2008

Determination of the optimal number of clusters in harmonic data classification

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

Determination of the Optimal Number of Clusters in Harmonic Data Classification

Ali Asheibi, David Stirling, Danny Sutanto

Abstract-- In many of clustering algorithms, such as K-means and Fuzzy C-mean, the value of the expected numbers of clusters is often needed in advance as an input parameter to the algorithm. Other clustering algorithms estimate this number as the clustering process progresses using various heuristic techniques; however such techniques can also lead to a local minima within the solution space without finding the optimum number of clusters. In this paper, a method has been developed to determine the optimum number of clusters in power quality monitoring data using a data mining algorithm based on the Minimum Message Length technique. The proposed method was tested using data from known number of clusters with randomly generated data points, with data from a simulation of a power system, and with power quality data from an actual harmonic monitoring system in a distribution system in Australia. The results from the tests confirm the effectiveness of the proposed method in finding the optimum number of clusters.

***Index Terms*-- classification, clustering, data mining, harmonics, monitoring system, power quality, segmentation.**

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 usefulness of clustering analysis is that it is easier to deal with groups or clusters rather than the complete data. An expert in the field is usually needed to interpret the discovered clusters. Further analysis is also needed, such as experimental work or simulation to verify the obtained knowledge. There are many different types of clustering in the literature, such as hierarchical (nested), partitional (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 (an object has a membership weight to all clusters) [1]. Clustering has been found to be a useful tool used in many disciplines, such as business, engineering, biology, psychology and medicine [1].

In using the clustering technique for harmonic monitoring data, each cluster can represent a specific operating condition, such as peak load, off-peak load, capacitor switching operation etc. The operating conditions of each of these clusters can be analysed and confirmed by the operation engineers [2]. In this way, clusters due to power quality issues can be identified and be used to identify future occurrence of the power quality problems. Repeated occurrence of these clusters may require countermeasures to be designed to reduce or eliminate the identified power quality issues. 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.

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

The aim of this paper is to develop a method to determine the optimum number of clusters, each of which represents a unique operating condition.

The 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, which is based on the Minimum Message Length (MML) principle. The paper discusses how the number of clusters is decided in ACPro, which shows the tendency of ACPro to overestimate the number of clusters. A method is then proposed to estimate the optimum number of clusters using the exponential method, and the Fitness Function. The proposed method is tested using three different types of data sets, and the results show that the proposed method is effective in finding optimum number of clusters, each of which represent a unique operating condition.

II. Harmonic Monitoring Program

A harmonic monitoring program [3], [4] was installed in a typical 33/11kV MV zone substation in Australia that supplies ten 11kV radial feeders. The zone substation is supplied at 33kV 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

978-1-4244-1770-4/08/$25.00 ©2008 IEEE

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 of a predominant load type. Site 5 was along the feeder route approximately 2km from the zone substation, feeds 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 product such as paper towels, toilet paper and tissues.

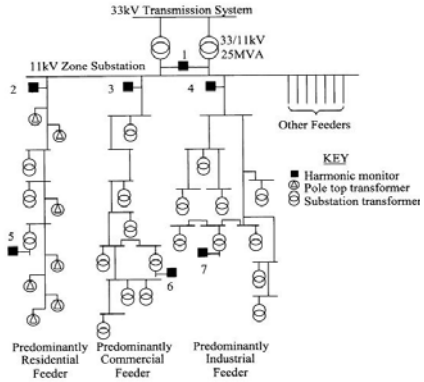

Fig. 1: Single line diagram illustrating the zone distribution system

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.

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 monitoring equipment used is the EDMI Mk3 [5]. 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 the 3rd, 5th and 7th 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 IEC standard IEC61000-4-30 for measurements of harmonic, inter-harmonic and unbalance waveforms. The standard regarded as best practice for power quality measurement recommends 10 min aggregation intervals for routine power quality survey. Each 10 min data represents the aggregate of the 10-cycle rms magnitudes over the 10 min period [6].

The data retrieved from the harmonic monitoring program spans from August 1999 to December 2002. Fig. 2 shows a typical output data from the monitoring equipment of the fundamental, 3rd, 5th and 7th harmonic currents in Phase 'a' at site 2, taken on 12 -19 January 2002. 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

Clustering using Data Mining 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.

