Data mining Applications in Egyptian Official Statistics

Salah S. Nassar
Chairman’s Information technology Consultant, CAPMAS, Egypt
snassar@capmas.gov.eg

Mohamed Sharkawy
Modelling Researcher, DWH Dept., CAPMAS, Egypt
m.sharkawy@capmas.gov.eg

Abstract

Data mining is used for a variety of purposes in both the private and public sectors. The different algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases. Electricity is more important in this world, the main objective of Electricity studies is to analyse carrying out generation, transmission and distribution in an effective manner and supply quality power to consumers. Availability of power is one of the biggest inputs necessary for the sustained growth of any economy. This becomes even more important for a country like Egypt, which is one of the most industrialized Countries in the Middle East. In this study focus on how the various data mining techniques are used in electricity forecasting for Egypt.

After measuring all those methods, they can detect a clear movement toward new, stochastic, and dynamic forecasting performances. It appears a lot of current research effort is absorbed on three such methods: Fuzzy Logic, Particularly Neural Networks and Expert Systems that prove that Egypt has been facing massive power deficits. According to the Central Agency for Public Mobilization and Statistics (CAPMAS) Electricity and Power surveys and Distribution Demand Function research, Egypt was expected to have a power deficit of around 18% in 2015-16[6]. As a result, most of Egypt is now facing huge power cuts. On an average, 3-4 hours of power cuts are being experienced by consumers in the governorate.

Keywords: Clustering, data warehouses applications, Load curve analysis, and Electricity Forecasting and time series analysis.

1. Introduction

In this study we will define the Mining operations as a new field at the frontiers of statistics and information technologies applied in the Data warehouse Department in Central Agency for Public Mobilization and Statistics that include survey design, processing and analysis, Applications of new mines such as texts, web and symbolic data mining, database management, artificial intelligence, machine learning, etc. and aim at discovering structures, patterns and conducting many innovative applications of data quality mining, computational mathematics, statistical analysis, Cognitive patterns mechanisms, mathematical CGEs models and fully descriptive Algorithms using high-precision and professional tools.

Data and text mining algorithms applied in CAPMAS official Databases Such as Foreign Trade, Labour, Electricity, services and many other inter-related data sources offer extensive possibilities of finding models relating variables together, linking Commodities codes, names and Units, Linking a wide range of questionnaires text questions or text notes which change periodically, and measure Data Quality with its relevant imputations procedures, and to support the decision makers with advanced analytics techniques such as neural networks and decision trees, which finds non-linear models, graphical models (or Bayesian belief networks) etc. give a valuable representation of relations between variables. Besides classical data-base where data are usually presented in form of a rectangular array, new kinds of data are now present:

1. Symbolic data applied in Electricity Survey inter-related with other surveys represent fuzzy data or intervals that not known precisely but belongs to an interval (Quartiles), with or without a probability distribution.
2. Text mining: most of the information which circulates is now digitalized as word processor documents, and powerful techniques are currently available for a wide variety of applications: for example: Classification tree models were also used in Foreign trade panel survey to identify characteristics of operations with specific reporting errors of commodities codes, units and description without requiring predefined classes, or assign documents to one or more user-defined categories.

2. Data Quality Mining

Data Quality Mining (DQM) applied in different types of CAPMAS databases with different 140 Questionnaire designs / 124 Statistical research and Publications Databases along ten years inter-related Time series (monthly, quarterly and annually) and different challenges to be managed by professional Tools like microstrategy and Teradata can be defined as the deliberate application of DM techniques for the purpose of data quality measurement and improvement, these databases and censuses cover with represented individuals and households samples the Egyptians characteristics, Expenditures, income and Consumption, Institutes and foundations in different Economic activities prospects with time series long 1996 to 2013, the goal of the DWH applied entities DQM is to detect, quantify, explain, and correct data quality deficiencies in very large, inter-related CAPMAS databases. There are many starting points to employ today’s common DM methods for the purposes of DQM such as:

1. Census Non-response Weighting in Egypt labour market Panel survey (ELMPS). Classification tree models were used to divide the 2006 census records into response propensity groups representing weighting adjustment cells.
2. Allocation of Survey Incentives in Inter-related Electricity Surveys Data sources.
3. Prediction of Survey commodities seasonality. We are also currently developing models using decision trees to predict Foreign Trade survey non-respondents seasonality checks of Imports and Exports Commodities based on several sets of historical data.
4. Questionnaire Design and Construction for Egypt labor market Panel survey (ELMPS).
5. Survey Data Edit Design in Foreign Trade Survey. Association analysis can be applied to survey data set, treating each record as a basket and individual data in each record as the items in the basket. This will generate many known relationships between items.

