Intelligent Information System as a Tool to Reach Unaproachable Goals for Inspectors

High-Performance Data Analysis for Reduction of Non-Technical Losses on Smart Grids

Juan I. Guerrero, Antonio Parejo, Enrique Personal, Félix Biscarri, Jesús Biscarri and Carlos León
Department of Electronic Technology
University of Seville
Seville, Spain
e-mail: juaguealo@us.es

***Abstract*—The Non-Technical Losses (NTLs) represent the non-billed energy due to faults or illegal manipulations in customer facilities. The objective of the Midas project is the detection of NTL through the application of computational intelligence over the information stored in utility company databases. This project has several research lines. Some of them are pattern recognition, expert systems, big data and High Performance Computing (HPC). This paper proposes module which use statistical techniques to make patterns of correct consumption. This module is integrated with a rule based expert system with other modules as: text mining module and data warehousing module. The correct consumption patterns are generated using rules which will be used in rule based expert system. Two implementations are proposed. Both implementations provided an Intelligent Information System (IIS) to reach unapproachable goals for inspectors.**

Keywords- non-technical losses; pattern recognition; expert system; big data analytics; high performance computing.

I. INTRODUCTION

The information systems have provided a new advantage: the capability to store, manage and analyze great quantities of information, without human supervision. This paper proposes one solution to a very difficult problem: the NTLs reduction.

NTLs represent the non-billed energy due to the abnormalities or illegal manipulations in client power facilities. The objective of Midas project is the detection of NTLs using computational intelligence and Knowledge Based Systems (KBS) over the information stored in Endesa databases. Endesa is the most important utility distribution company in Spain with more than 12 million clients. Initially, this project is tested with information about customers of low voltage. The system uses information of consumers with monthly or bimonthly billing. Although the system can analyze large volume of data, the system has a very high cost in time when there are more than 4 million consumers. Notwithstanding, this information volume will be unapproachable to analyze for an inspector. In order to reduce this cost, a hybrid architecture based on big data and high performance computing is currently applied to create a high-performance data analysis (HPDA). This architecture has been successfully applied in biomedical topics [1], text data classification [2], and other scientific datasets [3]. Smart Grids have provided a new scope of technologies, for example, Advanced Metering Infrastructures (AMI) with smart metering, Advanced Distribution Automation (ADA), etc. These new infrastructures increase the information about consumer, taking hourly or even quarterly measures.

In terms of consumption in utilities, a great spectrum of techniques can be applied; some techniques can be data mining, time series analysis, etc. Basically, it is essential the use of any type of statistical technique to detect anomalous patterns. This idea is not new. Several works usually apply statistical or similar techniques to make analysis of anomalous consumption [4][5][6][7][8]. Some techniques are based on study of consumption of the historical customer consumption; for example, Azadeh et al. [9] made a comparison between the use of time series, neural network and ANOVA, always with reference of the consumption of the same customer. But, these techniques have several problems, the main problem being that it is necessary to have a large historical data about consumption of customer. Other works use different studies to make good patterns of consumption, which compare the consumption of a customer with others who have similar characteristics. For example, Richardson [10] compared both neural networks and statistical techniques; in the tests performed, statistical techniques are 4% more efficient than neural networks. Hand et al. [11] proposed the identification of some characteristic which allow the identification of consumption patterns applying statistical techniques that use them as anomalous patterns. Other methods propose the use of advanced techniques to make other references or patterns of consumption. In this sense, Nagi et al. [12] used support vector machines and [13] applied rough sets, both of them in NTLs detection.

Other applications of advanced techniques, mainly Artificial Neural Network (ANN), which are not used for detection of NTLs, but could be used, are the applications for demand forecasting. In this sense, the forecasting can be made in short [14], medium [15], or long [16] term.

This paper proposes a model which uses statistical technique to detect correct consumption patterns. These patterns are used to generate rules which are applied in a Rule Based Expert System (RBES). The RBES is described in [17][18] and the module of text mining is described in

Copyright (c) IARIA, 2016. ISBN: 978-1-61208-518-0

[19]. In this paper, an increase of functionality of the data mining module is proposed. In Fig. 1 and Fig. 2 the system architecture is shown.

The proposed solution is described in the following sections. In Section II, the architecture and technical characteristics are described. In Section III, the characterization of correct consumption is proposed. In Section IV, the evaluation and experimental results are explained. In Section V, the conclusions are included. Finally, in Section VI, the future research lines are described.