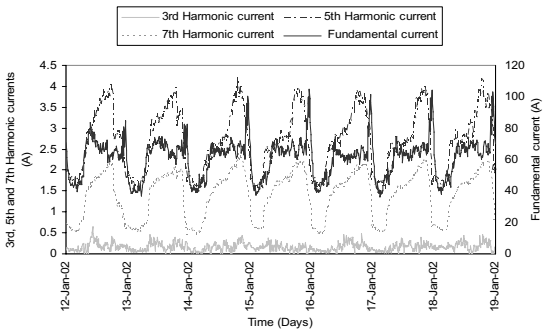

Fig. 2: Residential feeder (site 2) weekly harmonic Current data from the monitoring equipment

The partitioning of data into candidate subgroups is usually subject to some objective function like 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 the Minimum Message Length (MML), is used to evaluate each successive set of segmentations and monitor their progression towards a globally best model. The minimum message length of inductive inference is an invariant Bayesian point estimation and model selection technique based on information theory. In this technique, the measured data is considered as an encoded message. The Minimum Message Length 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 using a mixture of 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 [7], and the best model (shortest total message length) is selected.

The message length in MML method is given as:

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

where:

K : mixture of model clusters
L (K) : the message length of K
L(D/K) : the message length of the data given K
L (D, K): the total message length

Given a data set D and a given accuracy of measurement, AOM, the chosen statistical distribution is initially assumed, such as a Gaussian distribution. Starting from having all the data in one cluster having the chosen distribution (K=1) with a sample mean $\bar{x}$ and standard deviation s, the parameters μ, σ and α (mean, variance and abundance) of this model can be estimated using the Expectation Maximisation algorithm (EM) to fit the Gaussian distribution model [1]. The abundance value, α, 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. A small abundance may mean the cluster represents rare occurrences and this may point out instances when the system needs to be observed more carefully [8]. The single cluster may be subsequently be divided into a mixture of two clusters (K = 2) having the chosen distribution each with its own sample mean $\bar{x}$ and standard deviation s. EM is then used to optimise the parameters μ, σ and α (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 until increasing extra cluster does not result in additional reduction.

IV. EFFECT OF THE NUMBER OF CLUSTERS

To test the effect of the number of clusters, five clusters of data points (D's) were randomly generated (D1, D2,…, D5), 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 Figs. 3(a-d).

Figs. 3(a) and 3(b) show that underestimation of the number of clusters will result in having clusters with a combination of D's. Fig. 3(a) shows that one of the clusters represents D1 and D2 and the other D3, D4 and D5. Fig. 3(b) shows that D1, D2 and D3 are identified correctly, but D4 and D5 are identified as one cluster. Fig. 3(d) illustrates that the overestimation generated by ACPro, was due to its inadequate stopping criterion, producing spurious clusters representing the data of higher variances. Fig. 3(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 needs to have an optimal way of deciding the correct number of clusters from a given data set.

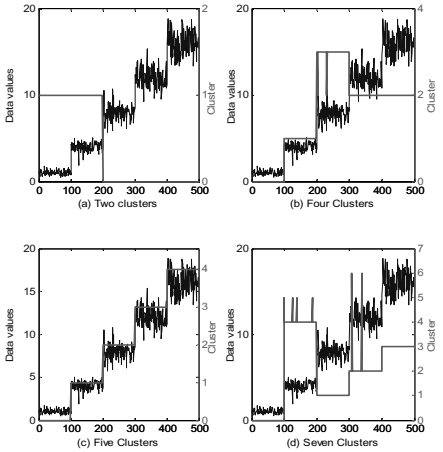

Fig. 3: The clusters obtained superimposed on the randomly generated data: (a) 2 clusters, (b) 4 clusters, (c) 5 clusters, and (d) 7 clusters

V. USING FITNESS FUNCTION TO DETERMINE THE OPTIMAL NUMBER OF CLUSTERS

From information theory, fitness function [9] can be used as a criterion to determine the optimum number of clusters when mixture modelling method is used for data fitting. The higher the fitness function value the better the data fit. Here, the fitness function gains maximum information from data by maximizing the entropy of its groupings. This maximum entropy is fulfilled if the data set can be modelled as a mixture of Gaussian distributions

The theoretical maximum entropy H_{max} of any distribution can be calculated as follows [10].:

$$H_{\max}(C_i) = \frac{1}{2}\log((2\pi e)^n \mid \mathrm{cov}(C_i) \mid) \quad (2)$$

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 can be calculated from the maximum entropy equation in (2) as follows:

$$ef_i = \frac{H(C_i)}{H_{\max}(C_i)} \quad (3)$$

where

$H(C_i)$ is the entropy of C_i

$$H(C_i) = -\sum_i P_i \log_2 P_i \quad (4)$$

The total fitness function EF_T from ef_i can be calculated from the individual fitness function ef_i given in (3) as:

$$EF_T = \sum_{i=1}^{k} \alpha_i \mid ef_i \mid \quad (5)$$

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. Thus, the largest value of the fitness function EF_T should correspond to the optimum number of clusters required for the data.

