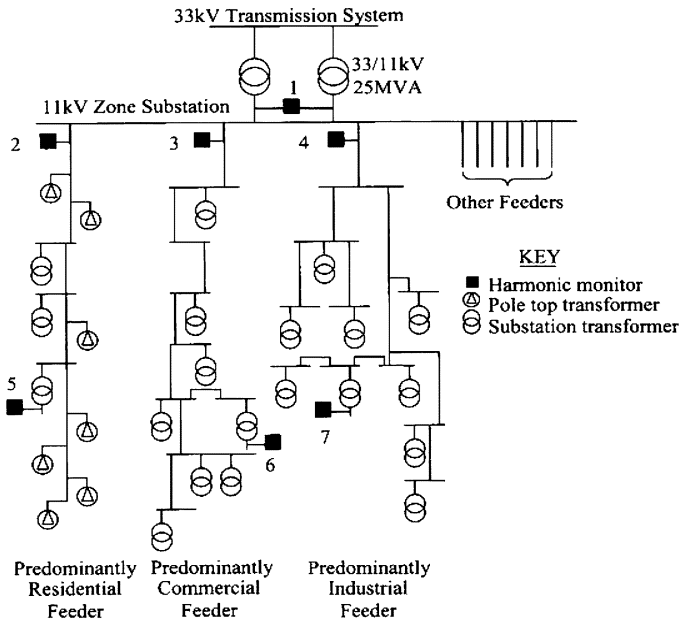


Fig. 1. Single-line diagram illustrating the zone distribution system.

Based on the distribution customer details, it was found that site 2 comprises 85% residential and 15% commercial, site 3 comprises 90% commercial and 10% residential, and site 4 comprises 75% industrial, 20% commercial, and 5% residential.

The monitoring equipment used is the EDMI Mk3 Energy Meter from Electronic Design and Manufacturing Pty. Ltd. [15]. Three-phase voltages and currents at sites 1–4 were recorded at the 11-kV zone substation and sites 5–7 were recorded at the 430-V side of the 11-kV/430-V distribution transformer, as shown in Fig. 1, The memory capabilities of the aforementioned 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 third, fifth, and seventh harmonic currents and voltages at each monitoring site, since these are the most significant harmonics. The memory restrictions of the monitoring equipment dictated that the sampling interval is 10 min. This follows the suggested measurement time interval by the International Electrotechnical Commission (IEC) standard as given in IEC61000-4-30 for measurements of harmonic, interharmonic, and unbalance waveforms. The standard regarded as best practice for power-quality (PQ) measurement recommends 10-min aggregation intervals for routine PQ survey. Each 10-min data represents the aggregate of the ten-cycle root mean square (rms) magnitudes over the 10-min period [16]. Further, a recent study [17] suggested that statistically, sampling at a faster rate will not provide additional significant extra insight.

The data retrieved from the harmonic monitoring program spans from August 1999 to December 2002. Figs. 2 and 3 show a typical output data from the monitoring equipment of the fundamental, third, fifth, and seventh harmonic currents in Phase "a" at sites 1 and 2, taken on January 12–19, 2002, showing a

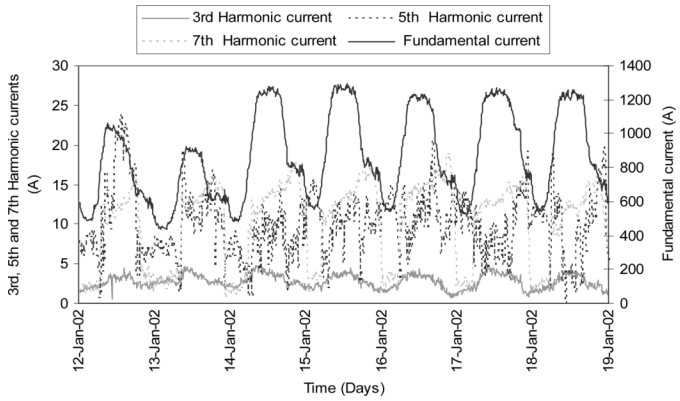


Fig. 2. Zone substation (site 1) weekly harmonic current data from the monitoring equipment.

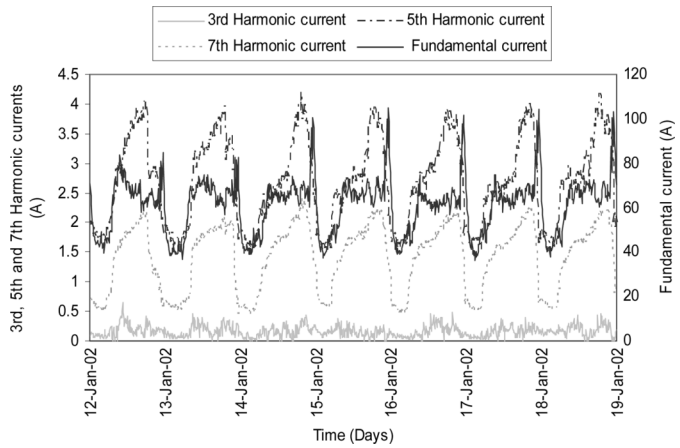


Fig. 3. Residential feeder (site 2) weekly harmonic current data from the monitoring equipment.

10-min maximum fundamental current at 1293 A and minimum fundamental current at 435 A. It is obvious that for the engineers to realistically interpret such large amounts of data, it will be necessary to cluster the data into meaningful segments.

