A Voltage Sag Pattern Classification Technique

Délio E.B. Fernandes, Mário Fabiano Alves, and Pyramo Pires da Costa Jr.

Pontifical Catholic University of Minas Gerais, PUC-MG,
Av. Dom José Gaspar, 500, Prédio 3, 30535-610, Belo Horizonte, MG, Brazil
{delio, mfabiano, pyramo}@pucminas.br

Abstract. This paper presents an investigation on pattern classification techniques applied to voltage sag monitoring data. Similar pattern groups or sets of classes, resulting from a voltage sag classification, represent disturbance categories that may be used as indexes for a cause/effect disturbance analysis. Various classification algorithms are compared in order to establish a classifier design. Results over clustering performance indexes are presented for hierarchical, fuzzy c-means and k-means unsupervised clustering techniques, and a principal component analysis is used for features (or attributes) choice. The efficiency of the algorithms was analyzed by applying the CDI and DBI indexes.

1 Introduction

As a result of customers' expectations on services level provided by power systems, utilities have begun to develop and apply extensive power quality (PQ) monitoring programs [1]. Voltage sags are reductions in the root mean square value of a voltage, and a great part of the related power quality problems faced by utilities and indus-tries are associated to this PQ disturbance [1]. Modern power quality measurement equipments used to monitor these disturbances can generate a large amount of data. A monitoring system, composed by anything from a few PQ monitors to a hundred or more units, covering several points of an electrical network, generates gigabytes of data. This great amount of data must be systematically and automatically organized in a power quality database, using structures of information developed for characterization of each particular PQ phenomenon, such as voltage sags. A computational system with these characteristics has been developed by the authors [2]. The information introduced in this database comes from a characterization process that defines attributes to the voltage sag event, in an automated pre-processing task, from a signal processing assessment.

Generally, voltage sag cause or effects on a power system plant is analyzed by comparing several event patterns on a bi-dimensional basis. Traditional analysis uses a pair of attributes, magnitude and duration of the voltage sag, to classify it. Several other attributes can be derived from a time analysis of a voltage sag digital recorder, with the possibility of further improving the classification process. Some of these attributes are: the point of the wave where the event begins, the deviation of the phase angle, the average value, hour of the day of event occurrence, energy loss, type of voltage sag (three phase, phase to phase, phase to ground) etc. The use of a larger number of attributes, in a multivariate analysis, increases the discriminatory capacity among the events. Real data sets (data collection) retrieved from the mentioned power quality database, are used as

S.K. Pal et al. (Eds.): PReMI 2005, LNCS 3776, pp. 248–253, 2005.

input classification data, and are aggregated by monitoring site, time period, monitor channel or magnitude.

Patterns classification consists on allocation of events to clusters or classes with similar characteristics. A classifier design cycle includes: data collection, features choice, model choice, classifier training and a classifier evaluation [3].

There are extensive investigations on pattern classification using both statistical and artificial intelligence techniques. A complete investigation review linking artificial intelligence and other mathematical tools to PQ is detailed in [4]. Classical statistics classifying and clustering formulations are presented in [3,5,6,7]. Artificial intelligence technique, specifically using neural networks and fuzzy logic on pattern classification are described in [3,5,8].

In order to obtain pattern classes we use different techniques of unsupervised learning data clustering: hierarchical clustering, k-means and fuzzy c-means. In the unsupervised learn classification tecniques, patterns form clusters in a ”natural grouping”, with results depending on the specific clustering system and data set used. Thus, for a given specific classification problem some systems are more appropriate than others. Classification procedure is followed by its test under different performance indexes [5]. The indexes applied to a number of data clusters that better fits a natural partitioning of the original data set space define which classifier model to be used.

The data set under analysis must be tested to define how useful each feature is to achieve discrimination. With the total variance as a measure, a principal component analysis - PCA applied over the data set, is then used to reduce the dimensionality of the problem [6]. After that, a new clustering session, with a lower dimensionality, defines feature choice assessment.

