Exponential Method for Determining Optimum Number of Clusters in Harmonic Monitoring Data

A. Asheibi, D. Stirling, and D. Sutanto

Abstract— **Clustering is an important process for finding and describing a variety of patterns and anomalies in multivariate data through various machine learning techniques and statistical methods. Determination of the optimum number of clusters in data is the main difficulty when applying clustering algorithms. In this paper, an exponential method has been proposed to determine the optimum number of clusters in power quality monitoring data using an algorithm based on the Minimum Message Length (MML) technique. The optimum number of clusters has been verified by the formation of super-groups using Multidimensional Scaling (MDS) and link analysis with power quality data from an actual harmonic monitoring system in a distribution system in Australia. The results of the obtained super-group abstractions confirm the effectiveness of the proposed method in finding the optimum number of clusters in harmonic monitoring data.**

Index Terms—**Harmonic monitoring, data mining, clustering**

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 counter measures 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 is 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 exponential method is then proposed to estimate the optimum number of clusters using the message length difference (MLD). The proposed method is verified using Multidimensional Scaling (MDS) and link analysis with harmonic monitoring data, and the results of the tests show that the proposed method is effective in finding the optimum number of clusters, each of which represent a unique operating condition.

II. Harmonics Monitoring System

To illustrate the use of the data mining analysis tools harmonic monitoring results from three MV electricity utility customers on a typical MV distribution system were obtained. Data from the source end of the MV feeders supplying the customers and the HV/MV substation transformer supplying the distribution network was also obtained, as shown in Fig. 1. Although not selected specifically for the application of data mining the test system involved capturing PQ data using standard parameters and monitoring intervals and thus it was perceived the true applicability of data mining to PQ data would be illustrated.

Manuscript received January 8, 2012; revised March 11, 2012.

A. Asheibi is with the Department of Electrical Engineering, Faculty of Engineering in Benghazi University, Benghazi, Libya (e-mail: ali.asheibi@benghazi.edu.ly).

D. Stirling and D. Sutanto are with the School of Electrical Engineering, University of Wollongong, and member of the Endeavour Energy Power Quality and Reliability Centre, NSW 2522, Australia (email: stirling@uow.edu.au; soetanto@uow.edu.au)

The monitored data included voltage and current readings of the fundamental, THD, and 3rd, 5th, and 7th harmonics every 10 minutes over a period of two weeks. Measurements were taken from the HV/MV zone substation transformer voltage transformers (VT) and current transformers (CT) at the MV feeder CTs and the LV side of each customer's 11kV/430V distribution transformer. The selected customers represented different load types, i.e. primarily residential, commercial or industrial sites. The locations of PQ monitoring devices at Sites 1-7 are illustrated in Fig. 1.

The residential site consists primarily of residential homes in an inner suburban location. The commercial site is a large shopping centre operating seven days a week. The industrial site is a medium sized factory manufacturing paper products such as paper towelling [3].

III. Unsupervised Clustering With 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 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.

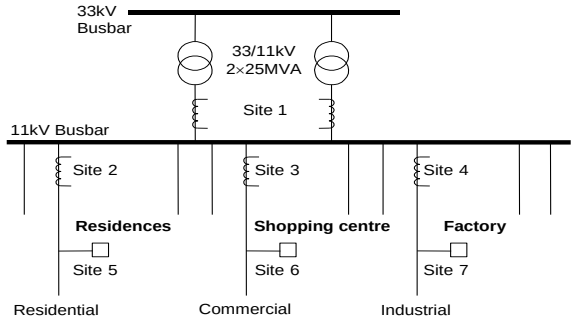

Fig. 1. Schematic layout of test system

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 towards a globally best model. 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 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 [4]:

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

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

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) [5].

$$L(K) = \log_2 \frac{range_\mu}{AOPV_\mu} + \log_2 \frac{range_\sigma}{AOPV_\sigma} \quad (2)$$

where:

$range_\mu$: range of possible μ values

$range_\sigma$: range of possible σ values

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

$$AOPV_\mu = \bar{s}\sqrt{\frac{12}{N}} \quad (3)$$

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

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

where:

N : number of data samples

$\bar{x}$: the sample mean

x_i : data points

AOPVσ : accuracy of the parameter value of σ

$$AOPV_\sigma = \bar{s}\sqrt{\frac{6}{N-1}} \quad (5)$$

The message length of the data using Gaussian distribution model can be calculated from the following equation [5]:

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

where:

Aom : accuracy of measurement

s : sample standard deviation

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

IV. Exponential Method of Determining Optimal Number of Clusters Using MML

One difficulty with the MML algorithm used in the mixture modelling method is the difficulty in establishing stopping criterion to secure optimum number of (mixture) clusters during the clustering process. During the

investigation, it was discovered early that a method has to be found to determine the optimum number of clusters using the MML technique, since overestimating the number of clusters will produce a large number of clusters each of which may not necessarily represent truly unique operating conditions, whereas underestimation leads to only small number of clusters each of which may represent a combination of specific events. (fewer) super–groups. In this study, it has been 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 [6].