III. Data Mining

There are two important learning strategies in machine learning and data-mining techniques: supervised learning (SL) and unsupervised learning (USL). SL, or data classification, provides a mapping from attributes to specified classes or concept groupings (i.e., classes are identified and prelabelled in the data prior to learning). USL generally amounts to discovering a number of patterns, subsets, or segments (clusters) within the data, without any prior knowledge of the target classes or concepts, that is, learning without any supervision.

In this paper, USL is first used to identify any naturally occurring cluster of a particular set of measured data from the harmonic monitoring system. SL is then used to obtain the relationship between the measured data in the clustering process (training set) and the cluster label. Once trained, the model can then be evaluated on an alternative data (a test set) which contains no prior cluster labels in order to predict which cluster each data point in the test data set should belong to.

A. Unsupervised Clustering Using MML

Unsupervised clustering is based on the premise that there are several underlying classes that are hidden or embedded within a data set which are not known *a priori*. The objective of such processes is to identify an optimal model representation of these intrinsic classes, by partitioning the data into multiple clusters or subgroups.

The partitioning of data into candidate subgroups is usually subject to some objective function such as a probabilistic model distribution (e.g., Gaussian). From any arbitrary set of data, several possible models or segmentations might exist with a plausible range of clusters.

In this paper, a technique based on minimum message length (MML) or minimum description length (MDL) encoding criterion is used to evaluate each successive set of segmentations and monitor their progression toward a globally best model. In this technique, the measured data are considered as an encoded message. The MML inductive inference, as the name implies, is based on evaluating models according to their ability to compress a message containing the 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 data values of the model. The model parameters and the data values are first encoded by using a probability density function (pdf) over the data range and assuming 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

$$\mathrm{L(D,K)} = \mathrm{L(K)} + \mathrm{L(D/K)} \tag{1}$$

where

K	mixture of clusters in the model;
L (K)	message length of model K;
$\mathrm{L(D/K)}$	message length of the data given the model K;
$\mathrm{L(D,K)}$	total message length.

Given a data set D, initially, 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 having Gaussian distributions each with its own mean (μ) and variance (σ) can be calculated from (2) [18]

$$\mathrm{L(K)} = \log_2 \frac{\mathrm{range}_\mu}{\mathrm{AOPV}_\mu} + \log_2 \frac{\mathrm{range}_\sigma}{\mathrm{AOPV}_\sigma} \tag{2}$$

where range_μ is the range of possible μ values; range_σ is the range of possible σ values; and AOPV_μ is the accuracy of the parameter value of μ

$$\mathrm{AOPV}_\mu = \bar{s}\sqrt{\frac{12}{N}}. \tag{3}$$

where $\bar{s}$ is the unbiased sample standard deviation

$$\bar{s} = \sqrt{\frac{1}{(N-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{4}$$

N	number of data samples;
$\bar{x}$	sample mean;
$\mathrm{x_i}$	data points;
$\mathrm{AOPV}\sigma$	accuracy of the parameter value of σ.

$$\mathrm{AOPV}_\sigma = \bar{s}\sqrt{\frac{6}{N-1}}. \tag{5}$$

The message length of the data using Gaussian distribution model can be calculated from the following equation [18]:

$$\mathrm{L(D/K)} = N\log_2 \frac{\bar{s}\sqrt{2\pi}}{\mathrm{Aom}} + N\frac{s^2 + \frac{\bar{s}^2}{N}}{2\bar{s}^2}\log_2(e) \tag{6}$$

where Aom is the accuracy of measurement and s is the sample standard deviation

$$s = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{7}$$

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 of the data in one cluster $(\mathrm{K} = 1)$ with a sample mean $\bar{x}$ and standard deviation s, the parameters μ, σ and π (mean, variance, and abundance) of this model can be estimated using the expectation maximization algorithm (EM) to fit the Gaussian distribution model [19]. The abundance valueπ for each cluster represents the proportion of data that are 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 importance of each cluster. A small abundance may mean the cluster represents a rare occurrence and this may point out instances when the system needs to be observed more carefully.

Once μ and σ are obtained, range_μ and range_σ can be estimated, and AOPV_μ and AOPV_σ can be calculated from (3) and (5). The total message length L (D, K) can then be calculated using (1), (2), and (6). The single cluster may be subsequently divided into a mixture of two clusters having the chosen distribution $(\mathrm{K} = 2)$ each with its own sample mean $\bar{x}$ and standard deviation s. EM is then used to optimize the parameters μ, σ and π (mean, variance, and abundance) of each new cluster. 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 optimization 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 $(\mathrm{K} = 3)$, and the aforementioned process is then repeated.

By itself, the splitting method is deficient to find the minimum message length in that the MML is often not found. To overcome this problem, other tactics are used in our program,

