2010

Clustering of harmonic monitoring data using data mining

Danny Sutanto

University of Wollongong, soetanto@uow.edu.au

A. Asheibi

University of Wollongong, ali_asheibi@uow.edu.au

David A. Stirling

University of Wollongong, stirring@uow.edu.au

Follow this and additional works at: <https://ro.uow.edu.au/infopapers>

Part of the [Physical Sciences and Mathematics Commons](#)

Recommended Citation

Sutanto, Danny; Asheibi, A.; and Stirling, David A.: Clustering of harmonic monitoring data using data mining 2010.

<https://ro.uow.edu.au/infopapers/3513>

Clustering of harmonic monitoring data using data mining

Abstract

A comprehensive harmonic monitoring program has been designed and implemented on a typical 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 data mining based on the minimum message length technique comprising unsupervised learning to assign cluster designation and supervised learning to describe the generated clusters and to predict the occurrences of unusual clusters in future measurement data. The clusters obtained from the unsupervised learning process provide the engineers with a rapid, visually oriented method of evaluating the underlying operational information contained within the clusters.

Disciplines

Physical Sciences and Mathematics

Publication Details

D. Sutanto, A. Asheibi & D. A. Stirling, "Clustering of harmonic monitoring data using data mining," in 14th International Conference on Harmonics and Quality of Power, ICHQP 2010, 2010, pp. 1-6.

Clustering of Harmonic Monitoring Data using Data Mining

D. Sutanto, *Senior Member, IEEE*, A. Asheibi, and D. Stirling, *Member, IEEE*

Abstract-- A comprehensive harmonic monitoring program has been designed and implemented on a typical 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 data mining based on the minimum message length technique comprising unsupervised learning to assign cluster designation and supervised learning to describe the generated clusters and to predict the occurrences of unusual clusters in future measurement data. The clusters obtained from the unsupervised learning process provide the engineers with a rapid, visually oriented method of evaluating the underlying operational information contained within the clusters.

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

I. INTRODUCTION

Harmonic monitoring is an important issue for electricity utilities and their customers due to the increasing use of power quality distorting loads, increasing penetration of equipment susceptible to power quality disturbances and the competition in the distribution networks [1]. Power quality disturbances resulting from distorting loads can cause significant financial impact due to loss of production, damage to equipment, and disruption on related manufacturing processes [2]. 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 [3]-[8].

A harmonic monitoring program was recently conducted by a local electricity distributor in Australia to measure the harmonic currents and voltages in a medium voltage (MV)

distribution system [9], [10]. The monitoring involved simultaneous measurements of the three-phase harmonic current and voltage from the residential, commercial, and industrial load sectors. 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 time. A more sophisticated analysis method is required to automatically segment the data into 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 [11] and Snob programs [12] and uses mixture models [13] 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.

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. The paper discusses the significance of the clusters obtained and how the associated operational conditions can be deduced from these clusters. The use of 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 33/11kV 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 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

This work was financially sponsored by higher education administration of Libyan government under grant HEA 614.

D. Sutanto, A. Asheibi and D. Stirling, are with the School of Electrical Engineering, University of Wollongong, NSW 2522, Australia, and are members of the Integral Energy Power Quality and reliability Centre, NSW 2522, Australia (email: soetanto@uow.edu.au)

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.

The monitoring equipment used is the EDM Mk3 Energy Meter from Electronic Design and Manufacturing Pty. Ltd. [14] shown in Fig. 2.

Fig.2. EDM 2000-04XX Energy Meter.

Three phase voltages and currents at sites 1-4 were recorded at the 11kV zone substation and at sites 5-7 were recorded at the 430V side of the 11kV/430V distribution transformer, as shown in Fig. 1. The memory capabilities of the above meters at the time of purchase limited recordings to

the fundamental current and voltage in each phase, the current and voltage THD in each phase, and three other individual harmonics in each phase.

For the harmonic monitoring program, the harmonics chosen to be recorded were the 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 minutes. This follows the suggested measurement time interval by the International Electrotechnical Commission (IEC) standard as given in IEC61000-4-30 for measurements of harmonic, inter-harmonic and unbalance waveforms. The standard regarded as best practice for power quality measurement recommends 10 minute aggregation intervals for routine power quality survey. Each 10 minutes data represents the aggregate of the 10-cycle rms magnitudes over the 10 minutes period [15]. Further, a recent study [16] suggested that statistically, sampling at faster rate will not provide additional significant extra insight.

The data retrieved from the harmonic monitoring program spans from August 1999 to December 2002. Fig. 3 shows a typical output data from the monitoring equipment of the fundamental, 3rd, 5th and 7th harmonic currents in Phase 'a' at site 1, taken on 12 -19 January 2002 showing a 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.

Fig. 3. Zone substation (site 1) weekly harmonic current data from the monitoring equipment.

III. CLUSTERING USING DATA MINING BASED ON MML

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.