Based on these innovative CAPMAS DWH applications there are many Methods for deviation and outlier detection seem promising, but it is also straight forward to employ clustering approaches and dependency analysis for data quality purposes. In addition, if we are able to supply training data prepared by a human then also classifiers might do a good job. It is even conceivable those neural networks and artificial intelligent utilized to recognize data deficiencies, we can classify the application of DM to improve data quality in these four important aspects:

1. Measuring and explaining data quality deficiencies,
2. Correcting deficient data,
3. Extension of KDD process models to reflect the potentials of DQM,
4. Development of specialized process models for pure DQM.

3. Data Mining Techniques in Electricity Load Profiling Official Statistics

An important feature present in most of these techniques is an ability to adapt to the local characteristics of the data. Such techniques are applied to electric load profiling tasks; load profiling consists of modelling the way in which daily load shape (load profile) relates to various factors such as weather, time and customer characteristics. An implementation of an adaptive load profiling methodology is presented. The main objective of Electricity mining is to carry out generation, transmission and distribution analysis in an effective manner and supply quality power analysis to consumers. Availability of power is one of the biggest inputs necessary for the sustained growth of any economy so Electrical Supply Industries (ESI) in Egypt has been modernized with the determination of introducing levels of competition into energy generation and retail energy sales. In any market with levels of competition information of future market conditions can contribute to giving market participants an economical advantage over their fellow market participants. With the increase in complexity and an estimated growth of 3-7% electric load per year according to CAPMAS Energy Forecasting Model Load Profiling Module and Classification Modules , the various factors that have become influential to the electric power generation and consumption are load management, energy exchange, spot pricing, independent power producers, non-conventional energy, generation units, etc.

One of the most important CAPMAS Official statistics portfolios presents an electricity consumer characterization framework based on a knowledge discovery in databases (KDD) method, sustained by data mining (DM) methods, applied on the different phases of the process. Two leading modules compose that framework:

Module 1: generates a set of consumer classes using a clustering operation and the demonstrative load profiles for each class. Module 2 uses this knowledge to construct a classification model able to allocate different consumers to the prevailing classes.

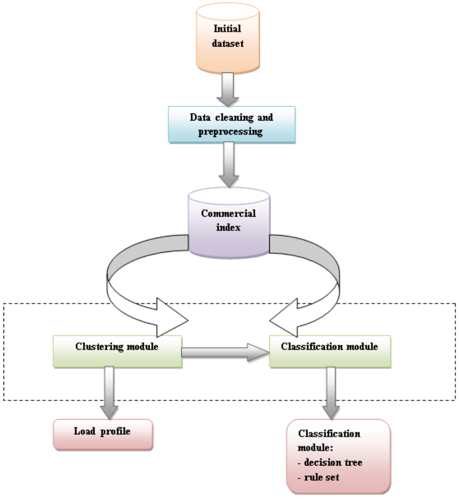

Figure 1: Structure of the Customer Characterization Framework

Figure 1 has following major steps:

1. Data Selection: the initial database, of important measurement, stores data related with electricity consumers. The first step is the selection of the data with more implication to the process. This selection is made allowing to the voltage level of the consumers. Separate readings must be directed to different voltage levels.
2. Data Cleaning and Pre-processing: In the cleaning phase we check for variations in the data and outliers are removed using the resulting technique. Inconsistent consumption values and outages are discovered and exchanged based on the information of similar days. In the pre-

processing phase disappeared values are discovered and exchanged using regression techniques.

3. Data Reduction: The data reduction is made using foregoing information about the way the loading environments, similar the season of the year and the type of weekday (working days or weekends), affect electricity consumption.
4. Data Mining: This step includes collection and submission of the data mining techniques. This is made using one isolated method or merging several methods, to form a model able to discover related knowledge about the changed consumption patterns found in the data. The application of the model contains several steps, like attribute selection, fitting the models to the data and evaluating the models.
5. Interpretation of the discovered knowledge: In this step the familiarity discovered by the DM model is enhanced. This information can provide new perceptions into associations between data elements and facilitate more productive and sophisticated DSS applications.

Electricity market price forecast is a changeling yet very important task for electricity market managers and participants. The proposals of data mining based electricity price forecast framework, which can predict the normal price as well as the price spikes. The spike prediction and explores the details for price spikes based on the amount of a future complex are Supply Demand Balance Index (SDI) Relative Demand Index (RDI).

The future model is based on a mining database including market clearing price, trading hour, electricity demand, electricity supply and backup. The mining outcomes are used to form the price spike forecast model. That future model is able to create forecasted price spike, level of spike and associated forecast self-confidence level.