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.

V. Results and Outcomes

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. 2.

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 s15 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 I. It is virtually impossible to obtain these 16 unique events by visual observation of the waveforms shown in Fig. 3.

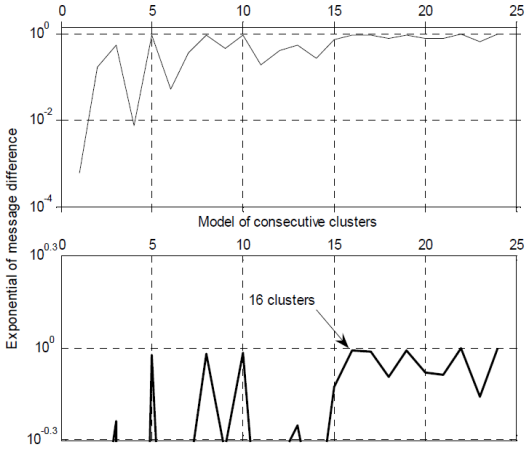

Fig. 2. Exponential curve detect sixteen clusters of harmonic data

TABLE I: 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

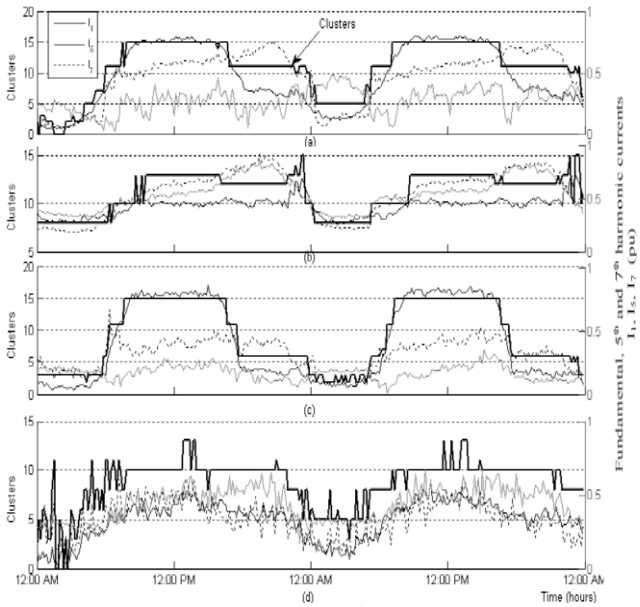

Fig. 3. Sixteen clusters superimposed on four sites (a) substation, (b) residential, (c) commercial and (d) industrial

VI. Verification of the Optimum Model Using Super-Groups

Fig. 2 shows that when the difference between the message

lengths of two consecutive mixture models is close to zero (or its exponential is close to 1) and stays close to zero (or its exponential stays close to 1), then it can be inferred that the two consecutive models are similar. The later model has been formed by splitting one or more of the clusters in the previous model into two or more similar clusters. These similar clusters can be re-merged into super-groups in order to return to the optimum model. This suggests that the super-group techniques can be applied to the MML algorithm to reduce the total number of clusters to the optimum value. This suggests that the super-group technique is a good technique to verify the proposed method of using the trend of the exponential of message length difference to obtain the optimum number of clusters.

To verify this, the same data from previous section was used as an input to ACPro, but now allowing ACPro to produce the maximum number of clusters (30 clusters). The trend in the exponential of message length difference for consecutive pairs of mixture models of the 30 clusters. The clusters are subsequently sorted in ascending order based on the mean value of the fundamental current, such that cluster s0 is associated with the off-peak load period and cluster s29 related to the on-peak load period. The Kullback Liebler distances (KL) [7] between the 30 clusters are sorted from the lowest to the highest (the most similar clusters to the most difference ones). A multidimensional scaling algorithm (MDS) [8, 9], which is a dimension reduction technique is then used to form a network from the KL distances. The MDS Knowledge Network Organising Tools (KNOT) software [10] is used to form the super-group abstractions by removing the links whose distances exceed a dissimilarity threshold (in this case when KL distance is less than 13.5 Bits). Using this technique and the defined dissimilarity threshold, it was found that sixteen super-groups are obtained as visualised in Fig. 4.