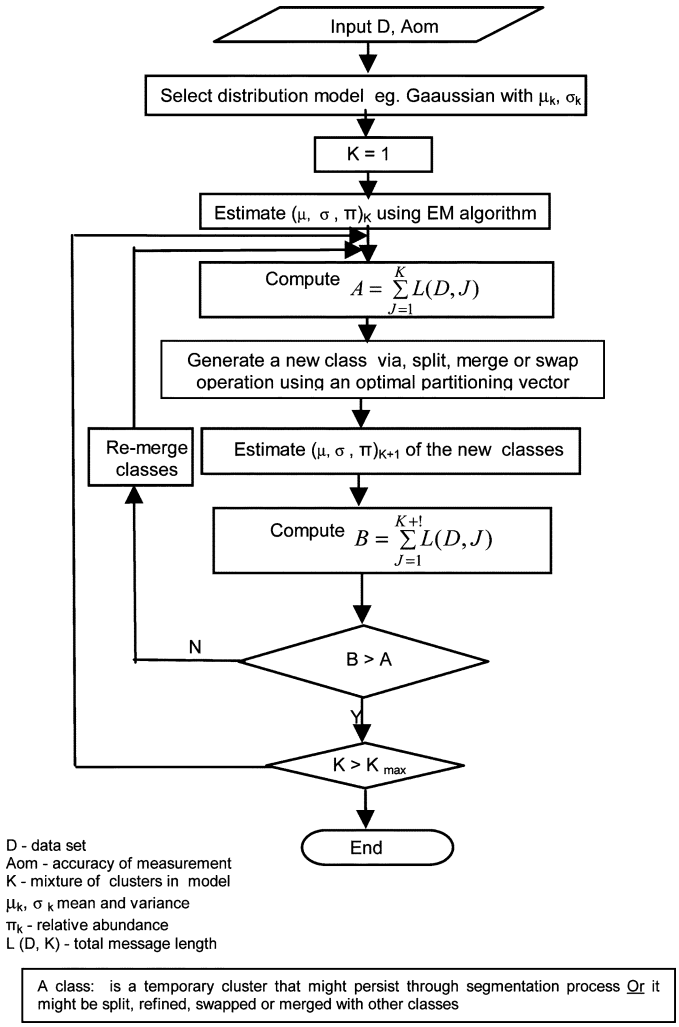


Fig. 4. Conceptual flowchart of clustering algorithm MML.

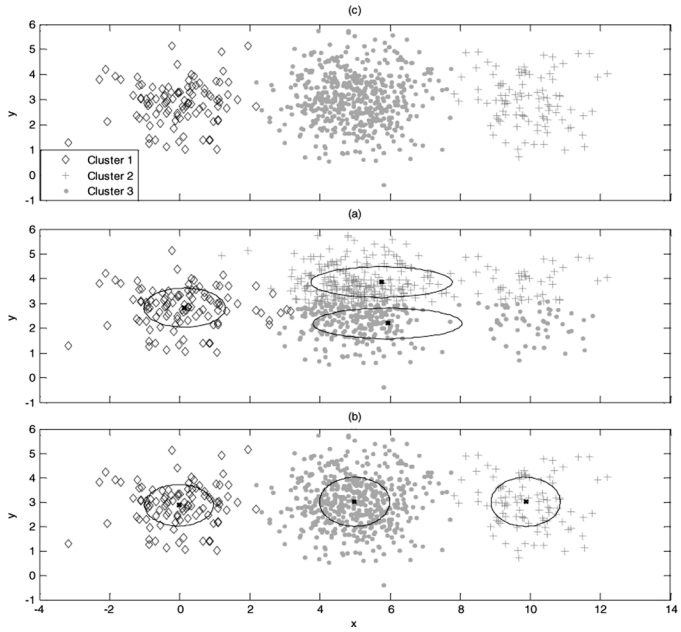


Fig. 5. (a) Three randomly generated clusters. (b) Clustering using K-means. (c) Clustering using mixture models.

such as merging, reclassifying, and swapping [20]. A conceptual flowchart of the MML clustering algorithm is given in Fig. 4. An illustrative example of how the MML algorithm can be applied to a small data set is given in the Appendix.

The use of the MML clustering algorithm for PQ classification has several advantages over traditional methods. One advantage of applying the MML technique in power-quality (PQ) monitoring data is that it does not require the full harmonic waveforms to do the classification, unlike the signal-processing techniques, such as Fourier transform (FT) or wavelet transform (WT), which first requires the waveform to perform the transformation to the relevant domain, and only then can the classification process be initiated.

The MML method used here is often also known as mixture modelling or intrinsic classification [21], [22]. Mixture models typically perform better than those based on *a priori* distance measures, such as a nearest neighbor algorithm, for example, K-means [23]. The mixture model clustering is a general form of K-means or fuzzy C-means because it can use other types of distributions beside Gaussian distributions, with various shapes of clusters [19]. K-means and fuzzy C-means are known to fail to obtain acceptable clusters when the clusters have different sizes, shapes, or covariances. For benchmarking purposes, three randomly generated clusters are shown in Fig. 5. Mixture models based on MML can classify these three natural clusters correctly as shown in Fig. 5(c), whereas K-means fails to detect the right cluster as shown in Fig. 5(b), because the center cluster is larger and denser than the other two.

For this reason, we have chosen the mixture modelling program based on MML for automatic clustering of the harmonic database [12]. The software allows the selection of the number of clusters with given data precision, and produces models structured as a collection of the means, variances, and relative abundance of each constituent cluster.

B. Supervised Learning Using the C5.0 Algorithm

In the last section, unsupervised clustering using MML was suggested to identify the natural classes from the measured data. Once classified using MML, these clusters can then be described and predicted from the measured data by using SL, providing a map from attributes to specified classes or concept groupings.

Decision trees are one example of these classification techniques, such as neural network or Bayes classifiers. In decision trees, a model is built proposing plausible relationships between the input data (training set) and the class, or here the cluster label obtained from MML. Once the model is trained with sufficiently good accuracy, it can then be applied to another data set (test data) having unknown classes in order to predict which data point in the test data set belongs to which recognized cluster. The optimum model is the one that has low errors in each of the aforementioned two steps.

In order to obtain high accuracy in the first step (training step), a large 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.

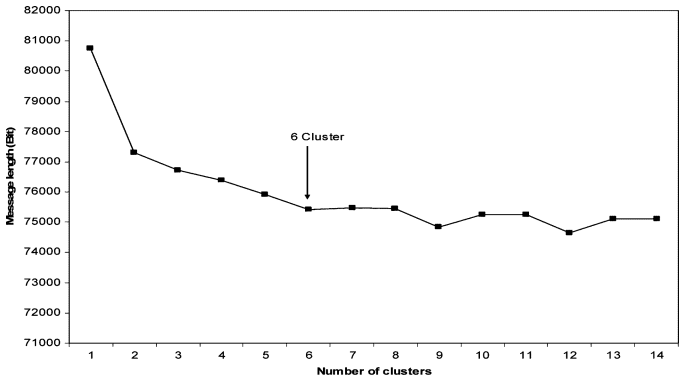


Fig. 6. Message length versus the number of generated clusters.