We will compare the accurateness of six univariate techniques for short-term electricity demand forecasting for lead times up to a day ahead. The very short lead times are of particular importance as univariate methods are often exchanged by multivariate methods for forecast beyond nearly six hours ahead. The methods considered include the recently proposed exponential smoothing method for double seasonality and a new method based on Principal Component Analysis (PCA). The PCA method implemented well, but general the best results were reached with the exponential smoothing method, leading us to arrange that humbler and more forceful methods, which involve little domain knowledge, can overtake more complex replacements.

Forecasting electricity prices in present-day competitive electricity markets is a must for both manufacturers and users because both need price evaluations to develop their separate market bidding approaches. The proposes an allocation function model to forecast electricity prices based on both previous electricity prices and demands, and deliberate the foundation to build it. The significance of electricity demand information is measured. Appropriate metrics to evaluate prediction quality are well-known and recycled. Representative and general imitations based on data from the PJM Interconnection for year 2010-2013 are accompanied. The planned model is associated with naive and other procedures.

Presenting different procedures have been applied to load forecasting. Nine approaches have been reviewed there are: Exponential Smoothing, Iterative Reweighted Least-Squares, Fuzzy Logic, Adaptive Load Forecasting, ARMAX Models Based On Genetic Algorithms, Multiple Regression, Neural Networks, Stochastic Time Series and Expert System.

There is also significantly less significance on techniques such as iterative reweighted least-squares and adaptive load forecasting.

Over the last few years, Egypt has been facing massive power deficits. According to the CAPMAS Electricity and Power surveys (Distribution Demand Function), Egypt was expected to have a power deficit of around 18% in 2015-16. As a result, most of Egypt is now facing huge power cuts. On an average, 3-4 hours of power cuts are being experienced by consumers in the governorate. The impact of this power shortage is being felt mainly by the industries, leading to a loss in efficiency and production. Forecast of power demand is crucial for an effective process of any utility corporation. A fuzzy version of neural network, namely Fuzzy back propagation network (Fuzzy BP) has been developed for short term electric load prediction. The planned construction consists of a unit with 51 inputs and 24 outputs. The inputs are fuzzified and the outputs are crisp values representing the expected load. The suggested method is implemented in MATLAB. The reproduction results are accessible for each day (24 hours) of the week. Besides that, a multi-layer perceptron (MLP) was also applied independently and the load was forecast using back propagation algorithm. The effects achieved from Fuzzy BP were established to be satisfactory when compared to those of MLP network. Grid Computing is a talented arrangement and knowledge that involves the cohesive and cooperative use of Computers, networks, databases and scientific instruments owned and managed by multiple organizations. Today over 21% of the total electrical energy produced in Egypt is lost in broadcast and delivery. It is possible to transport miserable the delivery losses to the minimum level in Egypt with the help of newer technological options in information technology called Grid Computing in the electrical power delivery area which will enable superior observing and control. When the Power factor is observed using this Grid computing method, the Substations can maintain their averages by giving suitable information to the operatives dynamically to add the capacitor banks to maintain the power feature so that the line loss will be reduced and the profits will automatically be enhanced.

Our case study introduces a new method for daily Peak Load forecasting using combinations of accomplished Artificial Neural Networks (ANNS): Constrained and Unconstrained.

ANNS will result these two different methods to identify various combinations of ANNS for peak load forecasting. A set of neural networks are qualified with different construction and with different knowledge restrictions. The neural networks are accomplished and verified for the definite peak load data of Egypt Governorates. A set of well qualified ANNS are nominated to develop various arrangements using these two methods as an alternative of using a single best expert neural network. Achieved test results using the Permutations of ANNs demonstrate its validity. Time series forecasting is an interesting task in many fields. Due to the multifaceted non-linear association between the multidimensional features of the time series data, enhanced time series forecasting needs a predicting model that combines multiple expectation models. Introduces a different two level combined learning method: Radial Basis Function networks (RBF), K - Nearest Neighbor (KNN) and Self Organizing Map (SOM).

These methods are used for time series prediction with the purpose of increasing the forecast accuracy. The valuation of the planned Pattern Prediction Ensemble Model (PAPEM) using three input datasets such as Mackey dataset Sunspots dataset Stock Price dataset.

