Applying Diverse Data Mining Methods in the Electric Power Industry

Manuel Mejía-Lavalle, Guillermo Rodríguez O.,
Gustavo Arroyo F. and Eduardo F. Morales[1]

Instituto de Investigaciones Eléctricas, Reforma 113, 62490 Cuernavaca, Morelos, México
[1] INAOE, L.E.Erro 1, 72840 StMa. Tonantzintla, Puebla, México
{mlavalle, gro, garroyo}@iie.org.mx
emorales@inaoep.mx

Abstract. We present our experiences in four Mexican power electric industry domains where we applied diverse data mining techniques. The first domain is about electric generator diagnosis. The second one is related to flashover forecasting in high-voltage insulators. The third case is about obtaining expert knowledge, applying data mining techniques to hydroelectric and thermoelectric utilities databases. The last case approaches a pattern recognition problem to detect potential electric illicit users. We outline successful and bad practices, and comment about possible solutions or future work that we think it have to be done to maximizing the usefulness of the data mining approach.

1 Introduction

Data mining has been employed with success in various fields and in many real world problems [1]. Data mining is applied to huge volumes of historical data mainly with the expectation of finding knowledge, or in other words, when it is sought to determine trends or behavior patterns that permit improve the current procedures of marketing, production, operation, maintenance, or others. In summary data mining, or more widely expressing, knowledge discovery, is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data [2]. Some of the traditionally used computer techniques to accomplish data mining are: neural networks, induction of decision trees, decision rules and case-based methods.

In this paper we present our experiences in several Mexican power electric industry domains where we applied diverse data mining techniques. The first domain is about electric generator diagnosis using expert systems plus a novel neural network paradigm. The second one is related to flashover forecasting in high-voltage insulators, where we proposed several tools to approach this problem. The third case is about obtaining expert knowledge, applying and comparing well known data mining techniques to hydroelectric and thermoelectric utilities databases. The last case approaches a pattern recognition problem to detect potential electric illicit users, where we proposed and realized a pre-processing feature selection method. We outline successful and bad

© M. Martínez, A. Alarcón (Eds.)
Advances in Computer Science and Engineering.
Research Computing Science 45, 2010, pp. 209-220

Received 30/04/10
Final version 19/05/10
Accepted 14/05/10

practices, and comment about possible solutions or future work that we think it have to be done to maximizing the usefulness of the data mining approach.

2 Electric Generator Diagnosis

In this Section we present an Electric Generator Failure Diagnosis (EGFD) system. The EGFD system combines two artificial intelligence approaches: expert systems and artificial neural networks. With our expert system shell we capture the human expertise. With our neural net paradigm we obtain knowledge from data. For instance, human experts on electric generation failures know that:

a) Internal partial discharges can occur within the ground wall insulation at delaminations or areas where the bonding material is missing or incompletely cured.
b) Such discharge activity is particularly common in older insulation systems such as mica folium and asphalt-mica.
c) The main characteristic of this mechanism is that the positive and negative partial discharge activity is about equal.

Then a production rule to detect this condition becomes as follows:

RULE 3
IF: - Insulation system is made of <mica folium> or <asphalt-mica>.
and - [Positive PD at 50 mV] * 1.1 >= [Negative PD at 50 mV].
and - [Positive PD at 50 mV] * 0.9 <= [Negative PD at 50 mV].
and - [Positive PD at 200 mV] * 1.1 >= [Negative PD at 200 mV].
and - [Positive PD at 200 mV] * 0.9 <= [Negative PD at 200 mV].
THEN:
- Bonding material is missing or incompletely cured. Certainty 7. and- Exe (PHAFII).

A number is assigned to each rule: *'3'* in this example. Then, the keyword *'IF'* indicates the beginning of the list of conditions, premises or antecedents of the rule. The first condition is true, if and only if, the user answer to the question: *'Insulation system is made of ?' is 'mica folium' or 'asphalt-mica'*. The second condition is true, if and only if, the variable *[positive PD at 50 mV]* multiply by 1.1 is greater or equal to the variable *[Negative PD at 50 mV]*. The same applies to the rest of the conditions. If one of these conditions happens to be false, because the user answer is different than expected, or because some variable value do not match the required condition, then the rule is false, and the inference machine of the expert system searches for another rule.