This work presents the results of the investigations performed on unsupervised classification techniques for data clustering. The data set used originated from digital recording of voltage sags on a specific bus from a regional transmission system (138 kV and higher) and it is composed of eighty patterns. This large power system can be divided into several regional systems, with most of the buses have a similar behavior regarding voltage sags. The authors are presently working with supervised classifiers to complete the design cycle mentioned previously, so as to reproduce the target knowledge, and test the classifier for other regional systems, with different voltage sag characteristics. The results will be presented in a future publication.

2 Classification Techniques

The following terms and notation are used trough this paper. A feature vector $x = (x_1, x_2, ..., x_H)$ is a pattern composed by H attributes (dimensionality). A pattern set is denoted $x = (X_1, X_2, ..., X_N)$, in a $N \times H$ pattern matrix, corresponding to N voltage sag events.

There are several metrics used to quantify the similarity of patterns on a feature space, each one resulting in different cluster extraction characteristics. Euclidian norm is used as a common metric for all the classifiers. Some Euclidean based distances measures that define clustering criterion functions are:

Vector to vector defines a distance between two patterns.

$$d(x,y) = \sqrt{\sum_{h=1}^{H} (x_h - y_h)^2} \tag{1}$$

Mean vector for cluster D_i, or cluster i centroid, where n_i is the cluster i number of patterns.

$$m_i = \frac{1}{n_i} \sum_{x \in D_i} (x) \tag{2}$$

Vector to set gives a mean distance between a sample and a cluster.

$$d(x,X) = \sqrt{\frac{1}{n_i} \sum_{y \in X} d^2(x,y)} \tag{3}$$

Intraset distance gives the mean distance between patterns inside of each cluster

$$\hat{d}(X) = \sqrt{\frac{1}{n_i} \sum_{j=1}^{n_i} d^2(x_j,X)} \tag{4}$$

In hierarchical clustering, considering a data set with N samples partitioned into C clusters, successively merged clusters based on a similarity measure, will be called agglomerative hierarchical clustering technique. A similarity level matrix provides a mean to group data into clusters, in a procedure with space complexity $O(CN2d)$, where d is the searching number needed to find the min/max distance. The most common hierarchical clustering algorithms are single linkage, complete linkage and average linkage. Single linkage performs the clustering using the minimum distance criterion, complete linkage uses maximum distance criterion. Single link is a more versatile algorithm, however complete link produces more useful hierarchies in many applications [7]. In the average linkage method, the distance between clusters is the average distance between all pairs of patterns, where one member of a pair belongs to each cluster. Depending on the clustering algorithm and metric used, a new result is reached. This iterative process defines a partition extremized by any criterion function.

The k-means clustering is a partitional algorithm that employs the squared error criterion, in relative low time, with space complexity $O(NHCt)$, where t is the number of iterations. Clusters centers are initially randomly chosen, the result de-pending on the choice, and each pattern assigned to its closest cluster center. The clusters centroids are recomputed and a convergence criterion tested to continue or not the process [7].

In the fuzzy c-means clustering, each pattern has a fuzzy membership grade, from a matrix u, where an element u_ij represents the probability of a pattern x_j to be classified as belonging to cluster c_i. The fuzzy c-means algorithm seeks a minimum of a global objective function. Patterns are reassigned to clusters in order to reduce this criterion function and recompute u, in an iterative process. The process stops when less significant changes in u is reached.

3 Clustering Validity Assessment

In order to examine the clustering techniques used for voltage sag patterns classification, intensive classifying sessions, with different criterion indexes and a progressive increase in the number of clusters, are used. Cluster dispersion indicator - CDI [5], based on a squared mean of K clusters centers distances set C:

$$CDI = \frac{1}{\hat{d}(C)}\sqrt{\frac{1}{K}\sum_{k=1}^{K}\hat{d}^2(X_k)} \tag{5}$$

Davies-Bouldin Index - DBI [5], represents the system-wide average similarity measure of each cluster with its most similar cluster.

$$DBI = \frac{1}{K}\sum_{k=1}^{K}\max_{i\neq j}\{\frac{\hat{d}(X^i)+\hat{d}(X^j)}{d(C^i,C^j)}\} \tag{6}$$

In these indexes, a smaller value represents a better clustering quality. When applying the indexes to k-means and fuzzy k-means algorithms, care must be taken regarding to the dependency on starting conditions. Random sorted centers can lead to different solutions.