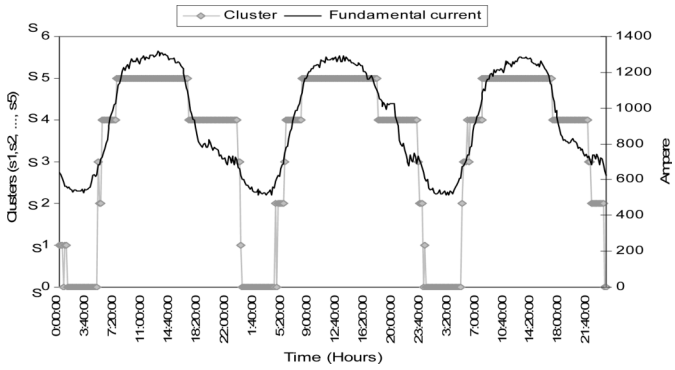


Fig. 7. Clusters obtained superimposed on the phase "a" fundamental waveform at substation site (site 1).

In this project, the C5.0 algorithm is used to carry out the supervised learning process, which can represent the results either as a decision tree or as a rule set structure, both of which are symbolic and can be easily interpreted. The C5.0 algorithm is an advanced SL tool with many features that can efficiently build the decision tree and also facilitate the pruning process [19]. Once trained, the decision tree or rule set obtained can then be used to subsequently infer or classify which class cluster any new data belong to. In this work, various data-processing and management tasks, including the supervised learning with the C5.0 algorithm, are supported within Clementine [24], an integrated data-mining work bench.

IV. Results and Outcomes

ACPRO was applied to the measured harmonic data from the monitoring program for the test system in Fig. 1. Three attributes (fundamental, fifth, and seventh harmonic currents) were selected from different sites (sites 1, 2, 3, and 4). The third harmonic current was excluded as its level was low due to the presence of Δ/Y transformers downstream, which block most of the third harmonic current from flowing up as shown in Figs. 2 and 3. The data were normalised to the range (0–1) and then used as input to the software with a given accuracy of measurement (Aom). Six different clusters, each with specific abundance, mean, and standard deviation were obtained. The reason behind selecting this number of clusters is that the decline in the message length significantly decreases at cluster 6, and the message length is fairly constant afterward as shown in Fig. 6.

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 offpeak load period and cluster s5 is related to the onpeak load period as shown in Fig. 7.

Each generated cluster can therefore be considered as a profile of the three variables (fundamental, fifth, and seventh harmonic currents) within an acceptable variance. If new data lie beyond the variance, another cluster is created (see Fig. 8).

From this sorting process, one can see that cluster s5 not only has the highest fundamental current, but also the highest fifth harmonic current. This infers that the high fifth harmonic currents are due to an overloading condition. Fig. 8 also shows that

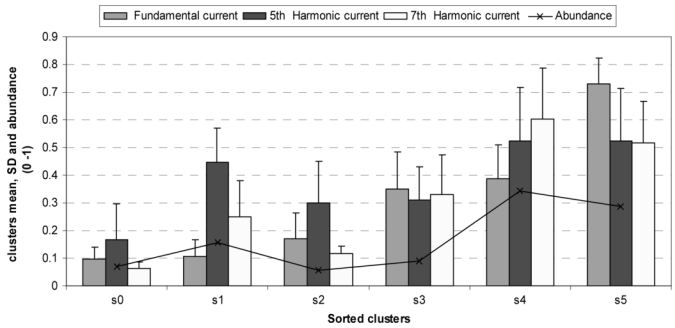


Fig. 8. Clusters' statistical parameters mean (μ), standard deviation (σ), and abundance ($\prod$).

TABLE I
Abundance Values for Each Generated Cluster

Cluster	S0	S1	S2	S3	S4	S5
Abundance (%)	9	16	5	8	33	29

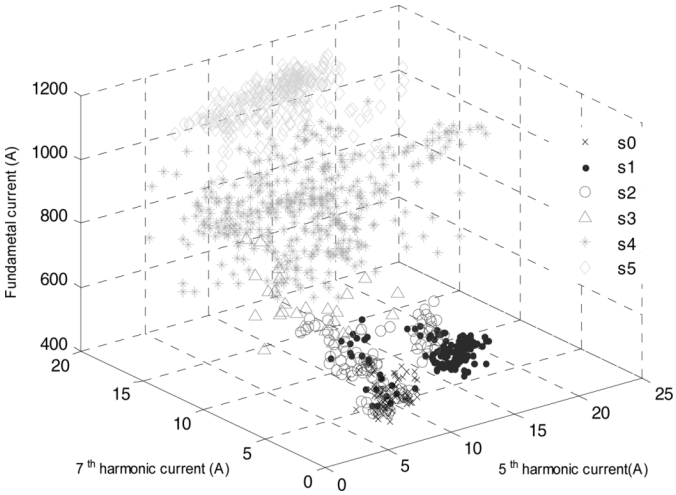


Fig. 9. Six clusters obtained at the substation site (site 1).

cluster s2 has very low abundance. This may be viewed as an anomalous, and potentially a problematic cluster as described later. Table I shows the abundance value of each cluster.

