2009

Discovery and pattern classification of large scale harmonic measurements using data mining

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

Discovery and Pattern Classification of Large Scale Harmonic Measurements using Data Mining

A thesis submitted in fulfilment of the

requirements for the award of the degree

Doctor of Philosophy

from

University of Wollongong

by

Ali Taher M. Asheibi, BSc(Eng), MSc(Eng)

School of Electrical, Computer and Telecommunications Engineering

March 2009

Dedicated to my parents...

Acknowledgements

It is my pleasure to thank the many people to whom I am indebted for the development of this thesis. First and foremost, thanks go to my supervisors, Dr David Stirling, Professor Danny Sutanto and Dr Duane Robinson. Their dedication, knowledge and experience could not have been surpassed.

Thanks to Sean Elphic, Neil Brown and Dr Vic Smith of the Integral Energy Power Quality and Reliability Center who have responded to many technical, administrative and software related requests for assistance.

Tim Brown, Ahsan Lateef, Matthew Field and Praboda Paranavithana, presently and previously with the Integral Energy Power Quality and Reliability Center , have been the sources of many interesting discussions which have contributed to the PhD experience.

Thanks to family who exercised considerable patience even in the face of typical thesis consequences. Particular thanks go to my wife Faesa Netfa who has suffered from many such consequences and been very kind to admit it.

Certification

I, Ali Taher M. Asheibi, declare that this thesis, submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Electrical, Computer and Telecommunications Engineering at the University of Wollongong, is wholly my own work unless otherwise referenced or acknowledged. The document has not been submitted for qualification at any other academic institution.

....

Ali Taher M. Asheibi

12 March 2009

Abstract

Harmonic monitoring is an important issue for electricity utilities and their customers. Continuous monitoring of voltage and current are required to identify any substantial harmonic events before they occur. This monitoring results in large volumes of multivariate data. Although researchers have realised that such large amounts of power quality (PQ) data hold much more information than that reported using classical statistical techniques for PQ monitoring, few have taken the opportunity to exploit this additional information. This hidden information might be of assistance in the identification of critical issues for diagnoses of harmonic problems such as, predicting failures in advance and giving alarms prior to the onset of dangerous situations.

Utility engineers are now seeking new tools in order to extract information that may otherwise remain hidden, especially within large volumes of data. Data mining tools are an obvious candidate for assisting in such analysis of large scale data. Data mining can be understood as a process that uses a variety of analytical tools to identify hidden patterns and relationships within data. Classification based on clustering is an important utilisation of unsupervised learning within data mining, in particular for finding and describing a variety of patterns and anomalies in multivariate data through various machine learning techniques and statistical methods. Clustering is often used to gain an initial insight into complex data and particularly in this case, to identify underlying classes within harmonic data.

The main data mining methodology used in this work is that of mixture modelling based on the Minimum Message Length (MML) algorithm which essentially searches for a model which best describes the data using a metric of an encoded message. This method of unsupervised learning, or clustering, has been shown to be able to detect anomalies and identify useful patterns within the monitored harmonic data set. Anomaly detection and pattern recognition in harmonic data can provide engineers with a rapid, visually oriented method for evaluating the underlying operational information contained within the data set.

A case study from power quality data upon which the MML method has been ap-

plied, was taken from a harmonic monitoring program installed in a typical 33/11kV MV zone substation in Australia that supplies ten 11kV radial feeders. Several patterns have been identified from using the MML technique on the harmonic data, such as significant high harmonic disturbances, footprints of the monitored sites, unusual harmonic events (capacitor switching, turn on televisions, air conditioners and the off peak hot water system) and detection of different abstractions (super-groups), each of which comprise similar clusters. The C5.0 supervised learning algorithm has been used to generate expressible and understandable rules which identify the essential features of each member cluster, and to further utilize these in predicting which ideal clusters any new observed data may best described by.

One difficulty with the MML algorithm when used to derive various mixture models is the difficulty in establishing a suitable stopping criterion to secure the optimum number of (mixture) clusters during the clustering process. A novel technique has been developed to overcome this difficulty using the trend of the exponential of message length difference between consecutive mixture models. First, the proposed method has been tested using data from known number of clusters with randomly generated data points and also with data from a simulation of a power system. The results from these tests confirm the effectiveness of the proposed method in finding the optimum number of clusters. Second, the developed method has been applied to various two-weekly data sets from the harmonic monitoring program used on this thesis. The optimum number of clusters has been verified by the formation of super-groups using Multidimensional Scaling (MDS) and link analysis. Third, the method was benchmarked against a commonly used fitness function technique, which has underestimated the optimal number of cluster in the measured harmonic data. This resulted from the theoretical maximum entropy equation used in calculating the fitness function that assumes the attributes are independent which is not the case in the correlated nature of the harmonic attributes. Finally, generated rules from the C5.0 algorithm were used for classification and prediction of future events to determine which cluster any new data should belong to.