A recent study [11] shows that the entropy fitness function can determine the right number of clusters to correctly identify the anomalies in intrusion detection data.

When applied to the five clusters randomly generated discussed in Section IV, Fig. 4 shows how the fitness function increases and reaches maximum when the total number of cluster is 5, suggesting that such a method is suitable to determine the optimum number of clusters.

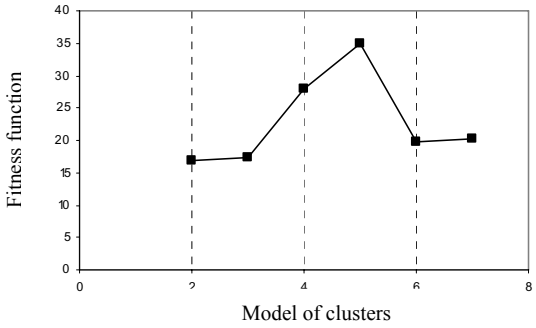

Fig. 4: Fitness function showing five clusters in random data

However the fitness function may fail to find the optimum number of clusters if the input attributes are correlated, because the maximum entropy equation in (1) assumes that the input attributes are independent variables.

Because the harmonic data in the four substations described in Section II, is correlated through the network equation, it is likely that the fitness function will have difficulty in determining the optimum number of clusters for the harmonic monitoring data. This will be discussed in Section VII.

VI. Proposed Method of Determining Optimal Number of Clusters Using MML

Section V shows that while 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 [12].

In our study, we have found that when the difference between the message lengths of two consecutive mixture models is close to zero and stays close to zero, then it can be inferred that the two models are similar. 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.

It has been shown that minimizing the message length in an MML technique is equivalent to maximizing the posterior probability in Bayesian theory [13].

However, we propose to further emphasize this difference by calculating the exponential of the change in message length for consecutive mixture models which 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 should be determined to be the optimum number of clusters.

When the proposed method is applied to the five randomly generated clusters given in Section IV as shown in Fig. 5, it is clear that 5 is the optimum number of cluster, since going to 6 and 7 clusters resulted in very small changes of the exponential of the message difference.

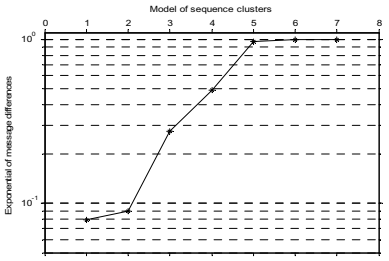

Fig.5: Exponential message difference curve with five clusters as optimum number

VII. Simulation Results

To test the proposed method, a simulation of a simplified power system (shown in Fig. 6) is carried out using PSCAD™/EMTDC®. Three switches are used to represent 8 operating conditions depending of which switch is turned ON or OFF. The switching operation and the times of switching are shown in Table I.

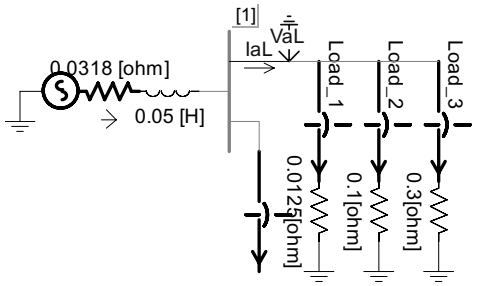

Fig. 6: A single line diagram of a simplified power system model used in the PSCAD Simulation.

Table I: The load switching operation and timing

Cluster No	Time		Load_1 on/off	Load_2 on/off	Load_3 on/off
	on (s)	off(s)			
6	0	0	0	0	0
5	1.25	2.50	0	0	1
7	2.5	3.75	0	1	0
1	3.75	5.00	0	1	1
0	5.00	6.25	1	0	0
2	6.25	7.5	1	0	1
3	7.5	8.75	1	1	0
4	8.75	10.00	1	1	1

Fig. 7(a) shows the rms voltage and current at phase ‘a’ at bus 1. Using these two variables as the two input attributes to ACPro, Fig. 7(b) shows the exponential of the difference of the message length of consecutive mixture models. Ten clusters were found to be the optimum number. Figure 7(c) shows the 10 clusters (s0,s1,s2, …s9) superimposed on the two input attributes.