The visualization of the six clusters at site 1 is shown in Fig. 9, showing the relationship among fundamental, fifth, and seventh harmonic currents and the obtained clusters.

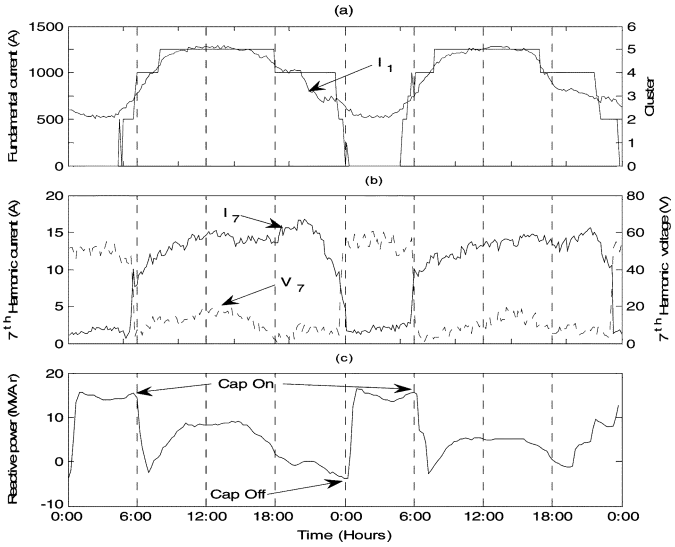


Fig. 10. Clusters at the substation site in two working days. (a) Clusters superimposed on the fundamental current waveform. (b) Seventh harmonic current and voltage data. (c) MVAr load at the 33-kV side.

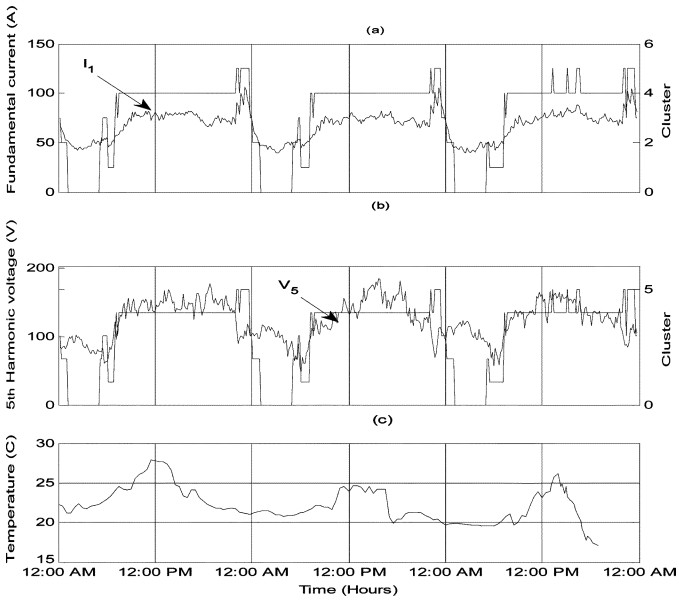


Fig. 11. Three normal temperature days at the residential site (site 2), fundamental current and generated clusters. (b) Fifth harmonic voltage and generated clusters. (c) The temperature near site 2.

A. Interpretations of the Results From Unsupervised Learning Using MML

By observing how the measured data are classified into various clusters, the power utility engineer can more readily deduce the PQ event that may have triggered a change from one cluster to another cluster. To confirm the observation, other available data can be used, such as temperature and reactive power measurements or by discussion with the system engineers or system operators.

For example, the MML clustering algorithm has identified sudden changes to cluster s2 at particular time instances during the day. Fig. 10(a) shows the clusters obtained from substation site (site 1) superimposed on the fundamental current measurement data for two days. Fig. 10(b) shows the seventh harmonic current and seventh harmonic voltage at the substation. By observation, it appears that this is due to sudden changes in the seventh harmonic current. After further investigation of the MVAr measurement at the 33-kV side of the power system shown in Fig. 10(c), it can be deduced that the second cluster (s2) is related to the capacitor switching event. Early in the morning, when the system MVAr demand is high as shown in Fig. 10(c), the capacitor is switched on in the 33-kV 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 seventh harmonic current and voltage plots as shown in Fig. 10(b) provide a clue that something is happening during cluster s2, in that the seventh harmonic current increases rapidly and the seventh harmonic voltage decreases, although the reason is still unknown. In this case, the clustering process 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 33-kV side can also be detected at the other sites (sites 2, 3, and 4) at the 11-kV side.

Although in this case the cause can be easily uncovered, there may be other cases where the clustering process can identify a cluster which can produce detrimental effects to the power system, which can 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 identify different operating conditions based on the different data attributes that are provided to the program (fundamental, fifth, and seventh harmonic currents). Once identified, more information can be gathered to deduce the reasoning why the cluster is generated. The deduction can then be confirmed by discussion with system engineers or system operators. In this way, anomalous cases can be quickly identified and analyzed.

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 fifth harmonic voltage are the characteristics of this cluster. This is shown in Fig. 11 for a period of three days when the temperature is normal for the time of the year. Fig. 12 shows the results for a period of three days when the weather is very hot, resulting in a significant use of air conditioners. There is usually a lag (human response) between the peak temperature and the onset of the peak use of an air conditioner, thereby causing a noticeable lag between the peak temperature and the sudden increase in the fifth harmonic as shown in Fig. 12(b) and (c).

