TWO STAGES CLUSTERING USED IN POWER QUALITY SURVEYS TO SELECT LOCAL FROM NETWORK SIGNATURES - FREQUENCY CASE STUDY

Dan APETREI[1]

Lucrarea propune o metodă nouă de extragere a evenimentelor din seriile de timp ce conțin date rezultate la supravegherea calității energiei electrice. Folosind tehnici de grupare ierarhică, evenimentele extrase sunt analizate în vederea extragerii de semnături. Analizele se sprijină pe rezultate practice obținute în campaniile de determinări efectuate în ultimii patru ani în trei amplasamente din țară.

The paper presents a method to extract events from the time series obtained in power quality surveys. By using hierarchical clustering techniques, the events extracted are analyzed in order to define signatures. The analysis is based on a real life survey made in the last four years in three nodes of the Romanian network.

Keywords: frequency, clustering, signature

1. Introduction

Using data mining techniques in power quality investigation is somehow recent domain of investigation. One of the key drivers in developing this techniques is the rapid development of computer technology [12,13].

This type of investigation is meant to extract information from data and to give decision support. Such information could be used for identification and diagnoses of power quality disturbance problems [14]. Other area of application is the prediction of system abnormalities or failure, and alarming of critical system situations [11].

Most of the investigation done by now was dedicated to local problem identification and data classification. This paper approach is different in terms of using high amount of real life data and defining based on these data practical methods for power quality investigation on multiple network nodes. As far as we know, clustering techniques were used to build classes within power quality data warehouses [15].

A new approach is proposed for using clustering residuals as feasible events filtering technique. Another original contribution is the second stage

[1] PhD student, SC Electrica SA, Bucharest, Romania, e-mail: dan.apetrei@electrica.ro

clustering that allows event classification and signature extraction. That is a proof that we are not in the noise domain, but the data extracted even from the normalized series reflect real life events.

The paper investigates the frequency as characteristic of the voltage waveform is regulated by quality standards [2,3]. These standards combine threshold prescription and measurement methods description.

The measurement method standard [3] defines the measurement process and the way results should be interpreted. Measurement methods are described for each relevant type of parameter in terms that will make it possible to obtain reliable, repeatable and comparable results. The requirements are set up by defining two classes of performance for voltage measurement, namely, class A and class S. This paper deals with data determined using class A equipment.

This paper shows a way to increase the sensibility of the process of selecting events beyond the limits of the classical filters described in the standard. Besides this on the event pool a selection is made in order to identify repeated event signatures. Considerations on local/network presence of the signature are made.

2. Brief description of experimental data context

The measurement system that we used stores every half cycle duration value for each measurement location during the survey. Fig. 1 gives a brief description of the idea behind the measurement experiment and the way results are expected from it.