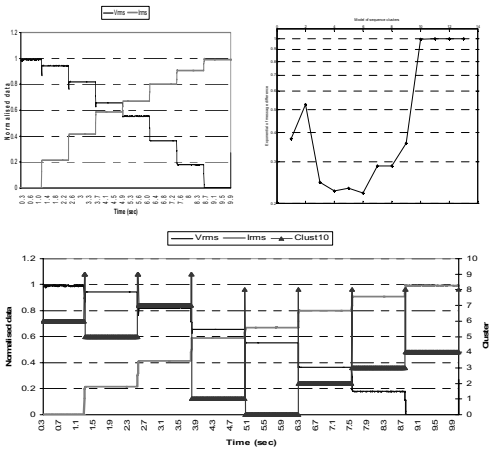

Fig. 7: a) The rms values of voltage and current in phase 'a', b) Exponential of the message length difference of consecutive clusters, c) clusters superimposed on simulation data

This is a very interesting result, because we were expecting to have only 8 clusters, however because of the inductance in the source, transient events can be observed in Fig. 7(a) and 7(c) at each switching point, and the MML method has identified these transients as two separate clusters – at the instant of switching at 1.25, 2.5 and 3.75 seconds – and another one at the other switching times. Looking at Table I, it can be observed that the first cluster is associated with load 1 being OFF and the second cluster is due to load 1 being ON. Fig. 7(a) 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.

Applying the fitness function to the same two attributes, produces the same optimum number as shown in Fig. 8. The highest fitness function is found at 10 clusters.

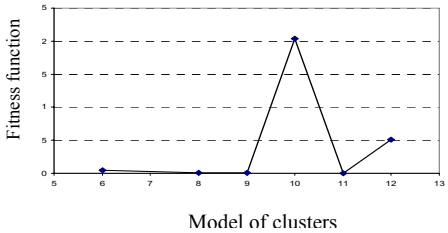

Fig. 8: Fitness function showing the optimum number of cluster

VIII. STUDY SYSTEM

To illustrate the use of the exponential of 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 from buses 1, 2, 3 and 4 taken on 12 -19 January 2002 were used as the input attributes to ACPro. The trend in the exponential message length difference for consecutive pairs of mixture models is shown in Fig. 9.

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, 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 when the exponential difference in message length 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 s 15 related to the on-peak load period.

With the help of the operation engineers, the sixteen clusters detected by this exponential method were interpreted as given in Table II. It is virtually impossible to obtain these 16 unique events by visual observation of the waveforms shown in Fig.10.

The fitness function method is then applied to the same data from the harmonic monitoring data as shown in Fig. 11. The highest fitness function is 5, which suggest the optimum number of clusters should be 5. The reduction in the number of 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 fitness function equations assume that the attributes are independent [10].

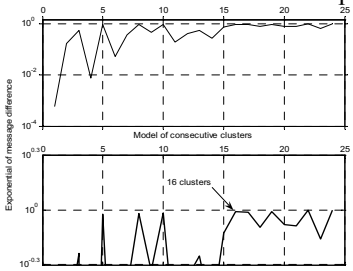

Fig. 9: Exponential curve detect sixteen clusters of harmonic data

IX. CONCLUSION

The optimal number of clusters in three different types of data sets was investigated using a 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 from a distribution system in Australia. A commonly used fitness function technique is found to produce underestimation because of the correlated natures of the attributes presented to the MML program. 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.

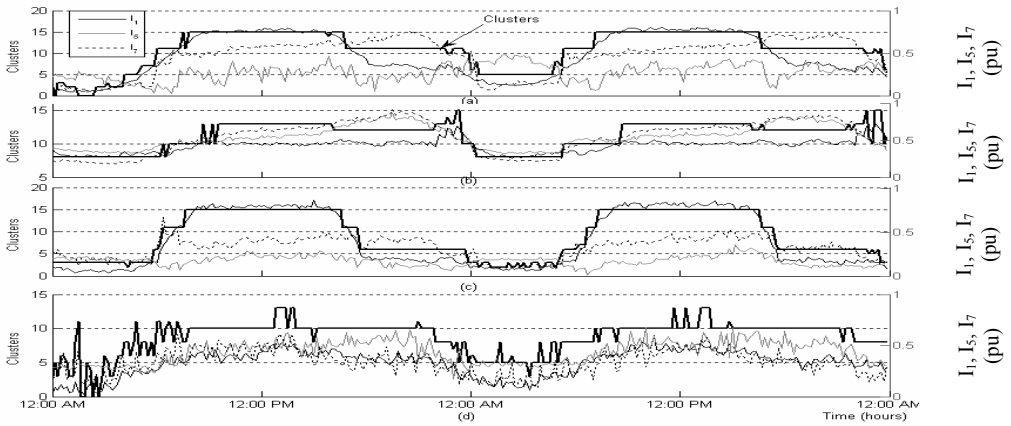

Fig. 10: Sixteen clusters superimposed on four sites (a) Substation, (b) Residential, (c) Commercial and (d) Industrial