Fig. 13 shows the difference in harmonic clusters at the residential site between the normal weather days and the hot days. It is evident that the MML has identified s5 cluster occurring

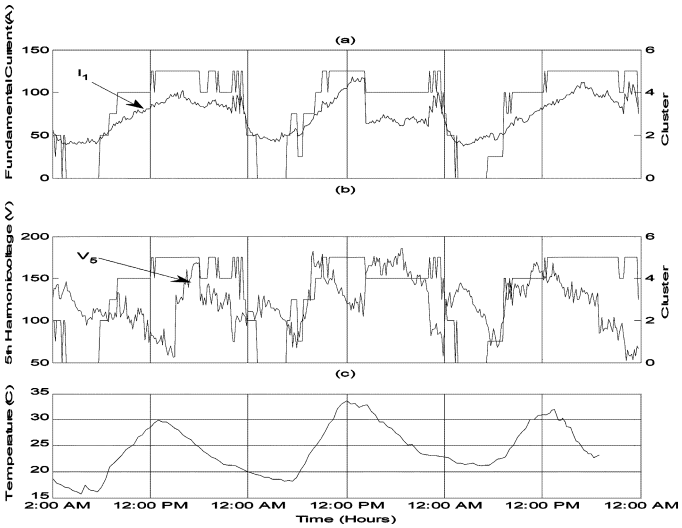


Fig. 12. Three hot days at the residential site (site 2). (a) Fundamental current and generated clusters. (b) Fifth harmonic voltage and generated clusters. (c) The temperature near site 2.

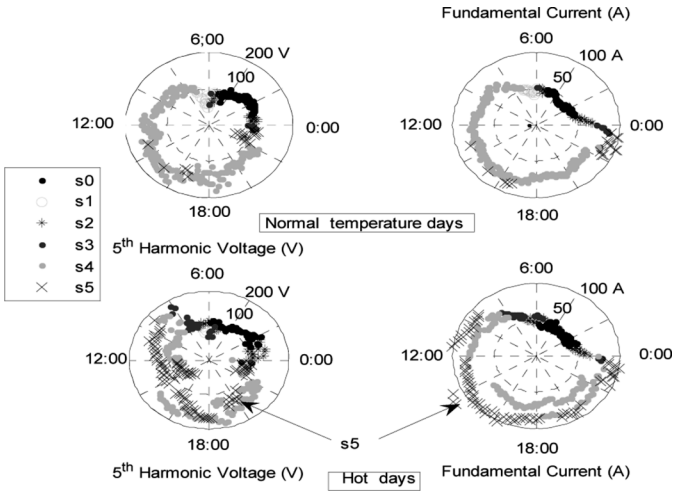


Fig. 13. Normal and hot days at residential site (site 2).

more often at daytime during the hot period compared to the days when the temperature is relatively mild.

From Figs. 11 and 12, it can also be observed that there is a period of peak load (cluster s5) around midnight, and following a discussion with the utility engineer, we were told that this is related to the turning-on of the offpeak water heaters.

B. Results From SL Using C5.0

To gain close insight into the obtained clusters as to what makes specific clusters differ from each other, the C5.0 algorithm classification tool was applied to the measured data set and the generated clusters from MML. The usefulness of this algorithm is that it can be used to describe and predict generated clusters with the results represented either as a decision tree or sets of "if ...then" rules without requiring much computation. A lagging time window of different ranges of time (30, 60, 90, and 120 min) is used in order to predict the occurrence of the clusters. This results in rules describing each cluster in terms of the values of the input attributes (fundamental, fifth, and seventh harmonic currents), time, and site locations.

TABLE II
RULES DESCRIBING S2 CLUSTERS GENERATED BY C5.0

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

Data-mining software, Clementine, was used in this section to produce the rule set related to each cluster. The most important rules are the ones associated with the least abundant clusters, as these clusters are considered to be anomalies among other clusters. Clusters s2 and s3, with proportions of 5% and 8%, respectively, are the least abundant (see Table I).

The discovered rules for cluster s2 with a window size of 60 min are shown in Table II. The range of attributes values is (0–1), as explained in the previous section. The accuracy of the model used to generate these rules was high at 98.8%. It should be realized that the C5.0 algorithm is used to generate rules explaining what the combining influences behind each cluster are. If different clusters are obtained from other sites, then new rules would obviously be required to formalize the different contexts associated with each new cluster.

The quality measure of each rule is described by two numbers (n, m) shown in Table II, 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.

The number of instances, from trained data, of Rule 1 is 286 with 267(286* 0.934) being correctly classified. Rule 1 means that if the fundamental current in phase a (C1a) was in the range (13.8%–.37.4%) 50 min ago [see Fig. 14(a)], and if the fifth harmonic current in the same phase (C5a) was in the range (9.5%–47.4%) 50 min before [Fig. 14(b)], and if the seventh harmonic current (C7a) was between (6.9%–15.3%) 50 min ago [Fig. 14(c)], then s2 will occur.

For example, if we consider a 3-h period between 3 A.M. and 6 A.M. on January 14, 2002 at the substation site, it can be observed that s2 occurs between 4:10 A.M. and 5:40 A.M. (Fig. 14). At 4:50 A.M., we can see that the value of I_1 50 min ago is between 0.138 and 0.374, and I_5 is between 0.095 and 0.474 and I_7 is between 0.069 and 0.153 and, hence, we can observe that at 4:50 A.M., C5.0 will predict that S2 will be generated as can be observed in Fig. 14. On the other hand, at 5:50 A.M., although I_1 and I_5 meet the rules associated with these two currents, but I_7does not meet the rule because it is not between 0.069 and 0.153 and, therefore, at time 5:50 A.M., S2 is not predicted by C5.0, instead s3 is predicted.