In Figures 1 and 2 the results for both CDI and DBI indexes indicate the superiority of the hierarchical algorithms in grouping voltage sag patterns for the data set under

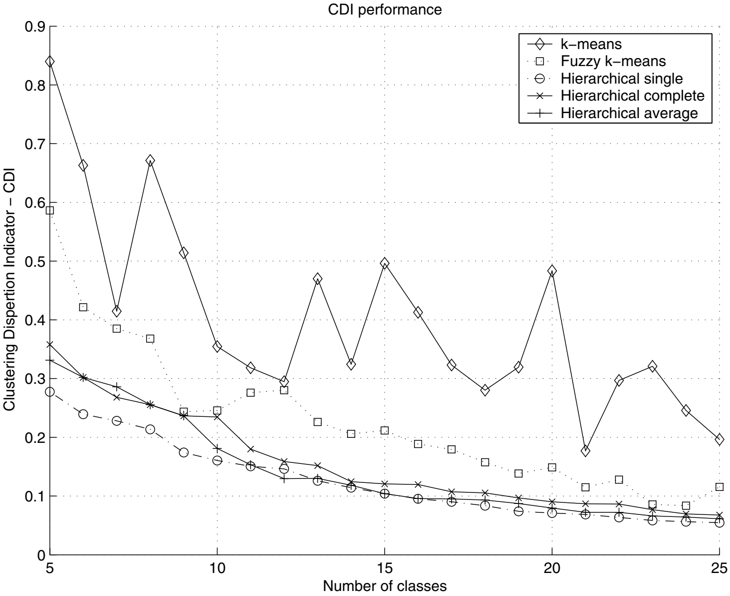

Fig. 1. Clustering dispersion Index - CDI computed from a 5 to 25 clustering using several clustering techniques

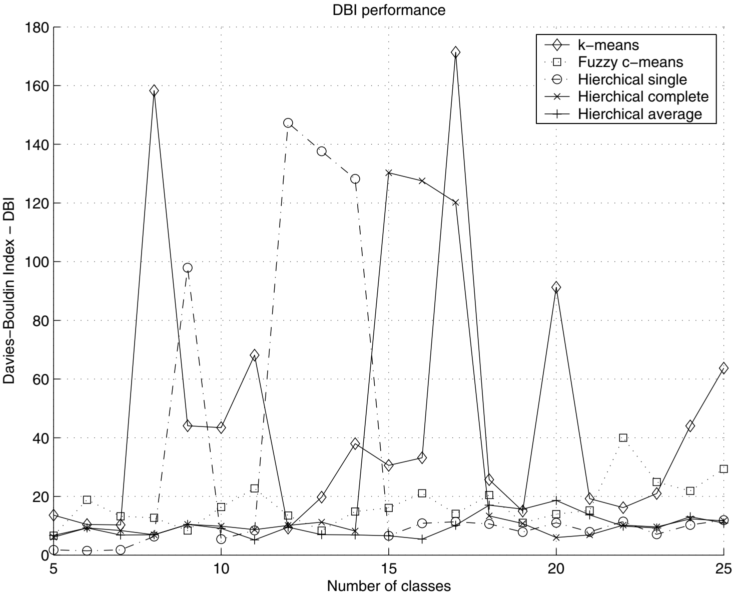

Fig. 2. Davies-Bouldin Index - DBI computed from a 5 to 25 clustering using several clustering techniques