II. ARCHITECTURE AND TECHNICAL CHARACTERISTICS

Initially, the architecture of applications is shown in Fig. 1. This architecture is detailed for Statistical Pattern Generator, showing the different stages of this process. The system was run in a single machine, and it has been successfully tested with four million clients. This volume of analysis forced the system to do partitions in order to analyze more than 4 million customers.

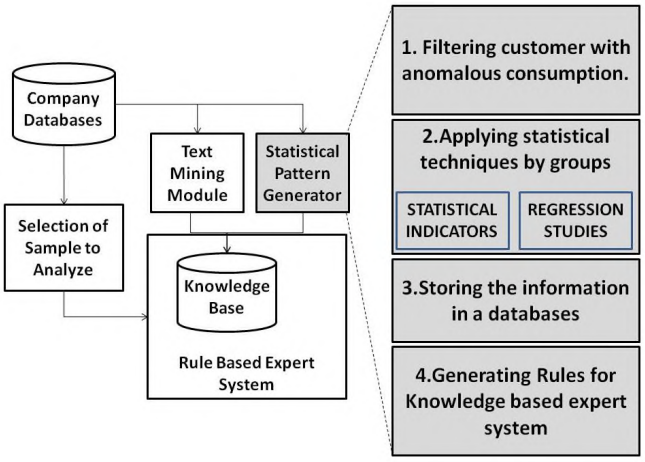

Figure 1. Local Expert System Architecture and Details of Statistical Pattern Generator Module

Currently, the new architecture applies big data and HPC. The big data architecture is based on Apache Spark with a database stored in HBase implemented in Apache Hadoop. The analytics are implemented in MLlib, GraphX, and library to send jobs to Graphics Processor Units (GPUs, based on Compute Unified Device Architecture or CUDA cores). The architecture is shown in Fig. 2.

The architecture is based on Apache Hadoop and Spark, enhanced with a new daemon to take advantage of HPC architectures. This daemon, named *gpulauncher*, is invoked by processes in order to send jobs to GPUs. The processes can be part of a MapReduce or Analytics. Although the system implemented statistical and regression algorithms, the new algorithms will also apply multivariable inference.

The *gpulauncher* daemon implemented several algorithms for analytics, functions for streaming the data to GPUs and functions for synchronization of nodes. These synchronization functions are in development. The main objective is the synchronization between GPUs of different nodes and working in near-real-time (NRT).

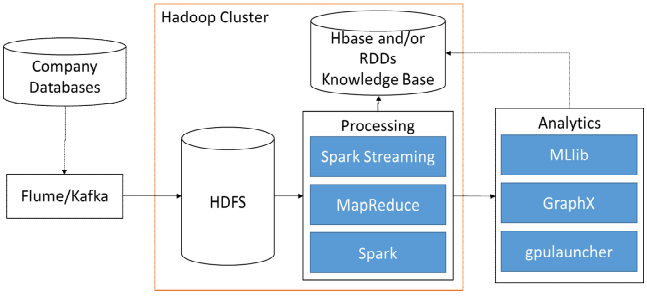

Figure 2. Architecture of Expert System in a High-Performance Data Analysis

The proposed system was not deployed in a real cluster of machines. The cluster was implemented with two virtualized servers. The first server has an Intel i7 (3GHz), 16GB RAM and GTX750 (2GB and 640 CUDA cores). The second server has an Intel Xeon E5 (2GHz), 64GB RAM and Quadro K1200 (4GB and 512 CUDA cores).

III. CONSUMPTION CHARACTERIZATION

To find out which characteristics have more influence in consumption is a very difficult task because there are a lot of consumption information available. An in-depth analysis shows that some characteristics have more influence over the consumption: time, geographical location, postal code, contracted power, measure frequency, economic activity, and time discrimination band. The importance of these characteristics has also been analyzed in other utilities as gas utility. Moreover, the results of these analysis have been compared with the knowledge provided by Endesa inspectors.

Each characteristic by itself is not efficient because the consumption depends on several characteristics at the same time. Thus, grouping characteristics can help find patterns of correct consumption, because these characteristics can determine the consumption with a low level of error rate providing, at least, one consumption pattern. These groups have in common a series of characteristics: Geographical location, time, contracted power, and measure frequency. These are named *Basis Group* because these are the main characteristics. The values for each of these characteristics are wide; therefore, each of them shows great variations of consumption. A description of Basis Group and the other groups are shown in Table I.

TABLE I. GROUPS OF CONSUMPTION CHARACTERISTICS

Consumption Characteristics	Description
Basis Group	This group provides consumption patterns by general geographical location: north, south, islands, etc.