There are some instances where more than one rule is applied at the same time. The C 5.0 algorithm then applies the rules in

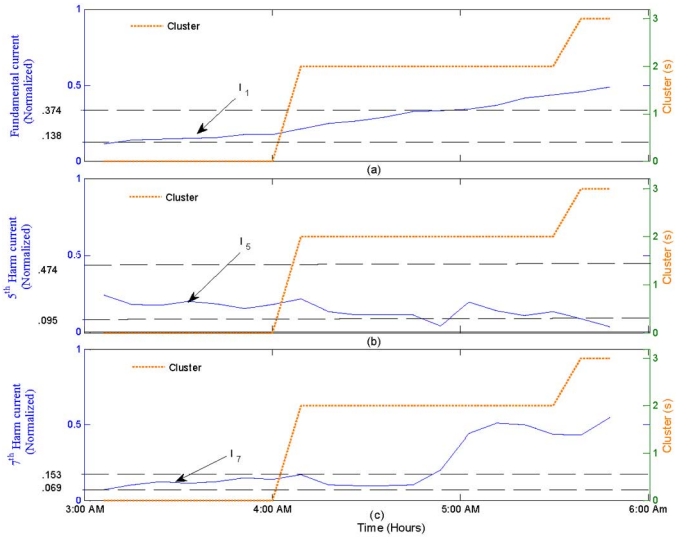


Fig. 14. Predicting cluster (s2) of capacitor switching.

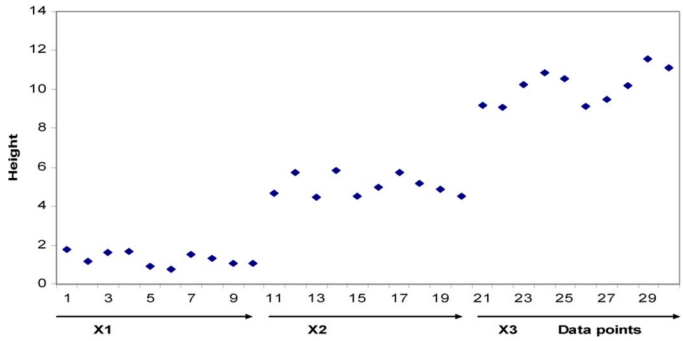


Fig. 15. Three cluster (30 data points) generated randomly from X1, X2, and X3.

TABLE III
DATA POINTS SHOWN IN FIG. 15

N	x1	x2	x3
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

the rule set, and makes a majority decision based on whether the cluster is generated. This means that one rule is not enough to predict the cluster (class) and so all of these rules should be considered.

V. CONCLUSION

PQ data from a harmonic monitoring program in an Australian MV distribution system containing residential, commercial, and industrial customers has been analyzed using data-mining techniques. The technique presented in this paper allows utility engineers to detect unusual PQ events from monitored sites, using clustering, then characterizing the obtained clusters using the classification techniques to infer information about future PQ performance at the monitored sites. Unsupervised learning and, in particular, cluster analysis using MML that selects the best model describing the data using a metric of an encoded message, has been shown to be able to identify useful patterns within the PQ monitored data set. The advantage of mixture models, in general, is that they can cope with different types of distributions. The main difficulty with the MML technique is to determine the optimum number of clusters associated with the global MML. Further work is currently being carried out to achieve this. The usefulness of the decision tree, unlike neural networks, is that it performs classification without requiring significant training and its 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. How to decide the most suitable rules is still an issue with this type of technique.

APPENDIX

This is an illustrative example to explain how MML is used to encode the data. Fig. 15 shows a set of 30 points generated randomly from three normal distributions of ten points each. The data for the 30 points are given in Table III. Assuming the Gaussian distribution model and the Aom is 0.1, and from the data values, the range of μ is chosen as 0–25 and for σ, the range is 0–5, the segmentation process can be described in the following steps:

A. Step 1

Considering the whole data set as one cluster and calculating the message length of this cluster from (1–7) yields 123.6 b.

B. Step 2

By splitting the data into two clusters $(\mathrm{N}_1 = 16, \mathrm{N}_2 = 14)$ by using optimal partition [25], the total message length is 104.7 b.

C. Step 3

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

D. Step 4

The previous generated clusters are remerged and step 3 is repeated until a smaller message length is found. The best message length for three clusters is found, which is 57.31 b.

The result of the aforementioned clustering steps is shown in Table IV.

From Steps 1 and 4 in Table IV, it can be seen that as the number of clusters increases from one to three clusters, the model length is increased accordingly from 5 b to 23.96 b $(10.1 + 6.41 + 7.45)$ to allow for a description of the new

TABLE IV
SEGMENTATION PROCESS OF DATA POINTS IN TABLE III

Step	N	Model L (K) (bit)	Data L(D/K) (bit)	Total Length L (D, K) /cluster (bit)	L(D,K) total (bit)
1	30	5	118.6	123.6	123.6
2	16	6.3	44.2	50.5	104.7
	14	5.3	48.9	54.2	
3	10	3.85	36.32	40.17	108.8
	10	7.72	25.21	32.93	
	10	5.22	29.99	35.21	
4	10	10.1	5.42	15.52	57.31
	10	6.41	10.1	16.51	
	10	7.45	17.83	25.28	

clusters. On the other hand, the data length significantly drops from 118.6 b to 44.52 b $(29.99 + 5.42 + 10.1)$ as the new model was able to compress the message containing the data, resulting in less total message length (i.e., 57.31 b compared to 123.6 b for the original single cluster).

REFERENCES

[1] G. T. Heydt, "Electric power quality: A tutorial introduction," *IEEE Comput. Appl. Power*, vol. 11, no. 1, pp. 15–19, Jan. 1998.