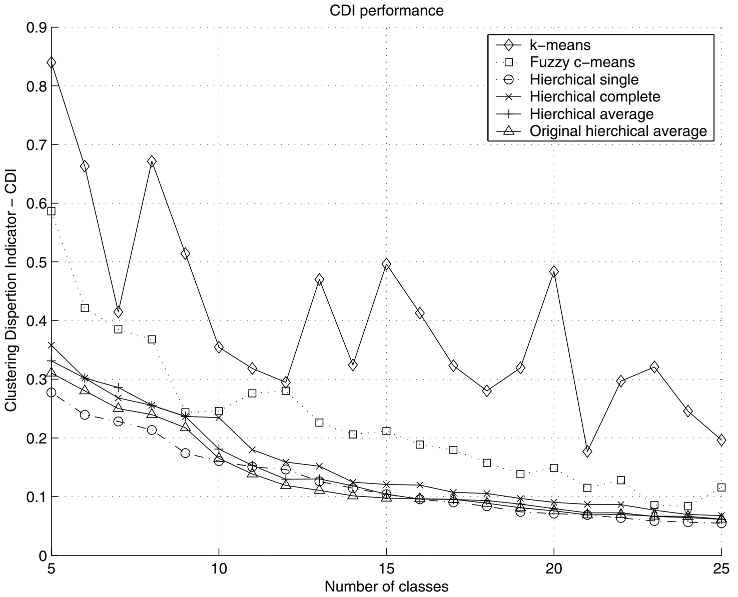

Fig. 3. Clustering dispersion Index - CDI applied over the five first principal component subspace of original data set. Original hierarchical average linkage plotted as a comparative parameter.

analysis. It can also be concluded that use of the CDI index resulted in a better tendency of cluster definition.

The use of principal component analysis, or $Karhunen - Lovetransform$, leads to a feature choice selection by projecting H dimensional data onto a lower R dimensional subspace (the inherent dimensionality). PCA eigenvalues show that the amount of total variance of the used data set is explained by using the first R components. PCA analysis reduced the dimensionality of the problem from eleven variables to five first principal components. Clustering validity over the new R dimensional data set was then performed using CDI and DBI index. Hierarchical single clustering obtained with the original data set, was used for comparison. Figure 3 shows the results of this new clustering process computing the CDI performance index.

4 Conclusions

Voltage sag pattern classification has been investigated using five unsupervised algorithms. The efficiency of the algorithms was analyzed by applying the CDI and DBI indexes. The hierarchical algorithms presented the best performance. This conclusion holds for the two situations considered, that is, for the case of the original data set and for the case with the dimensionality reduced by PCA analysis.

References

1. Dugan, R.C. McGranaghan, M.F. Beaty, H.W.: *Eletrical Power Systems Quality*, McGraw-Hill, (1996).
2. Alves, M.F., Fernandes, D.E.: *Development of an Automated Power Quality Management System*, 1999 IEEE/PES Transmission and Distribution Conference, New Orleans, (1999).
3. Duda, R.O., Hart, P.E., Stork, D.G.: *Pattern Classification*, John Willey & Sons, New York, USA, (2001), 654p.
4. Morcos, M. M., Ibrahim, W. R. A.,: *Artificial Intelligence Advanced Mathematical Tools for Power Quality Applications*, IEEE Transactions on Power Delivery, v17, n2, june, (2002).
5. Chicco,G., Naopli,R., Piglioni,F.: *Comparisons among Clustering Techniques for Electricity Customer Classification*, IEEE Bologna Power Tech, Bologna, Italy, june, (2003).
6. Johnson, R.A., Wichern, D.W.: *Applied Multivariate Statistical Analysis*, Prentice Hall, 3rd edition, New Jersey, USA, (1992), 642p.
7. Jain, A.K. , Murty, M.N., Flynn, P.J.: *Data Clustering: A Review*, ACM Computer Survey, 31, 3, june, (1999), 264-323.
8. Haykin, S.: *Neural Networks - A Comprehensive Foundation*, Macmillan Publishing Company, USA, (1994), 696p.