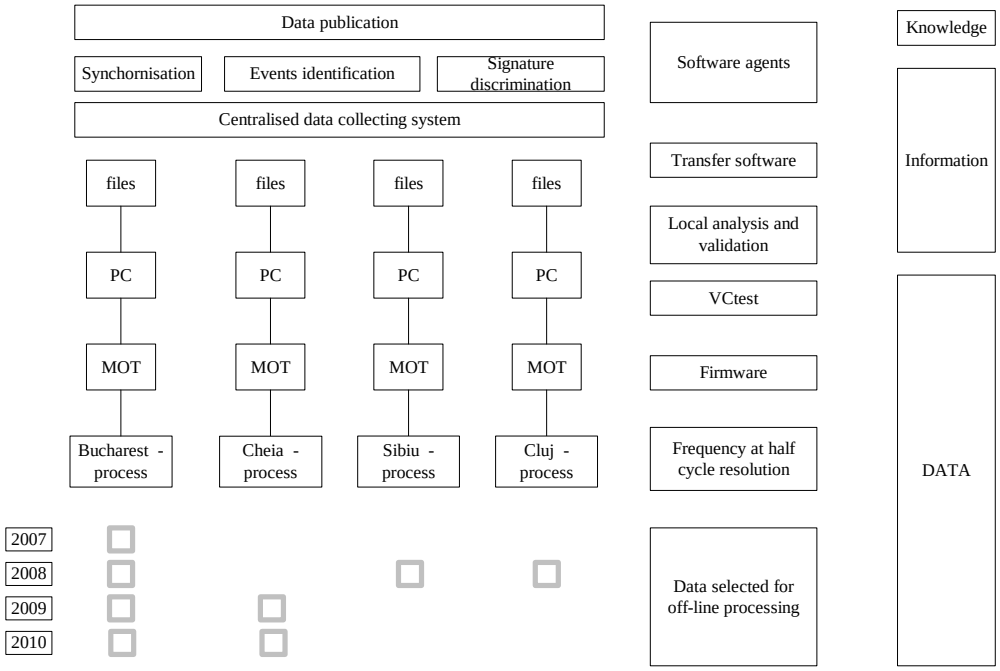


Fig. 1. General description of the measurement environment

As could be seen in the lower part of Fig. 1, the locations we are talking about are: Bucharest, Cheia, Sibiu and Cluj. First test were in 2006 and the first coordinated measurement campaign was in 2007. Since then, the process of acquiring data from the field was continued.

The measurement campaign took almost four years and covered four measurement locations. Some of the spots are still active.

3. Short considerations on Clustering

The Clustering Method [4] includes many alternatives of classification. Among them one can identify: between-groups linkage, within-groups linkage, nearest neighbour, farthest neighbour, centric clustering, median clustering, Ward's method.

In order to apply one of the above clustering methods, one important selection criteria is the definition of a distance or similarity measure. A common choice for the available software tools is computing: interval data, Euclidean distance, squared Euclidean distance, cosine, Pearson correlation, Chebychev, block, Minkowski, and customized definition of distance.

Processing steps for hierarchical clustering are:

- Considering a number of points ***N*** in a ***m*** dimensional space we form ***N*** initial groups called clusters;
- Iteration after iteration the number of groups is reduced based on distance selection criteria;
- The process is repeated until the goal number of clusters is reached or we get only one common cluster.

Grouping is based on a distance definition and selection [5]. In order to take advantage of this algorithm, a square matrix is built having one line and one column for every group. The matrix cell contains the distance between the groups according to the metric that was selected. The similarity is calculated according to distances based on a criterion a priori selected.

4. First stage clustering - Events extraction

From the data collected, we selected 191 ten minute intervals starting on Monday 14th of January 2008 00:00. Having data for three out of four locations (Bucharest, Cluj and Sibiu) was one of the selection criteria.

The purpose of the analysis is to select the best clustering method that extracts sag events or sag events traces.

For each interval, the 60000 values measured during 10 minutes survey were split into 60 vectors of 1000 values. Each vector describes the measurement for 10 seconds. After all, this is the time interval recommended by the standard [3].

We applied different scenarios in order to determine the best choice for the distance definition. The following methods were used [6,7]:

• Euclidean distance: is the most common distance measure. A given pair of cases is plotted on two variables, which form the x and y axes. The Euclidean distance is the square root of the sum of the square of the x difference plus the square of the y distance.

• Chebychev distance: is the maximum absolute difference between a pair of cases on any one of the two or more dimensions (variables) which are being used to define distance.

• Minkowski distance: is the generalized distance function (1) representing the p-th root of the sum of the absolute differences to the p-th power between the values for the items:

$$d_{ij}^{(p)} = \left(\sum_{k=1}^{K} \left(x_{ik} - x_{jk} \right)^p \right)^{1/p} . \qquad (1)$$

When $p = 2$, the Minkowski distance is the same as the Euclidean distance. In order to determine the best distance definition used for clustering in order to separate the sags, we built the dendrograms according to the distance definition as could be seen in Fig. 2.

The dendrograms show the relative size of the proximity coefficients at which cases were combined.

