DATA MINING TECHNIQUES APPLIED TO SPATIAL LOAD FORECASTING

F. H. Pereyra Zamora C. M. V. Tahan

Escola Politécnica da Universidade de São Paulo – Departamento de Engenharia de Energia e Automação Elétricas
Av. Prof. Luciano Gualberto, Trav. 3, 158, Cidade Universitária – São Paulo - Brasil
franzhpz@pea.usp.br – cmvtahan@pea.usp.br

INTRODUCTION

Load forecasting in distribution systems is essentially important and it is a challenge due to its spatial diversity associated to the land-use and consumers that inhabit on it. Many factors should be considered, such as: which type of land-use exists and will exist and which type of energy consumption the territory in analysis has or will have; when new substations and new feeders should be built or when existent facilities should be reinforced; where to plan new lines and structures.

This work presents the Data Mining techniques applied to a methodology for Spatial Load Forecasting which is aimed to the planning of distribution systems. This methodology addresses aspects such as: consumers’ data, network load data, end-use load curve standards, consumers load curves aggregation, small areas, classification of electric loads and patterns recognition.

For the classification of the electric loads in the small areas, it is utilized the Kohonen Self-Organizing Map, which allows to obtain, the automation of the electric loads classification process in each small area defined for the base year.

The patterns recognition is accomplished with clusters analysis methods through the application of Kohonen Self-Organizing Map to obtain the patterns of the small areas.

The growth of the global market is determined by the Back Propagation Neural Network. The Back Propagation Neural Network is used to accomplish the global load forecasting, regarding the historical series values for the different type of consumers (i.e. residential, commercial, industrial, rural and others).

The statistical methods of fit curves, regression and time-series are also used with the purpose of defining alternatives to make the right decisions, with a smaller risk.

The allocation process of the future load in the geographical area depends on the growth of the determined global market, considering different scenarios and the load evolution per small area, as well as the land-use evolution according to the urban planning and the data base integration of the power distribution companies.

The future load allocation in the geographical space, can be accomplished after there are obtained the predominant types of land-use, the small area patterns, the respective rate of growth and global forecasting.

Artificial Neural Networks are probably the most common data mining techniques. Neural Networks learn from a training set, generalizing patterns inside them for classification and prediction.

KDD AND DATA MINING

The expressive growth of database has created a need for a new generation of tools and techniques for automated and intelligent database analysis. In this context, the process of Knowledge Discovery in Databases (KDD) and Data Mining help the database analysis [1].

Data Mining is the exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules [2].

The KDD process is interactive and iterative, involving the following steps: selection, preprocessing, transformation, data mining and interpretation/evaluation.

First it is necessary to make a selection of the data. For this, one shall understand the domain of the application and the objectives of the final user. The data set target must be looked for, and then it must be select the subset of variables or attributes on which will take place the knowledge discovery.

The preprocessing supposes tasks, such as: noise elimination, strategies for managing null values or missing values and normalization of the data.

In the following step, the transformation and reduction of the data are accomplished. This is a critical stage of the process of KDD that requests a good knowledge of the problem. Data Mining can make a limited set of tasks and only under limited circumstances. Classification, estimation, prediction, affinity grouping, clustering and description are the tasks accomplished in data mining.
After the discovery of patterns, the results are monitored and it is formulated a way to make use of the extracted knowledge. Only after the interpretation of the obtained information, the knowledge is founded.

The choice of a particular combination of techniques to apply in a particular situation depends on both the nature of data mining task to be accomplished and the nature of the available data. Familiarity with a variety of techniques is necessary to provide the best approach to solve data mining problems.

The main tasks of data mining in the methodology of spatial load forecasting are: classification, cluster analysis, prediction and time-series [1 - 3].

Classification analyzes a set of training data and constructs a model for each class based on the features in the data. There have been many classification methods developed in the fields of machine learning, statistics, database, neural network and others.

Prediction involves the finding of the set of attributes relevant to the attribute of interest and predicting the value distribution based on the set the data similar to the select objects. Usually, regression analysis, generalized linear model, correlation analysis and decision trees are useful tools in quality prediction. Genetic algorithms and neural networks models are also popularly used in prediction.

Clustering analysis is to identify clusters embedded in the data, where a cluster is a collection of data objects that are similar to one another. Similarity can be expressed by distance functions. A good clustering method produces high quality clusters to ensure that the inter-cluster similarity is low and the intra-cluster similarity is high.

Time-series analysis is to analyze large set of time-series data to find certain regularities and interesting characteristics, including search for similar sequences or subsequences, and mining sequential patterns, periodicities, trends and deviations.

**LOCAL DATABASE**

The database used in the application of spatial load forecasting should allow the storage of consolidated information in relation to the data target, the pre-processed data, the transformed data, the found patterns and the obtained knowledge, as it happens in the evolution of KDD process.

In the development phase of a database there are three types of data models that must be considered: the conceptual data model, the logical data model and the physical data model [4].

The objects, their characteristics and relationships that faithfully represent the observed environment define the conceptual data model. Independently of limitations imposed by the technologies, implementation techniques or physical devices, this model simply considers the conceptual aspect of a certain environment to represent the concepts and characteristics.

In the **conceptual data model** the particularities of implementation of the system and the way of future implementation are unknown. Thus, the conceptual data model remains unalterable in relation to an implementation at a relational or hierarchical database management system.

The **logical data model** is the one in which the objects, their characteristics and relationships are represented according to the implementation rules and limitations due to a type of technology. But, this representation is independent of the devices or means of physical storage of the data structures.

The concepts of access keys, control of duplicated keys, normalization, etc. are respected for their posterior implementation in information systems. The logical model results from the application of derivation rules upon a conceptual model of data already developed.