In this paper, a technique based on Minimum Message Length (MML) [12], 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 first application of mixture modelling method using the MML technique was suggested by Wallace et al [17] and based on this proposal a classification program called Snob was written. The program was successfully used to classify groups of six species of fur seals. Since then, the program has been extended and utilised in different areas, such as psychological science, health science, bioinformatics, protein and image classification [18]. In this project, ACPro software [13], which is based on MML technique and uses mixture models, is used to represent clusters.

The use of MML clustering algorithm for Power Quality classification has several advantages over traditional methods. One advantage of applying MML technique in power quality 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. As mentioned in Section II, the data available was measured at 10 minute aggregation interval, with each measurement represents the aggregate of the 10-cycle rms magnitudes over the 10 minutes period.

For this reason, we have chosen the mixture modelling program based on MML for automatic clustering of the harmonic database [12]. The software allows 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 of the constituent clusters.

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, 5th and 7th harmonic currents) were selected from different sites (sites 1, 2, 3 and 4). The 3rd harmonic current was excluded as its level was low due to the presence of Δ/Y transformers downstream, which block most of the 3rd harmonic current from flowing up as shown in Fig. 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. 4.

Fig. 4. Message length vs. number of generated clusters.

Using a basic spreadsheet tool the clusters are subsequently sorted in ascending order (s0, s1, s2, s3, s4 and s5) based on the mean value of the fundamental current, such that cluster s0 is associated with the off peak load period and cluster s5 related to the on-peak load period as shown in Fig. 5.

Fig. 5. The clusters obtained superimposed on the phase 'a' fundamental waveform at substation site (site 1).

Each generated cluster can therefore be considered as a profile of the three variables (fundamental, 5th, 7th harmonic currents) within an acceptable variance. If new data lies beyond the variance, another cluster is created (see Fig. 6).

Fig. 6. The clusters statistical parameters mean (μ), standard deviation (σ) and abundance (p).

From this sorting process one can see that cluster s5 not only has the highest fundamental current, but also the highest 5th harmonic current. This infers that the high 5th harmonic currents are due to an overloading condition. Fig. 5 also shows

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

TABLE I
THE ABUNDANCE VALUES FOR EACH GENERATED CLUSTER

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

The visualisation of the six clusters at site 1 is shown in Fig. 7, showing the relationship among fundamental, 5th and 7th harmonic currents and the obtained clusters.

Fig. 7. The six clusters obtained at substation site (site 1).

A. Interpretations of the results from clusters obtained using MML

By observing how the measured data are classified into various clusters, the power utility engineer can more readily deduce the power quality event that may have triggered a change from one cluster to another cluster. To confirm the observation, other available data 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. 8(a), shows the clusters obtained from substation site (site 1) superimposed on the fundamental current measurement data for two days. Fig. 8(b), shows the 7th harmonic current and 7th harmonic voltage at the substation.

By observation, it appears that this is due to sudden changes in the 7th harmonic current. After further investigation of the MVar measurement at the 33kV side of the power system shown in Fig. 8(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. 8(c), the capacitor is switched on in the 33kV side to reduce bus voltage and late at night when the system MVar demand is low, the capacitor is switched off to avoid excessive voltage rise.

Fig. 8. Clusters at substation site in two working days: (a) Clusters superimposed on the fundamental current waveform, (b) 7th harmonic current and voltage data. (c) MVar load at the 33kV.

By just observing the fundamental current, it is difficult to understand why the second cluster has been generated. The 7th harmonic current and voltage plots as shown in Fig. 8(b), provide a clue that something is happening during cluster s2, in that the 7th harmonic current increases rapidly and 7th 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 33kV side can also be detected at the other sites (sites 2, 3 and 4) at the 11kV side.

Although in this case the cause can be easily uncovered, there may be other cases where the clustering process can identify a cluster which can produce detrimental effect 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, 5th and 7th 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 a high fundamental currents and high 5th harmonic voltage are the characteristics of this cluster. This is shown in Fig. 9, for a period of three days when the temperature is normal for the time of the year. Fig. 10, shows the results for a period of three days when the weather is very hot, resulting in significant use of air conditioners. There is usually a lag (human response) between the peak temperature and the onset of the peak use of air conditioner, thereby causing a noticeable lag between the peak

temperature and the sudden increase in the fifth harmonic as shown in Figs. 10(b) and 10(c).

Fig. 9. Three normal temperature days at the residential site (site 2): (a) Fundamental current and generated clusters, (b) 5th harmonic voltage and generated clusters, (c) The temperature near site 2.

Fig. 10. Three hot days at the residential site (site 2): (a) Fundamental current and generated clusters, (b) 5th harmonic voltage and generated clusters, (c) The temperature near site 2.

Fig. 11, shows the difference in harmonic clusters at residential site between the normal weather days and the hot days. It is evident that the MML has identified s5 cluster occurring 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 discussion with the utility engineer, we were told that this is related to the turning-on of the off-peak water heaters.

V. CONCLUSION

Power quality (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 based on MML. The technique presented in this paper allows utility engineers to detect unusual PQ events from monitored sites, using clustering to infer information about future PQ performance at the monitored sites. The clustering analysis using MML has been shown to be able to identify useful patterns within PQ monitored data set. The advantage of mixture models, in general, is that it can cope with different types of distributions.