List of Symbols and Abbreviations

ANN	Artificial Neural Network.
Aom	Accuracy of measurement.
DFT	Discrete fourier Transform.
D	Data set.
DWT	Direct Wavelet transform.
CT Fund.	Fundamental current.
CT Harm 3	Third harmonic current.
CT Harm 5	Fifth harmonic current.
CT Harm 7	Seventh harmonic current.
CT Harm 19	Nineteenth harmonic current.
CT Harm 49	Forty-ninth harmonic current.
CT THD	Total harmonic current distortion.
f	Frequency.
FT	Fourier transform.
FFT	Fast Fourier transform.
IEC	International Electrotechnical Commission.
K	Mixture of clusters model.
KDD	Knowledge Discovery in Databases.
KL	Kulback Lieber distance.
LV	Low voltage.
MV	Medium voltage.
MDS	Multidimensional scaling.
MML	Minimum Message Length.
MVAr	Reactive power Q.
PCA	Principle Component Analysis.
PQ	Power Quality.
Ph Fund	Fundamental voltage.
Ph Harm 3	Third harmonic voltage.
Ph Harm 5	Fifth harmonic voltage.
Ph Harm 7	Seventh harmonic voltage.
Ph Harm 19	Nineteenth harmonic voltage.
Ph Harm 49	Forty ninth harmonic voltage.
Ph Total H Dist	Total harmonic voltage distortion.
rms	Root mean square.
SOM	Self Organising Map.
ST	S-transform.
SVM	Support Vector Machine.
WT	Wavelet transform.

Publications arising from this Thesis

1. A. Asheibi, D. Stirling, and D. Sutanto and D. Robinson, *"Clustering, classification and explanatory rules from harmonic monitoring data"*, Book Chapter in *"Theory and Novel Applications of Machine Learning"*, Men Joo Er and Yi Zhou, Eds., I-Tech Education and Publishing, Vienna, Austria, February 2009.

2. A. Asheibi, D. Stirling, and D. Sutanto, *Analyzing Harmonic Monitoring Data using Supervised and Unsupervised Learning.*, IEEE Transactions on Power Delivery, Vol. 24, No.1, pp. 293-301, January 2009.

3. A. Asheibi, D. Stirling, and D. Sutanto, *Classification and Explanatory Rules of Harmonic Data*, Proc. Australasian Universities Power Engineering Conference (AUPEC 2008), 14-17 December 2008 Sydney, Australia, Paper ID: 259.

4. A. Asheibi, D. Stirling, and D. Sutanto, *Determination of the Optimal Number of Clusters in Harmonic Data Classification*, Proc. of the 13$^{\text{th}}$ International Conference on Harmonics and Quality of Power (ICHQP 2008), 28 September-1 October 2008, Wollongong, NSW, Australia, Paper 1045.

5. A. Asheibi, D. Stirling, and D. Sutanto, *Analyzing Harmonic Monitoring Data using Data Mining*, Proc. Fifth Australasian Data Mining Conference(AusDM06), 29-30 November, 2006, Sydney, NSW, Australia, pp: 63-68.

6. A. Asheibi, D. Stirling, and D. Sutanto, *Analyzing Harmonic Monitoring Data using Data Mining*, Conferences in Research and Practice in Information Technology (CRPIT), 61. Peter, C., Kennedy, P.J., Li, J., Simoff, S.J. and Williams, G.J., Eds., Australian Computer Society Inc. (ACS), 2006, pp: 63-68.

7. A. Asheibi, D. Stirling, and D. Robinson, *Identification of Load Power Quality Characteristics using Data Mining. Proc. of the Canadian Conference on Electrical and Computer Engineering, 2006. CCECE '06.*, 7-10 May 2006, Ottawa, Canada, pp: 157-162.

8. A. Asheibi, D. Stirling, S. Perera and D. Robinson, *Power quality data analysis using unsupervised data mining*, Proc. Australasian Universities Power Engineering Conference (AUPEC 2004), 26-29 September 2004, Brisbane, Australia, Paper ID: 187.

Table of Contents

List of Figures

List of Tables