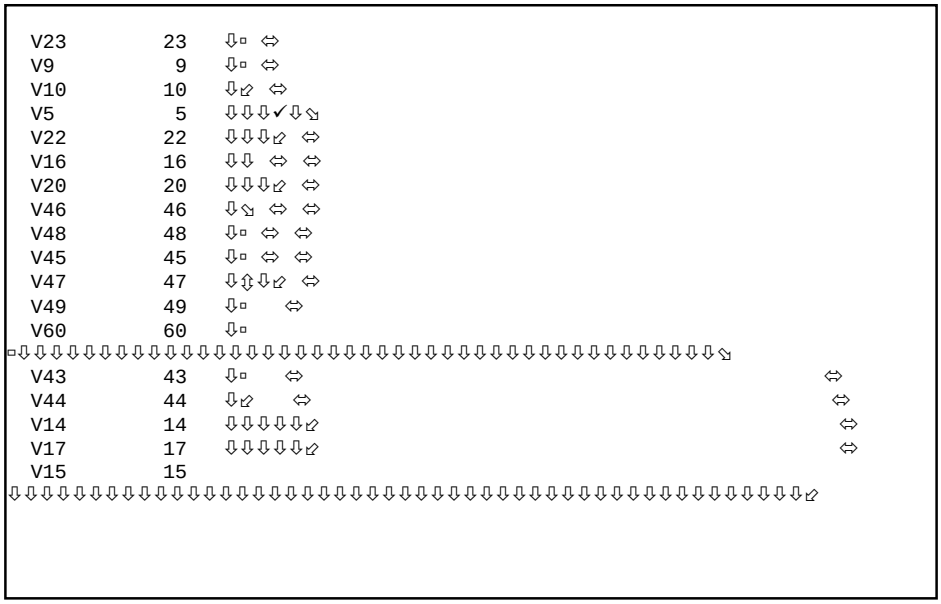


Fig. 2. Dendrogram of the interval 118 registered in Cluj

Cases with low distance/high similarity are close together. Cases showing low distance are close to each other. Similar cases are connected with a line linking them at a short distance from the upper part of the dendrogram. This indicates that they are agglomerated into a cluster at a low distance coefficient, indicating alikeness. The 15th vector in Fig. 2 is obviously different with respect to the others, since it participates in the grouping process at a latter stage. Visual confirmation of the fact that we managed to select sag or a sag trace is presented in Fig. 3.

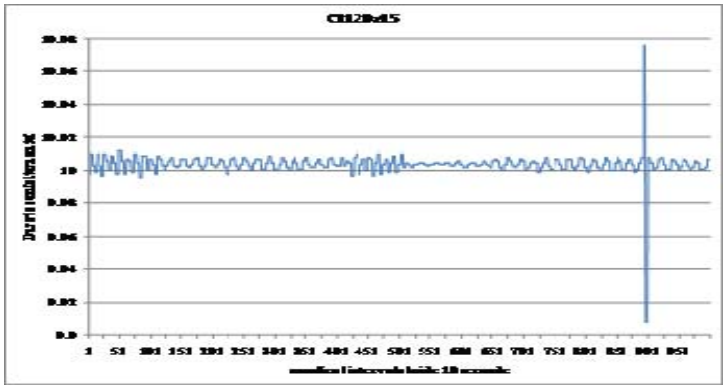


Fig. 3. The 15th 10 seconds interval of the 118th 10 minutes determination in Cluj

For the dendrograms in Fig. 4, vectors number 3, 10 and 46 are marked with a blue line. These are the events that need to be extracted since they contain sags [6]. From the dendrogram structure, it appears that the most effective of the three methods to identify sag is clustering with Chebychev definition of the metric.

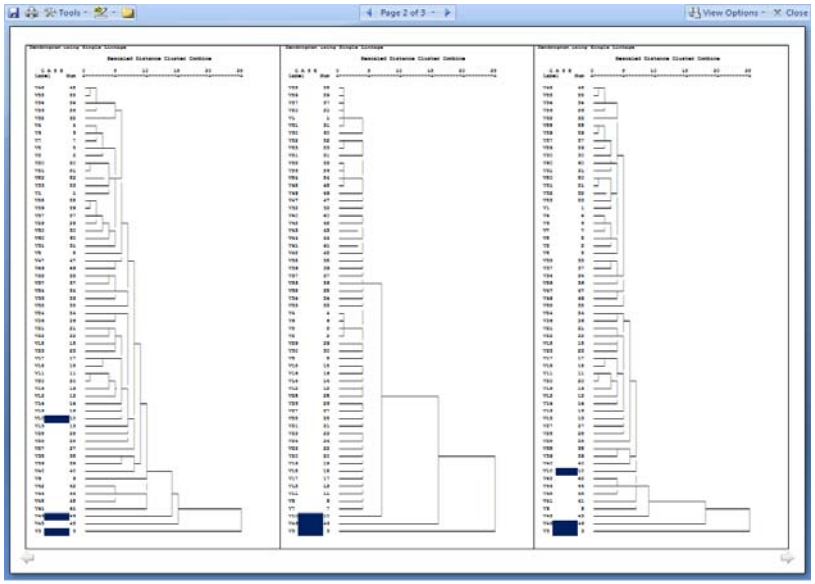


Fig. 4. Dendrograms with different metrics. From left to right distances were: Euclidian, Chebychev, Minkowski (p=3)

Because of its popularity in power quality surveys, the method used to form clusters was the “nearest neighbour”. In short, this means that the distance between two clusters is the distance between their closest neighbouring points.