On the other hand, if all the conditions of a rule are true, then the rule is true and its conclusion is 'fired': *'Bonding material is missing or incompletely cured'*. The word *'Certainty'* at the end of a rule means 'the degree of certainty' or belief that the human expert has on the rule and it ranges from 0 to 10, where 10 means that the expert is absolutely certain of what the rule states.

With this production rule, the expert system can identify, with 70% certainty or reliability, that 'Bonding material is missing or incompletely cured' if the 'insulation system is made of mica folium or asphalt-mica', as stated in first condition, and if the 'positive partial discharge activity' is similar (within ten percent) to the 'negative partial discharge activity' at 'pulse magnitudes' of 50 mV and 200 mV, as stated in the rest conditions.

Then, our neural net paradigm is called using the command *Exe(PHAFII)*, where PHAF II is the module that handles the neural net. Algorithmic details of PHAF II are in [3]. We used the neural net to take advantage of the enormous amount of information currently available in many electric generator databases. Data from the partial discharge graphs are normalized within the range [0,1] and then fed to the PHAF II neural net. The neural net, previously trained with normalized data from graphs which are typical patterns of abnormal situations, performs the recognition of the fed graph and computes the percentage of similarity using three criteria:

a) 10,000 base, where the graphs are compared using a lineal scale from 0 to 10,000 of frequency units. With this criterion, the differences or likenesses of the graphs have the same weight at high and low frequencies.
b) Logarithmic base, where the graphs are compared using a logarithmic scale from 0 to 10,000 of frequency units. With this criterion, the differences or likenesses of the graphs are adjusted with more weight given to the differences in the low frequencies (0 to 100) and less weight to the differences at the high frequencies (100 to 10,000).
c) Central base, where the graphs are compared using as a reference the pattern graph.

With the mean (average) of these three criteria, we obtain a final certainty factor. This factor indicates the similarity of the fed graph and the pattern graph. If the final factor is greater than 70% the system displays the screen shown in Fig. 1.

From Fig. 1, it is observed that the system displays the graph being recognized, the diagnosis, and eight certainty factors. Four of these correspond to a 'Global' analysis (GCF), where the certainty factor is computed as the mean of the likenesses or differences at all the points of the graphs. The other four certainty factors, called 'Local' (LCF), are obtained from the same point on the graphs where there exists the greatest distance between the graphs (the test graph and the pattern graph).

We planning, as future work, incorporate more human and data knowledge to this system. To efficient this phase we will investigate about automatic elicitation tools.

3 Flashover Forecasting

To approach the flashover on high-voltage insulators forecasting, we developed and integrated four data mining tools that combine the ID3 algorithm [4] and the nearest neighbor case-based reasoning method [5]. The first tool uses data mining to build a classification or decision tree from historic data, the second generates production rules, the third operates the decision tree as an expert system, and the last, makes tests with unseen cases to evaluate forecasting accuracy. The results were compared against other

classic machine learning tools like C4.5, FOIL, CN2 and OC1, and we obtain similar or better solutions.

To perform the experiments, data from an N-120P high-voltage insulator were registered during 21 days (504 examples of meteorological and surface resistance values). The attributes used were: hour of the day, wind direction, wind velocity, temperature, precipitation, dew temperature, barometric pressure, relative humidity and absolute humidity. The class attribute is the *surface resistance*, a variable correlated with the flashover phenomenon.