Basis Group and Postal Code	This group provides patterns useful for cities with coastal and interior zones.
Basis Group and Economic activity.	The granularity of geographical location is decreased. In this way, the economic activity takes more importance. Nevertheless, the geographical location cannot be despised because, as for example, a bar has not the same consumption whether it is in interior location or coastline location.
Basis Group and time discrimination band.	There are several time discrimination bands. Each band registers the consumption at a different time range. This group provides consumption patterns in different time discrimination bands. These are usefull because there exists customers who make their consumption in day or night time.

Some characteristics have different granularity because they have continuous values or have a lot of possible values. The granularity is used because there are some problems related with the measures. For example, the proposed framework performs a discretization of contracted power in 40 ranges. In the graph of Fig. 3, the 14th range of contracted power is shown. This range groups the contracted power between 46,852 kW and 55,924 kW in North of Spain. This figure shows an abnormal level of consumption at 2002; this fact represents errors in measures which cannot be filtered.

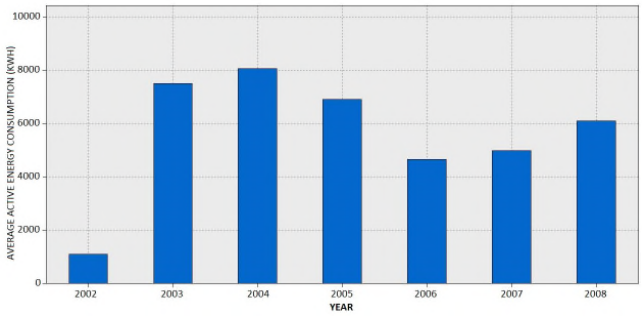

Figure 3. Average yearly consumption graph in different power ranges

In the graph of Fig. 4, the average consumption in monthly periods for the 14th range is shown. In this case, the granularity of time is increased; therefore, it is possible to get another pattern, which is better than the one obtained from the graph of Fig. 3. In this case, the consumption can be analyzed monthly.

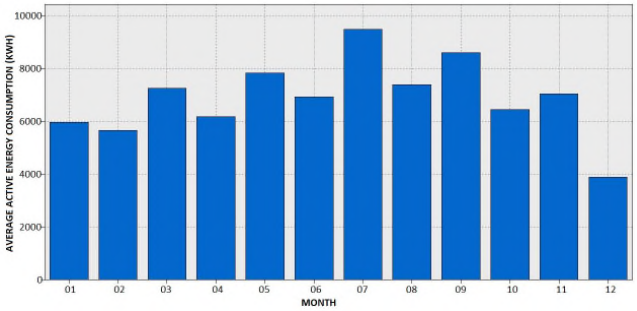

Figure 4. Average monthly consumption graph in different power ranges

Thus, several time ranges are used: absolutely, monthly, yearly and seasonally. For example, the average consumption calculation provides different results: total average consumption (absolutely), twelve/six average, monthly/bimonthly consumption, one average yearly consumption (when the measures are available), and four average seasonally consumption. In the same way, the contracted power has to be discretized in equal consumption ranges. In lower contracted power, the ranges are very narrow because there are a lot of consumers. When the contracted power is higher, the quantity of consumers is smaller, although the consumptions are very different. The reason for aggregating the consumption (of supplies without NTL) in different groups is because there are scenarios in which it is necessary to have other patterns.

These groups provided dynamic patterns, which can be updated according to the time granularity. Once the characteristics are identified, it is necessary to design a process which finds patterns automatically. Initially, these studies were made bimonthly and were applied as a part of an integrated expert system to model correct consumption patterns (Fig. 1). Currently, the process can be performed hourly, through the architecture proposed in Fig. 2. The system applies statistical techniques to get consumption patterns using the process detailed in Fig. 1.

When the rules are created, they are used to analyze the customers in order to determine if there exist any NTLs. There are defined series of rules in RBES which use the information generated by the proposed module. The antecedents of the rules are generated dynamically using the patterns generated in the described process and according to the characteristics of the customer who will be analyzed. In this way, the use of memory resources is minimized because only the necessary antecedents of the rules are generated.