The representation of the objects under the aspect of the physical level of implementation of the occurrences, or instances of the entities and their relationships, constitute the **physical data model**.

Two levels of representation can be included in the physical data model. Thus, the occurrences or instances, their relationships and physical disposition of elements form a first level of representation, and, the allocation of physical spaces in the several grouping levels (tables, blocks, registrations, fields) form a second level of representation.
Figure 2 [7] presents the general diagram about how the information flows during the several processes necessary for the solution of the problem of spatial electric load forecasting.

**GENERAL METHODOLOGY FOR SPATIAL LOAD FORECASTING**

In order to achieve the aims supposed to do load spatial forecast, with the help of GIS, it is necessary to develop the activities grouped into 4 phases. Such phases require methodologies and techniques to solve the spatial forecast problem, in an integrated way [5 - 7]. Figure 3 illustrates the phases in the development of the methodology, which are briefly described herein:

**Phase 1** includes activities related to information gathering, including: 1) Consumers’ Data, 2) Patterns of Load Curves of Final Use, 3) Electrical Network Load data; and 4) Historical data of consumptions per class and number of consumers, exogenous factors and predominant land-use.

**Phase 2** refers to data pre-loading in data bank for spatial forecast: 5) Data Adjustment, 6) Load curve data, including addition of several consumers 7) Year Base Map.

**Phase 3** aims the building of a spatial model and the allocation and classification of the transformer station in small areas. It basically carries out the following activities: 8) Spatial Model; 9) Pattern Identification; 10) Global Model; 11) Exogenous Factors Model and 12) Scenarios - Planner. It builds itself within this methodology.

**Phase 4** The following activities are herein carried out: 13) Future Load Allocation Model and 14) Forecast Maps.

Activity 12) involves the planner, by proposing and validating scenarios, based on the information generated in activities 8), 9), 10) e 11).

**CLASSIFICATION AND CLUSTERING APPLIED TO SMALL AREA ANALYSIS**

Phase 3 in Spatial Forecast includes the spatial model and pattern identification of small areas. The spatial Model demands allocation (classification) of transformer utilities within the geographical area given by small areas. The pattern identification process precedes the use of cluster analysis techniques, in order to reach typical patterns of small areas taking the geographical area studied.

The use of statistical & artificial intelligence techniques might solve these problems. For Phase 3 to carry on, one needs definite elements in Phases 1 and 2 (data collection and treatment), when the addition of load curves, calculated with basis on network consumers and institutes, such as transformer stations, circuits and distribution transformer stations.

The Kohonen Self-Organizing Map (SOM) was used to classify electrical network transformers within a given geographical area. SOM neural network [7, 8] is used as a classifier when relating electrical loads represented by transformer stations to particles. The classification takes place over the operation phase of the neural network. It is carried out with basis on geographical coordinates, to which the transformers within the geographical area are associated.

After reaching a spatial model for the year-base, every small area bear the information needed, including attributes dependent on the allocated transformers. One then determines a series of small area patterns. A pattern identification process is then carried out for that purpose.

When classification problems are concerned, one must beforehand know class identification, as well as the components of a few samples from these classes.

For clustering problems, it is necessary classify similar objects, contemplating a sample set and distance relationships that might derive from descriptions on the samples. A Kohonen Neural Network [8] might be used in order to carry clusterings out. These approach was chosen for implementation, aiming at the use of artificial intelligence techniques in spatial load forecasting.

Being composed of two layers of knots (input and competition layers), the Kohonen Neural Network is used to verify the potentiality and adaptation of such technique to solve the pattern identification problem for small areas. The
second layer represents the very heart of the network, inasmuch as the unsupervised learning process takes place there.

A few neural networks carry out clustering by initially having one knot haphazardly react to the presentation of input samples. Knot bearing higher exits for an input sample learn how to react more soundly to that sample and to other input samples neighboring that one. Hence, different knots specialize and strongly respond to dissimilar input samples of clustering. This method is analogous to statistical approximation of k-nearest neighbor clustering, where each sample is placed in the same cluster, as its most immediate neighbors.

Portraying the network type based on a map of characteristics validates the Kohonen network’s competence for organizing the attributes of input data in homogeneous clusters.

For this particular case, the Kohonen neural network’s structure presents an input layer bearing a variable amount of neurons (2 to 7), following the variables examined, according to the formulation of the method.

GLOBAL MODEL - APPROACH VIA NEURAL NETWORK

In Spatial Load Forecasting, determining the Global Model corresponds to another activity in Phase 3 of proposed methodology. The long time Forecast Global Model aims at forecasting, based on the consumption background and on the number of consumers [7].

A Back-Propagation Neural Network is used to carry out Global Load Forecast, considering background values for the different consuming types: residential, commercial, industrial and rural among others.

Multilayer Perceptron Networks (MLP) [8] characterize for bearing a much higher computational power than that born by networks without hidden or intermediary layers. Theoretically, intermediary double layered neural networks might implement any function, be it linearly detachable or not. The implementation of an objective function and the precision reached depend on the number of neurons used in hidden layers.

A variety of activation functions have been proposed to be applied to MLP Networks. The sigmoid logistic activation function is mostly used. An MLP learns a predetermined input/exit sample set, by employing a propagation/adaptation cycle. Once a vector of input data is applied as stimulus for the first layer of network units, it will propagate through all higher layers, until generating an exit sign. The exit sign then compares to the desired exit and an error sign is calculated for each exit unit.

CONCLUSIONS

This paper presented the data mining techniques applied to a spatial load forecasting methodology, aiming at its application in distribution systems.

Data preparation is a fundamental stage in the knowledge discovery process.

Neural networks have show its effectiveness in classification, prediction, and cluster analysis tasks.

ACKNOWLEDGMENTS

To FAPESP, for the financial help offered for the development of this research.

REFERENCES