This inputs demonstration that the proposed PAPEM model achieves better than the specific classifiers. Mining Time Series data has a marvellous progress of importance in today’s world. To provide a suggestion various applications are considered and shortened to identify the different complications in remaining applications. Clustering time series is a trouble that has uses in an extensive range of fields and has recently involved a large amount of research. Time series data are commonly large and may contain outliers. In addition, time series are a special type of data set where features have a progressive ordering. Therefore clustering of such data stream is a significant problem in the data mining procedure. Numerous methods and clustering algorithms have been planned previous to support clustering of time series data streams. They present a survey on various clustering algorithms available for time series datasets.

4. Conclusions

This applied DM techniques demonstration that the proposed PAPEM model achieves better than the specific classifiers. Mining Time Series data has a marvellous progress of importance in today's world. To provide a suggestion various applications are considered and shortened to identify the different complications in remaining applications. Clustering time series is a trouble that has uses in an extensive range of fields and has recently involved a large amount of research. Time series data are commonly large and may contain outliers. In addition, time series are a special type of data set where features have a progressive ordering. Therefore clustering of such data stream is a significant problem in the data mining procedure. Numerous methods and clustering algorithms have been planned previous to support clustering of time series data streams. They present a survey on various clustering algorithms available for time series datasets.

5. References

[1] Amit Jain, and B. Satish," Clustering Based Short Term Load Forecasting Using Support Vector Machines". International Institute of Information Technology, October 14, 2009.

[2] Chitra and S. Uma, "An Ensemble Model of Multiple Classifiers for Time Series Prediction". International Journal of Computer Theory and Engineering, Vol. 2, No. 3, June, 2010, ISSN: 1793-8201.

[3] FJ Nogales and AJ Conejo, "Electricity price forecasting through transfer function models". Journal of the Operational Research Society (2006) 57, pp.350–356, ISSN: 0160-5682.

[4] Francisco Martínez_Álvarez, Alicia Troncoso, José C. Riquelme and Jesús S. Aguilar_Ruiz, "Energy Time Series Forecasting Based on Pattern Sequence Similarity", 2011.

[5] Hari Seetha and R. Saravanan, "Short Term Electric Load Prediction Using Fuzzy BP". Journal of Computing and Information Technology - CIT 15, 2007, vol 3, pp. 267–282 doi:10.2498/cit.1000987.

[6] Hesham K. Alfares and Mohammad Nazeeruddin, "Electric load forecasting: literature survey and classification of methods". International Journal of Systems Science, 2002, volume 33, number 1, pp. 23±34.

[7] James W. Taylor, Lilian M. de Menezes, Patrick E. McSharry, "A comparison of univariate methods for forecasting electricity demand up to a day ahead". International Journal of Forecasting 22 (2006), pp.1– 16, ISSN: 0169-2070.

[8] Kavitha, M.Punithavalli, "Clustering Time Series Data Stream – A Literature Survey". (IJCSIS) International Journal of Computer Science and Information Security, Vol. 8, No. 1, April 2010.

[9] Sansom, T. Downs and T. K .Saha," Support Vector Machine Based Electricity Price Forecasting for Electricity Markets utilising Projected Assessment of System Adequacy Data". The Sixth [1] International Power Engineering Conference (IPEC2003), pp. 27-29 November 2003, Singapore.

[10] Senthamarai, A.Krishnan and R. Hemalatha, "Performance Monitoring of Energy Flow I in the Power Transmission and Distribution System Using Grid Computing". Journal of Computer Science 3 (5), pp. 323-328, 2007, ISSN 1549-3636.

[11] Subbaraj and V.Rajasekaran, "Peak Load Forecasting Using Optimal Linear Combinations of Artificial Neural Networks". International Journal of Electrical and Power Engineering 2 (l): pp. 50-54, 2008, ISSN: 1990-7958.

[12] Vera Figueiredo, Fátima Rodrigues, Zita Vale, "An Electric Energy Consumer Characterization Framework Based On Data Mining Techniques". IEEE Transactions On Power Systems, Vol. 20, No. 2, May 2005.

[13] Xin Lu, Zhao Yang Dong, Xue Li, "Electricity market price spike forecast with data mining techniques". Electric Power Systems Research Volume 73, Issue 1, January 2005, pp.19–29.

[14] Barsoum, Ghada. 2007. Egypt Labor Market Panel Survey 2006: report on methodology and data collection. Economic Research Forum Working Paper No. 0704.

[15] Evaluating Efficiency of Statistical data editing, United Nations statistical commission and economic commission for Europe.

[16] Statistical Data Editing, Impact on data Quality, United Nations statistical commission and economic commission for Europe.

[17] K.Perritt. 2000. A look into AGGIES, An Automated Edit and Imputation System.

[18] J.Kovar, P.Whitridge and J.Mac2Lillan. 1988. Generalized edit and imputation system for economic surveys at statistics Canada.