This is the same number of clusters obtained from the proposed method of determining optimum number of clusters usig the trend of the exponential of message length difference. Table II shows the relationship between the sixteen super-groups obtained using the MDS method and the optimum sixteen clusters obtained from the proposed method based on the trend of message length difference.

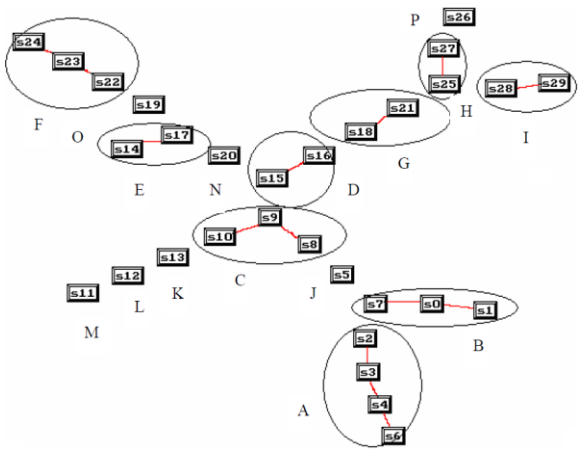

Fig. 4. Sixteen super-groups obtained from the generated 30 clusters

TABLE II. ALIGNMENT BETWEEN OPTIMUM 16 CLUSTERS AND SUPER-GROUPS

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

VII. CONCLUSION

The optimal number of clusters in power quality data was investigated using a proposed method based on the trend of the exponential difference in message length between two consecutive mixture models. The results show that the suggested method is effective in determining the optimum number of clusters in harmonic monitoring data from a distribution system in Australia. To validate the optimum number of clusters obtained, the MDS method was used to form super-groups with a defined dissimilarity threshold. It was found that similar number of super-groups is obtained as the optimum number of clusters obtained from the proposed method based on the trend of the exponential difference in message length between two consecutive mixture models. The MDS method shows how the clusters obtained from an overestimation of the number of clusters can be merged to form the optimum number of clusters. Correct determination of the number of system unique operating conditions is important in the diagnosis of power quality disturbances as well for prediction of these events in the future.

REFERENCES

[1] T. Pang, M. Steinbach, and V. Kumar "Introduction to Data Mining," *Pearson Education*, Boston, 2006.
[2] A. Asheibi, D. Stirling, and D. Soetanto "Analyzing Harmonic Monitoring Data using Data Mining" Australian Data Mining Conference ADMC06, Nov. 2006, *Sydney*, Australia
[3] D. Robinson, "Harmonic Management in MV Distribution System" PhD Thesis, *University of Wollongong*, 2003.
[4] C. Wallace, "Intrinsic Classification of Spatially Correlated Data," *The Computer Journal*, Vol. 41, No. 8, 602-611, 1998.
[5] J. J. Oliver and D. J. Hand, Introduction to Minimum Encoding Inference, [TR 4-94] *Dept. Stats.* Open University.
[6] C. S. Wallace and D. L. Dowe. "MML clustering of multi-state, Poission, von Mises circular and Gaussian distributions," S*tatistics and Computing*, 10(11):73-83, 2000.
[7] K. Solomon, Information *Theory and Statisics*. New York: Dover Publications, Inc. 1997.

[8] J. B. Kruskal and M. Wish. Multidimentional scaling. Sage University Papers, 7, no11, 1978.
[9] W. Schvaneveldt, *Pathfinder Associative Netorks*. New Jersey: Alpex. 1990.
[10] Knowledge Network Organisation tool KNOT. www.interlinkinc.net/KNOT.html, Accessed, 24 August 2009.

A. Asheibi obtained his BSc. and MSc. degrees in electrical engineering from the University of Benghazi, 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 Benghazi from1999 to 2002. He then joined the University of Wollongong in 2003 and completed his PhD in Power Quality data analysis using Data Mining in 2009. He is currently a lecturer at Benghazi University, Benghazi, Libya. Dr. Asheibi is a member of IACSIT.

D. 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. He is a member of IEEE.

D. Sutanto 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.