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
s4	Off peak at Industrial site
s5	Off peak at Substation site
s6 and s7	Switching on and off of capacitor at Substation site
s8	Ramping load at industrial site
s9	Switch on harmonic load at industrial
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
S14	Ramping load at substation due to commercial
S15	On peak load at substation due to commercial

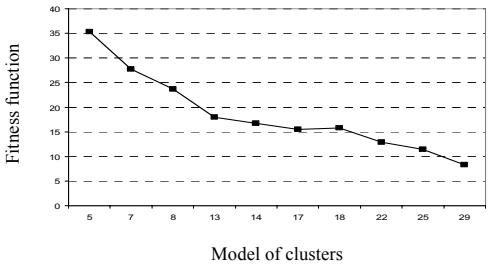

Fig.11: Fitness function showing only five clusters as optimum number

X. REFERENCES

[1] T. Pang, M. Steinbach, V. Kumar "Introduction to Data Mining", Pearson Education, Boston, 2006.

[2] A. Asheibi, D. Stirling, D. Soetanto "Analyzing Harmonic Monitoring Data using Data Mining" Australian Data Mining Conference ADMC06, Nov. 2006, Sydney, Australia

[3] V. Gosbell, D.Mannix, D. Robinson, and S. Perera, , "Harmonic Survey of an MV distribution system." Proc. AUPEC, 23-26 September 2001, Perth, pp. 338-342.

[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.

[6] IEC Standard for Electromagnetic Compatibility (EMC) – part 4-30: Testing and measurement Techniques – Power Quality Measurement Methods, IEC61000-4-30, 2001.

[7] J. J. Oliver and D. J. Hand, Introduction to Minimum Encoding Inference, [TR 4-94] Dept. Stats. Open University.

[8] A. Asheibi, D. Stirling, D. S. Perera, D. Robinson "Power quality data analysis using unsupervised data mining" AUPEC'04, Brisbane, Australia

[9] W. Lu and I. Traore "Determining the Optimal Number of Clusters Using a New Evolutionary Algorithm" Proceedings of the 17th IEEE International Conference on Tools with Artificial Intelligence (ICTAI05). 2005.

[10] N.J. Hoboken, "Elements of information theory" Wiley-Interscience, 2006

[11] W. Lu and I. Traore. "An unsupervised anomaly detection framework for network intrusions" Dept. of Electrical and computer engineering, University of Victoria, October 2005.

[12] A. Asheibi, D. Stirling, D. Robinson "Identification of Load Power Quality Characteristics using Data Mining" IEEE Canadian Conference on Electrical and Computer Engineering", May 2006 Ottawa, Canada

[13] C. S. Wallace, D. L. Dowe. "MML clustering of multi-state, Poission, von Mises circular and Gaussian distributions", Statistics and Computing, 10(11):73-83,2000.

XI. BIOGRAPHIES

Ali Asheibi Mr. Asheibi received his BSc. and MSc. degrees in electrical engineering from the University of Garyounis, Libya in 1991& 2001. His work experience was with G.E.C of Libya as a projects and planning engineer in distribution systems between 1992 and 1998. He was an academic at the University of Garyounis from1999 to 2002. He then joined the University of Wollongong in 2003 and studying towards his PhD in Power Quality data analysis Data Mining.

David Stirling (M' 2001): Dr Stirling obtained his BEng degree from the Tasmanian College of Advanced Education (1976). He further obtained his MSc degree (Digital Techniques) in Digital Techniques from Heriot-Watt University, Scotland UK (1980), and his PhD from the University of Sydney (1995). He has worked for over 18 years in wide range of industries, most recently as a Principal Research Scientist with BHP Steel. He has recently taken up a position as Senior Lecturer at the University of Wollongong. His research interests are in Machine Learning and Data Mining.

Danny Sutanto (SM '77) obtained his BEng. (Hons) and PhD from the University of Western Australia. He is presently the Professor of Power Engineering at the University of Wollongong, Australia. His research interests include power system planning, analysis and harmonics, FACTS and Battery Energy Storage systems. He was the PES Region 10 Regional Representative in 2002-2004. He is a Senior Member of IEEE.