[2] M. McGranagahn, "Trends in power quality monitoring," *IEEE Power Eng. Rev. J.*, vol. 21, no. 10, pp. 3–9, Oct. 21, 2001.

[3] A. K. Khan, "Monitoring power for the future," *IEEE Power Eng. Rev. J.*, vol. 15, no. 2, pp. 81–85, Apr. 2001.

[4] R. E. Morrison and E. Duggan, "Long term monitoring of power system harmonics," *IEEE Colloq. (Dig.)*, no. 120, pp. 2/1–2/3, 1993.

[5] T. C. Shuter, H. T. Vollkommer, and T. L. Kirpatrick, "Survey of harmonic levels on the american electric power distribution system," *IEEE Trans. Power Del.*, vol. 4, no. 4, pp. 2204–2213, Oct. 1989.

[6] E. Duggan and R. E. Morrison, "Prediction of harmonic voltage distortion when a nonlinear load is connected to an already distorted supply," in *Inst. Elect. Eng. Colloq. (Digest)*, 1993, vol. 40, no. 3.

[7] J. Lachaume, T. Deflandre, and M. Meunier, "Harmonics in MV and LV distribution systems—Present and future levels," *Inst. Elect. Eng., Disturbances and Protection in Supply Systems*, 1993.

[8] V. Gosbell, D. Mannix, D. Robinson, and S. Perera, "Harmonic survey of an MV distribution system," in *Proc. AUPEC*, Perth, Australia, Sep. 23–26, 2001, pp. 338–342.

[9] D. Robinson, "Harmonic management in MV distribution system," Ph.D. dissertation, Univ. Wollongong, Wollongong, Australia, 2003.

[10] D. D. Sabin, D. L. Brooks, and A. Sundaram, "Indices for assessing harmonic distortion from power quality measurements: Definitions and benchmark data," *IEEE Trans. Power Del.*, vol. 14, no. 2, pp. 489–496, Apr. 1999.

[11] A. E. Emanuel *et al.*, "A survey of harmonic voltages and currents in customer's bus," *IEEE Trans. Power Del.*, vol. 8, no. 1, pp. 411–421, Jan. 1993.

[12] P. Cheeseman and J. Stutz, "Bayesian classification (AUTOCLASS): Theory and results," in *Advances in Knowledge Discovery and Data Mining*, U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusanny, Eds. Menlo Park, CA: AAAI, 1995.

[13] J. Oliver, R. Baxter, and C. Wallace, "Unsupervised learning using MML," in *Proc. 13th Int. Conf Machine Learning*, 1996, pp. 364–372.

[14] G. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*. New York: Wiley, 1992.

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

[16] *IEC Standard for Electromagnetic Compatibility (EMC)—Part 4–30: Testing and Measurement Techniques—Power Quality Measurement Methods*, IIEC61000-4-30.

[17] S. Elphick, V. Gosbell, and S. Perera, "The effect of data aggregation interval on voltage results," in *Proc. Australasian Universities Power Eng. Conf.*, Perth, Australia, Dec. 2007, pp. 15–02.

[18] J. J. Oliver and D. J. Hand, "Introduction to minimum encoding inference," [TR 4–94], Dept. Stats., Open Univ.

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

[20] D. M. Boulton and C. S. Wallace, "A program for numerical classification," *Comput. J.*, vol. 13, no. 1, pp. 63–69, 1970.

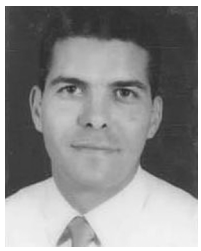
[21] C. Wallace, "Intrinsic classification of spatially correlated data," *Comput. J.*, vol. 41, no. 8, 1968.

[22] R. Xu and D. Wunsch II, "Survey of clustering algorithms," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 645–678, May 2005.

[23] M. A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Trans. Pattern Anal. Mach. Learning*, vol. 24, no. 3, pp. 381–396, Mar. 2002.

[24] Clementine 8.0 User's Guide. Chicago: SPSS Inc., SPSS Inc., 2003.

[25] C. M. Stow, A. C. T. Kenington, C. Milona, and W. Fitzgerald, "Experimental issues of functional merging on probability density estimation," in *Proc. 5th Int. Conf. Artificial Neural Networks*, 1997, pp. 123–128.



Ali Asheibi received the B.Sc. and M.Sc. degrees in electrical engineering from the University of Garyounis, Benghazi, Libya, in 1991 and 2001, respectively, and is currently pursuing the Ph.D. degree in power-quality data analysis data mining at the University of Wollongong, Wollongong, Australia.

He was a Projects and Planning Engineer in distribution systems with G.E.C of Libya from 1992 to 1998. He was an academic at the University of Garyounis from 1999 to 2002.

David Stirling (M'01) received the B.Eng. degree from the Tasmanian College of Advanced Education, Tasmania, Australia, in 1976, the M.Sc. degree in digital techniques from Heriot-Watt University, Scotland, U.K., in 1980, and the Ph.D. degree from the University of Sydney, Sydney, Australia, in 1995.

He has worked for more than 18 years in a wide range of industries, most recently as a Principal Research Scientist with BHP Steel. Currently, he is a Senior Lecturer at the University of Wollongong, Wollongong, Australia. His research interests are in machine learning and data mining.

Danny Sutanto (SM'08) received the B.Eng. (Hons.) and Ph.D. degrees from the University of Western Australia, Perth, Australia.

Currently, he is the Professor of Power Engineering at the University of Wollongong, Wollongong, Australia. His research interests include power system planning, analysis and harmonics, flexible ac transmission systems (FACTS), and battery energy storage systems.

Dr. Sutanto was the Power & Energy Society Region 10 Regional Representative in 2002–2004.