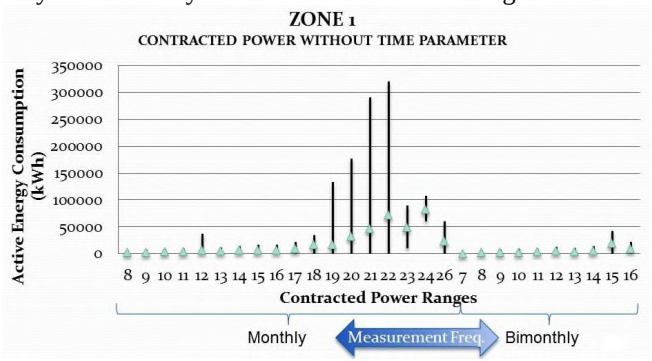

Figure 5. Graph of Active Energy Consumption Ranges vs. contracted power ranges without time parameter for specific geographical zone.

When the consumption of a customer is analyzed, several rules can fit with the characteristics of that customer. Initially, the rules are applied in the most restrictive way; this means, the customer consumption will be correct if it fits in any correct consumption pattern. Moreover, the system notifies if the pattern fails for each customer. For example, the correct consumption ranges of active energy for specific geographic zone, different contracted power

Copyright (c) IARIA, 2016. ISBN: 978-1-61208-518-0

ranges, and different measurement frequency (monthly or bimonthly) are shown in Fig. 5.

IV. EVALUATION AND EXPERIMENTAL RESULTS

The proposed module provides patterns of correct customer consumption. The analysis made by the mentioned expert system uses this module to create rules. The customer consumption analysis applies these rules according to the contract attributes: contracted power, economic activity, geographical location, postal code, and time discrimination band. Traditionally, the systems used to detect frauds or abnormalities in utilities make patterns for NTLs detection. But in the proposed system, models of correct consumption ranges and trends are made. The use of these patterns increases the efficiency of the RBES. This module is essential to analyze the customer. The RBES has been applied in real cases getting better results in zones with a lot of clients. The success of the RBES is between 16,67% and 40,66% according to the quantity of clients of the corresponding location. This fact is shown in Fig. 6. The new architecture based on high-performance data analysis allows the application of the expert system in NRT. Thus, this system will be useful in the new Smart Grid infrastructures, based on AMI.

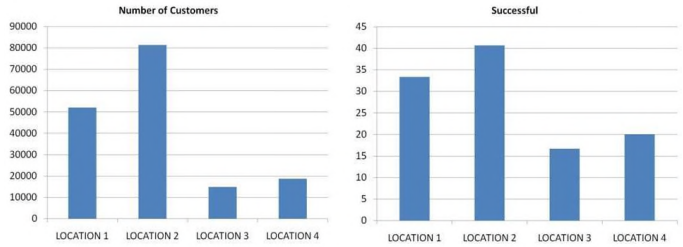

Figure 6. Number of customers vs. successful

V. CONCLUSION AND FUTURE WORK

The proposed framework was implemented, tested and deployed in a real Power Distribution Company. This framework is part of a RBES. This model establishes a series of similarities with other utilities. For example, the utilization of frequency billing, geographical location, and time can be made in all utilities. However, the contracted power can be replaced by the contracted volume of flow in gas or water utilities.

The proposed module can be added to other systems of NTLs detection to increase their efficiency by using rules or a translator of the knowledge generated by the module.

Usually, an inspector takes between 5 to 30 minutes to analyze the information about a customer, in order to confirm whether there exists a NTL. This period depends on the quantity of information to be analyzed; the average time of the analysis process takes 16,3 minutes. This means that the time to analyze 4 million customers (the maximum number of customers in case proposed in Fig. 1) would be 1086666,6 hours of work. In the first case, the proposed system in Fig. 1 takes 22 milliseconds per customer in the analysis process. The HPDA provides the possibility to analyze the information in NRT, without limit in the number of customers. Notwithstanding, the analysis of the inspector will be always better than the machine analysis because inspectors usually work in the same zone and they have additional knowledge of facilities, that is not stored in the system.

Finally, several research lines for improving the efficiency of the proposed framework will be addressed:

- Application of techniques related with Information Retrieval, to increase the information about consumers.
- Test the new approach in a big scenario, based on AMI infrastructure and with hourly measures.
- Application of the proposed framework in other utilities.
- Enhance the analysis with application of multivariable inference.

ACKNOWLEDGMENT

The authors would also like to thank the backing of SIIAM project (Reference Number: TEC2013-40767-R) which is funded by the Ministry of Economy and Competitiveness of Spain.

REFERENCES

[1] E. Elsebakhi et al., “Large-scale machine learning based on functional networks for biomedical big data with high performance computing platforms,” *J. Comput. Sci.*, vol. 11, pp. 69–81, Nov. 2015.