5. Second stage clustering, events grouping and signature extraction

From each vector of the extended pool of 42 stepping events identified in the previous paragraph, we extracted the relevant 32 records. That means 100ms before the step and 200ms after it. Fig. 5 presents the simplified form of the dendrogram that resulted in the clustering process of these new vectors. The 42 events are distributed as follows: Bucharest 18; Sibiu 17; Cluj 7. Since signature identification is a shape matching process we normalized the vectors before clustering.

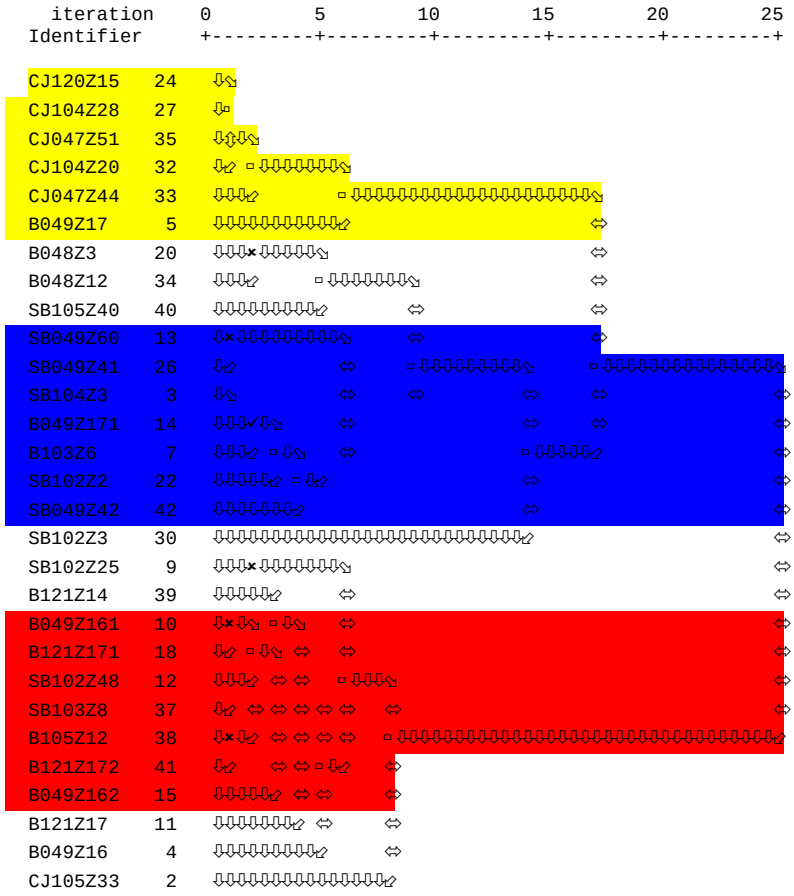


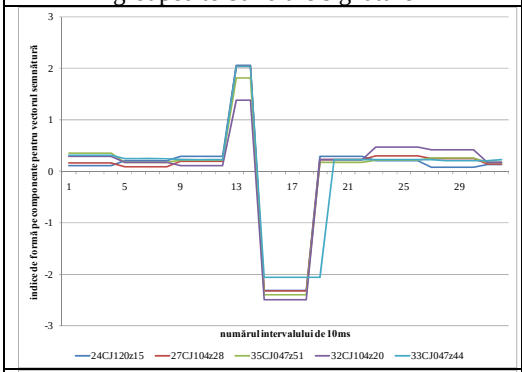
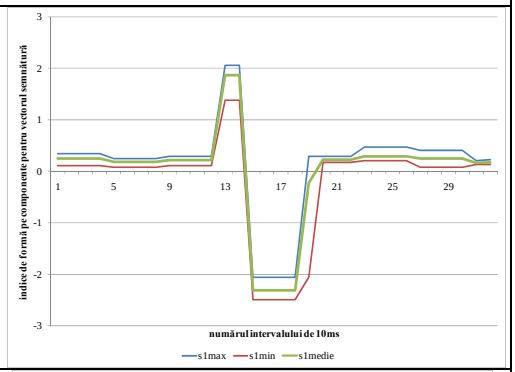
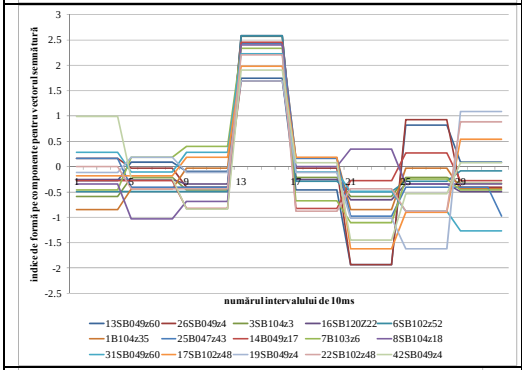
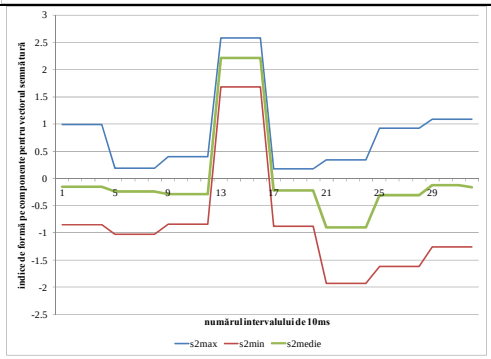
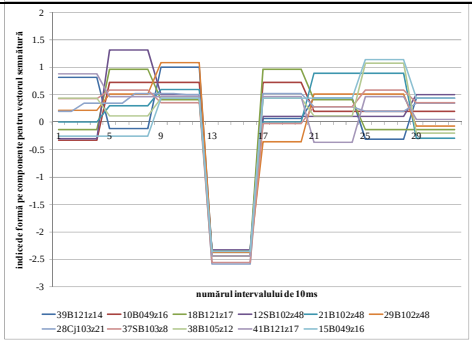
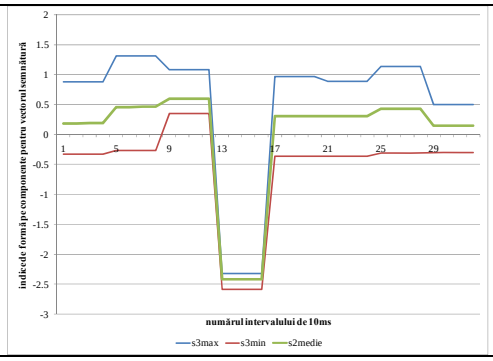
Fig. 5. Signature identification dendrogram

In Table 1 we present the time series of the events grouped using the dendrogram. The first column shows the normalised curves and the second column shows the signature and the limits of the signature signal.

As it can be seen in Fig. 5, there are three distinct areas in the dendrogram. Each area of the dendrogram determines one line in the table. We marked the dendrogram boxes with yellow, blue and red colours.

Table 1

Signatures identified in the analysis of half cycle duration

Graphical representation of the data that were grouped to build the signature	Signature, maximum and minimum
	
	
	

The first box, the yellow one, contains only events that took place in Cluj. That could lead to idea that the signature associated to these events is a *local* one, related to a particular process of balancing production and consumption.

The events in the blue box are recorded in Bucharest and Sibiu. There is a possibility (that can be further investigated) that these events are specific to the Southern area of the Romanian network.

The red box of the dendrogram is grouping events that appear in all three measurement locations. Further investigation could show whether those signatures can describe *system-wide* events.

In the end, we got one signature that appears only in one measurement location (first line of Table 1), one signature that appears only in a specific area of the network (second line of Table 1) and a signature that is common to all measurement points (last line of Table 1) [10].

7. Conclusion

The static definition of the threshold in detecting events does not correlate the node parameters to the power quality analysis criteria [9]. For instance, if the measurement is made at high voltage close to a power generator, there is a lower probability to detect sags than for a rural area low voltage determination.

In order to avoid the drawbacks of the static definition of sags coming from the standards [2,3] we developed a clustering based method to extract events from time series of power quality parameter survey.

After testing the selectivity of different definitions of the distance, we selected the Chebychev distance as the best choice for event and event trace detection. Based on the events detect, a second clustering stage allowed the detection of the signatures of the events.