Fig. 11. Normal and hot days at residential site (site 2).

The main difficulty with the MML technique is to determine the optimum number of clusters associated with the global minimum message length. Further work is currently being carried out to achieve this.

VI. REFERENCES

- [1] B.W. Kennedy. *Power Quality Primer*. New York: McGraw-Hill, 2000.
- [2] R. Lamedica, G. Esposito, E. Tironi, D. Zaninelli, and A. Prudenzi, "A survey on power quality cost in industrial customers," in *Proc. 2001 IEEE Power Engineering Society Winter Meeting*, Vol. 2, No. 3, pp. 938 – 943.
- [3] McGranaghan, M, "Trends in Power Quality Monitoring", *IEEE Power Engineering Review Journal*, vol. 21, no. 10, pp. 3-9, 21, October 2001.
- [4] Khan, A.K., "Monitoring Power for the future", *IEEE Power Engineering Review Journal*, vol. 15, No. 2, pp.81-85, April 2001.
- [5] Morrison, R.E. and Duggan, E., "Long Term Monitoring of power system harmonics", *IEEE Colloquium (Digest)*, 1993 (Issue 120), pp 2/1-2/3.
- [6] Shuter, T.C., Vollkommer, H.T., Kirpatrick, T.L., "Survey of harmonic levels on the American Electric Power Distribution System", *IEEE Trans. Power Delivery*, 1989, Vol. 4, No. 4, pp. 2204-2213
- [7] Duggan, E. and Morrison, R.E., "Prediction of harmonic voltage distortion when a nonlinear load is connected to an already distorted supply," *IEE Colloquium (Digest)*, 1993, Issue 40, No 3.
- [8] Lachaume, J., Deflandre, T. and Meunier, M., "Harmonics in MV and LV distribution systems - Present and Future Levels", in *Proc 1993 IEE Disturbances and protection in Supply Systems Conf.*, pp. 2.1.1 – 2.1.5.
- [9] Gosbell, V., Mannix, D., Robinson, D. and Perera, S., "Harmonic Survey of an MV distribution system." *Proc. 2001 AUPEC*, 23-26 September 2001, Perth, pp. 338-342.
- [10] Robinson, D., "Harmonic Management in MV Distribution System" PhD Thesis, University of Wollongong, 2003.
- [11] Cheeseman, P., and Stutz, J., "Bayesian Classification (AUTOCLASS): Theory and Results, in Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. and Uthurusanny, R., eds, "Advances in Knowledge Discovery and Data Mining," the AAAI press, Menlo Park, 1995.
- [12] Oliver, J., Baxter, R. and Wallace, C., "Unsupervised Learning using MML, in "Machine Learning: *Proc. 13th Int. Conf ICML '96*, pp. 364-372, 1996.
- [13] McLachlan, G., "Discriminant Analysis and Statistical Pattern Recognition", Wiley, New York, 1992.
- [14] EDM, *Users Manual - EDM 2000-04XX Energy Meter*. Electronic Design and Manufacturing International.
- [15] *IEC Standard for Electromagnetic Compatibility (EMC) – part 4-30: Testing and measurement Techniques – Power Quality Measurement Methods*, IEC61000-4-30, 2001.
- [16] S. Elphick, V. Gosbell, S. Perera, "The Effect of Data Aggregation Interval on Voltage Results" *Proc. Australasian Universities Power Engineering Conference AUPEC07*, Dec. 2007, Perth, Australia, Paper 15-02.
- [17] Wallace C. S. and D.M. Boulton, "An information measure for classification," *The Computer Journal*, 11, no 2:185–194, 1968.

- [18] Y. Agusta. Minimum message length mixture modelling for uncorrelated and correlated continuous data applied to mutual funds classification. PhD thesis, School of Computer Science and Software Engineering, Monash University, Melbourne, Australia, 2004.
- [19] D. M. Boulton and C. S. Wallace. A program for numerical classification. The Computer Journal, 13(1): 63-69, 1970.
- [20] C. Wallace, "Intrinsic Classification of Spatially Correlated Data", The Computer Journal, Vol. 41, No. 8, 1968.
- [21] R. Xu, D. Wunsch II, "Survey of Clustering Algorithms", IEEE Transactions on Neural Networks, Vol. 16, No 3, pp. 645-678, May 2005.
- [22] T. Pang, M. Steinbach, V. Kumar "Introduction to Data Mining", Pearson Education, Boston, 2006.
- [23] SPSS Inc. 2003. Clementine 8.0 User's Guide. Chicago: SPSS Inc.

VII. BIOGRAPHIES

Danny Sutanto (SM) 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.

Ali Asheibi received his BSc. and MSc. degrees in electrical engineering from the University of Garyounis, Libya in 1991 & 2001 and PhD from the University of Wollongong, Australia. 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 from 1999 to 2002. He then joined the University of Wollongong in 2003 and received his PhD in Power Quality Data Analysis Data Mining in 2009.

David Stirling (M' 2001) obtained his BEng from the Tasmanian College of Advanced Education (1976, his MSc (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 is now a Senior Lecturer at the University of Wollongong. His research interests are in Machine Learning and Data Mining.