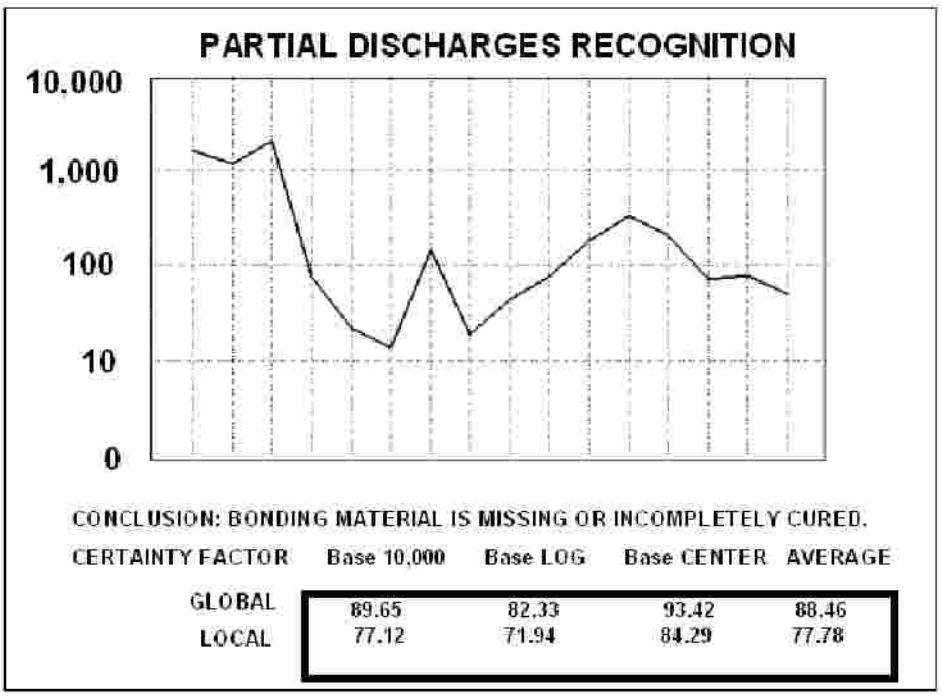

Fig.1 EGFD final display

Three classes were assigned to the surface resistance: "low" (for values between 1,013 and 3,489 kilo ohms), "medium" (for the interval 3,490/ 7,992) and "high" (for the interval 7,993/ 11,482).

With these tools, we could successfully determine:

- The relationship among the environmental variables and the surface resistance class.
- The variables that have more impact on the phenomenon, and the variables that hardly affect the surface resistance behavior.

- The circumstances that cause the surface resistance to be low.
- The surface resistance value 24 hours ahead, with an accuracy of 83%.

Table 1 shows comparison results. In the future we will work in the simplification and reduction of the production rule quantity generated for the proposed tool, to facilitate the interpretation of the discovered knowledge. We will work too with more data to try to forecast with more anticipation the flashover event.

Table 1. Results comparison.

TOOL	**Proposed**	C4.5	FOIL	CN2	OC1
TEST-Acc	82.9	82.9	62.9	82.4	80.1
TRAIN-Acc	99.6	85.3	96.8	92.0	81.5
# Rules	496	9	61	52	NA
Default	Unknown	Low	NA	High	NA
Time(secs)	3.68	143	10.7	74	643

where:
TEST-Acc = Accuracy to forecast unknown cases (%).
TRAIN-Acc = Accuracy to forecast the same examples used in learning stage (%).
# Rules = Number of generated rules.
Default = Default class (if no rule applies).
Time (secs) = Seconds required to generate the results.
NA = Characteristic not available by the tool.

4 Electric Utilities Data Mining

This case is about extract knowledge from data. Some well-known data mining tools (C4.5, CN2, FOIL and PEBLS) were applied and evaluated for the task to obtain expert knowledge. This was done on a real power generation database with thermoelectric and hydroelectric Mexican electric utilities information over eight years of historic data. We evaluated accuracy, knowledge amount reduction and processing time. We analyzed the expert system rules (extracted knowledge) and we propose an architecture of an integrated knowledge discovery system for this electric power generation database [6].

For this research, personnel of the Performance Control and Informatics Unit of Federal Commission of Electricity in Mexico (CFE) selected the data. One table was built with 32 variables and 1,110 records corresponding to thermo and hydroelectric information for the years 1988 to 1995.