[2] A. Rauber, P. Tomsich, and D. Merkl, “parSOM: a parallel implementation of the self-organizing map exploiting cache effects: making the SOM fit for interactive high-performance data analysis,” in *IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks, 2000*, vol. 6, 2000.

[3] J. Liu and Y. Chen, “Improving Data Analysis Performance for High-Performance Computing with Integrating Statistical Metadata in Scientific Datasets,” in *High Performance Computing, Networking, Storage and Analysis (SCC), 2012 SC Companion:*, 2012.

[4] I. Monedero et al., “Using regression analysis to identify patterns of non-technical losses on power utilities,” in *Knowledge-Based and Intelligent Information and Engineering Systems*, Springer, 2010, pp. 410–419, 2010.

[5] C. C. Ramos, A. N. de Sousa, J. P. Papa, and A. X. Falcão, “A New Approach for Nontechnical Losses Detection Based on Optimum-Path Forest,” *IEEE Trans. Power Syst.*, vol. 26, no. 1, pp. 181–189, pp. 181–189, Feb. 2011.

[6] M. E. de Oliveira, D. F. Boson, and A. Padilha-Feltrin, “A statistical analysis of loss factor to determine the energy losses,” presented at the Transmission and Distribution Conference and Exposition: Latin America, 2008 IEEE/PES, 2008.

[7] M. Gemignani, C. Tahan, C. Oliveira, and F. Zamora, “Commercial losses estimations through consumers’ behavior analysis,” presented at the 20th International Conference and Exhibition on Electricity Distribution - Part 1, 2009. CIRED 2009, 2009.

Copyright (c) IARIA, 2016. ISBN: 978-1-61208-518-0

[8] A. H. Nizar and Z. Y. Dong, "Identification and detection of electricity customer behaviour irregularities," presented at the Power Systems Conference and Exposition, 2009. PSCE '09. IEEE/PES, 2009.

[9] A. Azadeh, S. F. Ghaderi, and S. Sohrabkhani, "Forecasting electrical consumption by integration of Neural Network, time series and ANOVA," *Appl. Math. Comput.*, vol. 186, no. 2, pp. 1753–1761, Mar. 2007.

[10] R. Richardson, "Neural networks compared to statistical techniques," presented at the Computational Intelligence for Financial Engineering (CIFEr), 1997., Proceedings of the IEEE/IAFE 1997, 1997.

[11] D. J. Hand and G. Blunt, "Prospecting for gems in credit card data," *IMA J. Manag. Math.*, vol. 12, no. 2, pp. 173–200, Oct. 2001.

[12] J. Nagi, A. M. Mohammad, K. S. Yap, S. K. Tiong, and S. K. Ahmed, "Non-Technical Loss analysis for detection of electricity theft using support vector machines," presented at the Power and Energy Conference, 2008. PECon 2008. IEEE 2nd International, 2008.

[13] J. E. Cabral and E. M. Gontijo, "Fraud detection in electrical energy consumers using rough sets," presented at the 2004 IEEE International Conference on Systems, Man and Cybernetics, vol. 4, 2004.

[14] B. F. Hobbs, U. Helman, S. Jitprapaikulsarn, S. Konda, and D. Maratukulam, "Artificial neural networks for short-term energy forecasting: Accuracy and economic value," *Neurocomputing*, vol. 23, no. 1–3, pp. 71–84, Dec. 1998.

[15] M. Gavrilas, I. Ciutea, and C. Tanasa, "Medium-term load forecasting with artificial neural network models," in *Electricity Distribution, 2001. Part 1: Contributions. CIRED. 16th International Conference and Exhibition on (IEE Conf. Publ No. 482)*, vol. 6, 2001.

[16] K. Padmakumari, K. P. Mohandas, and S. Thiruvengadam, "Long term distribution demand forecasting using neuro fuzzy computations," *Int. J. Electr. Power Energy Syst.*, vol. 21, no. 5, pp. 315–322, pp. 315–322, Jun. 1999.

[17] C. León et al., "Integrated expert system applied to the analysis of non-technical losses in power utilities," *Expert Syst. Appl.*, vol. 38, no. 8, pp. 10274–10285, Agosto 2011.

[18] J. I. G. Alonso et al., "EIS for Consumers Classification and Support Decision Making in a Power Utility Database," *Enterp. Inf. Syst. Implement. IT Infrastruct. Chall. Issues Chall. Issues*, p. 103, 2010.

[19] J. I. Guerrero Alonso et al., "Increasing the efficiency in non-technical losses detection in utility companies," 2010.

Copyright (c) IARIA, 2016. ISBN: 978-1-61208-518-0