Acknowledgments

The idea of this paper came from the results of the research on hierarchical clustering applied in load profiles analysis by Professor Gianfranco Chicco. He helped me to develop the idea and to apply the clustering concept in Power quality research.

Most of the analysis that is presented in the paper was discussed with Professor Mihaela Albu. Her expertise in Power quality measurements allowed me to clarify the concepts and put order in the development of the presentation. Besides this, her kindness in sharing a vast area of contacts allowed me to access some of the experts in this domain.

Besides opening research gates and encouraging me to extend the investigation area in “marginal” zones of the Power quality domain Professor

Petru Postolache, was an invaluable information resource for a broader knowledge area then Power quality domain. He taught me the pleasure of doing the undone…

Professor Nicolae Golovanov is my tutor in the struggle for one step further in power quality domain. Simple “thank you” is not enough for all the help I got from him in the difficult moments of the research and the good advices that allowed obtaining the results presented in the paper. He had the art to make me feel that the research could be attractive despite the fact it needs to be disciplined.

R E F E R E N C E S

[1] „ - "CEI 60050-161:1990, International Electrotechnical Vocabulary (IEV) - cap. 161: Electromagnetic compatibility", CEI, 0

[2] „ - "EN 50160 Voltage characteristics of electricity supplied by public distribution systems", CENELEC, 1999

[3] „ - "IEC 61000-4-30 Ed. 1: Electromagnetic compatibility (EMC) - Part 4-30: Testing andmeasurement techniques - Power quality measurement methods", CEI, 2003

[4] *Guojun Gan, Chaoqun Ma Jianhong Wu*, "Data Clustering: Theory, Algorithms, and Application", - ASA-SIAM Series on Statistics and Applied Probability, 2007

[5] *M.R. Anderberg*, "Cluster Analysis for Applications", - Academic Press, New York, 1973

[6] *D. Apetrei, P. Postolache, N. Golovanov*, Mihaela Albu, Gianfranco Chicco, - "Hierarchical Cluster Classification of Half Cycle Measurements in Low Voltage Distribution Networks for Events Discrimination ", - ICREPQ, 2009

[7] *D. Apetrei, G. Chicco, P. Postolache, N Golovanov, M. Albu*, "Cluster Analysis of Half-Cycle DurationMeasurements to Classify Local and Network events", IEEE Powertech , 2009

[8] D*. Apetrei, G. Chicco, R. Neurohr, M. Albu, I. Silvas, P. Postolache*, "Voltage and Frequency time series analysisImprovements in transforming data from quality survey into information", - EEEIC - Prague Czech Republic - IEEE, 2010

[9] D*. Apetrei, P. Postolache, I. Silvas*, "Power Quality Time Series Analysis - Improved Means to Extract Events", - MPS - Modern Power Systems - Cluj, 2010

[10] *D. Apetrei, I. Silvas, V. Rascanu, C. Stanescu, D. Stanescu, A. Nicoara*, "Aspects Of Frequency As System Parameter - Low Voltage Measurement Case Study", Conference On Electricity Distribution Serbia, Vrnjacka Banja, CIRED, 2008

[11] *A. Asheibi, D. Stirlingy, D. Robinsonz*, "Identification of Load Power QualityCharacteristics using Data Mining", IEEE CCECE/CCGEI, Ottawa, 2006

[12] *P.K. Dash, I.L.W. Chun, M.V. Chilukuri*, "Power quality data mining using soft computing and wavelet transform", - TENCON 2003. Conference on Convergent Technologies for Asia-Pacific Region -ISBN: 0-7803-8162-9, 2003

[13] *Jun Meng, Dan Sun, Zhiyong Li*, "Applications of Data Mining Time Series to Power Systems Disturbance Analysis", - Advanced Data Mining and Applications Lecture Notes in Computer Science, Springer Link, 2006

[14] *K. Vivek, M. Gopa, B.K. Panigrahi*, "Knowledge Discovery in Power Quality Data Using Support Vector Machine and S-Transform", Third International Conference on Information Technology: New Generations (ITNG'06) - Las Vegas, Nevada, 2006

[15] *Le Xu, Mo-Yuen Chow, Leroy S. Taylor*, "Power Distribution Fault Cause Identification WithImbalanced Data Using the Data Mining-BasedFuzzy Classification E-Algorithm", IEEE TRANSACTIONS ON POWER SYSTEMS, VOL. 22, NO. 1, FEBRUARY, 2007.