The 32 variables include the following: power plant identifier, date, plate and effective capacity; unavailability by type of failure; outage equivalent hours due to decrements,

number of outages and outage hours due to failure and due to routine and corrective maintenance and other causes; fuel kilocalories; net and gross generation; permanent workers used in maintenance, in operation, and in other activities; additional workers used in maintenance, in operation, and other activities; equivalent substitution workers in maintenance, in operation, and in other activities; total personnel positions; accidents that cause lost of time; accidents in transit; days lost due to accidents; days lost due accidents in transit; sum of disabilities in percent; and various expenses.

With this data set, a supervised data mining knowledge extraction was outlined. We used the variable Power Plant Factor (PF), as the "class" or focus of attention for the experiment. To calculate the PF the following formula was used:

$$PF = \frac{\text{Gross Generation}}{\text{HP} * \text{Net Generation}} * 100 \quad (1)$$

where HP (hours per period) is equal to 8,760 hours (365 days by 24 hours). It was found that only one rule describes the knowledge for 'excellent' plant factor for hydroelectric utilities, with 85.7% certainty:

IF:
Unavailability due to failure (%) <= 9.375 and
Unavailability due to maintenance (%) <= 0.520 and
Unavailability due to other causes (%) <= 32.560 and
Permanent Workers (Rest) > 822
THEN: The Plant Factor is Excellent [certainty 85.7%]

Only one rule describes the knowledge for 'excellent' plant factor for thermoelectric utilities, with 92.3% certainty:

IF:
Effective Capacity (MW) <= 298 and
Gross Generation (GWH) > 1,721
THEN: The Plant Factor is Excellent [certainty 92.3%]

We found that the variables that most affect the hydroelectric plant factor turned out to be: Unavailability due to other causes (63%), Gross generation (59%), and Effective capacity (48%). For thermoelectric utilities the variables were: Gross generation (87%), Effective capacity (83%), and Unavailability due to failure (48%). A summary of the results is shown in Table 2. From the experience obtained in the development of the experiments described, the need of having a system to facilitate the process of knowledge discovery using data mining algorithms and the exploration of various alternatives that improve the quality of the extracted knowledge (expert system rules) was evident.

So we proposed the creation of the following data mining modules:

- User Interface: allows the user to have an integrated environment, which shows the user a screen from which he can choose different options to accomplish the data mining and to obtain the results.

Table 2. Comparison of the number of Errors*

Plant Factor	C4.5	C4.5*	CN2	FOIL	FOIL*
Very-Low	4	7	8	1	1
Low	37	83	125	3	14
Regular	31	54	79	3	23
Average	23	67	87	10	7
Good	28	42	61	1	9
Very-Good	12	29	39	6	7
Excellent	5	22	23	0	0
Total	140\13.5%	304\29.2%	422\40.6%	24\2.3%	61\5.8%

Errors* = Number of cases misclassified using unseen data
Commentaries: FOIL has the better classification efficiency, followed by C4.5
C4.5 = results of the 'composite rule set'
C4.5* = results of the 'trial 0'
FOIL* = using similar attributes grouping

EXECUTION TIMES:
C4.5 = more than 30 mins.
CN2 = 10 mins.
FOIL = 128.1 secs.
FOIL* = 189.6 secs.

- Pre-Processing: this module handles different options to prepare the information of the database before the application of the mining algorithm. This module allows, among other things, the addition or deletion of columns and rows, clustering (using several methods like ChiMerge, 1R, Chi2, etc.) of continuously valued variables to group them in (a few) labeled classes, feature selection methods and to automatically prepare the data to the format required by the mining tool.
- Mining tools: the user selects from among several data mining tools, the one to be applied to the preprocessed data. Usually, it is necessary to try different algorithms due to the fact that there does not exist a perfect tool, but rather, depending on the data, some algorithms perform better than others.
- Post-Processing: through this module, the user may request the conversion of the extracted knowledge by the mining tool in a representation that it will be easier for him to understand; again, it does not exist "the best" representation of knowledge, since it depends on the user preferences. Some knowledge representations are: production rules, decision trees, graphics (OLAP), characteristic tables (prime relation tables and feature tables), Horn clauses, and prototypes.

These ideas were proposed by us before tools like Weka, Orange, Elvira, and others, arrived to the data mining community. However, still nowadays several issues related with the proposed modules are open for research, like data quality tools (profiling, cleansing, etc.), knowledge representation and visualization tools, etc.

5 Illicit Users Pattern Recognition

To process this problem, we have to realized feature selection pre-processing task due to the database size and because of noise data problems. The problem domain can be expressed thus: CFE faces the problem to accurately detect customers that illicitly use energy, and consequently, CFE tries to reduce the losses due to this concept. At present time, a lot of historical information is stored in the Commercial System (SICOM), an electric billing database. SICOM was created mainly to register the users contract information, and the invoicing and collection data; this database has several years of operation and has a great amount of accumulated data (millions of records).

To make feasible the mining of this large database, in an effective and efficient way, we firstly realized an evaluation of different filter-ranking methods for supervised learning. The evaluation took into account not only the classification quality and the processing time obtained after the filter application of each ranking method, but also it considered the discovered knowledge size, which, the smaller, the easier to interpret.

Also the boundary selection topic to determine which attributes must be considered relevant and which irrelevant was approached, since the ranking methods by themselves do not give this information. We proposed an extension, simple to apply, that allows unifying the criterion for the attributes boundary in the different evaluated ranking methods [7].

Based on the experimentation results, we proposed a heuristic that looks for the efficient combination of ranking methods with the effectiveness of the wrapper methods. Although our work focuses on the SICOM data, the lessons learned can be applied to other real world databases with similar problems.

Recently, to process this problem more efficiently and accurately, we proposed several competitive metrics and algorithms for feature selection considering inter-dependencies among nominal attributes (*buBF* method) [8] or numeric attributes (*dG* method) [9]. Some results and comparisons against other feature selection methods in Weka [10] and Elvira [12] tools are shown in Table 3.

Table 3. J4.8´s accuracies for 10-fold-cross validation using the features selected by each method (Electric billing database).

Method	Total features selected	Accuracy (%)	Pre-processing time
CFS	1	90.18	9 secs.
dG	2	90.70	43 secs.
vG	3	94.02	0.7 secs.
Bhattacharyya	3	90.21	6 secs.
Matusita distance	3	90.21	5 secs.
ReliefF	4	93.89	14.3 mins.
Euclidean distance	4	93.89	5 secs.
Kullback-Leibler 1	4	90.10	6 secs.
Mutual Information	4	90.10	4 secs.
buBF	5	97.50	1.5 secs.
Kullback-Leibler 2	9	97.50	6 secs.
OneR	9	95.95	41 secs.
Shannon entropy	18	93.71	4 secs.
ChiSquared	20	97.18	9 secs.
All attributes	24	97.25	0

Furthermore, these ideas was applied successfully to other well known databases [12], as Table 4 shows. So we can conclude that the proposed metrics and feature selection methods are valuable tools to detect relevant attributes.

In the near future we will work in developed feature selection methods for mixed data, this is to say, for nominal and numeric attributes at the same time.

6 Conclusions and Future Work

We have presented four data mining applications in the Mexican power industry, and the way that we approached each one of them. From the experimentations presented we think that our proposed methods represents promising alternatives, compared to other methods, because of its acceptable performance. At Table 5 we resume our experiences: we outline advantages, drawbacks and possible solutions that we think it have to be done in the near future to maximizing the usefulness of the data mining techniques that we used in our works.

Table 4. J4.8´s accuracies using the features selected by each method for five UCI datasets.

Method	Autos (25/205/7)			Horse-c (27/368/2)			Hypothyroid (29/3772/4)			Sonar (60/208/2)			Ionosphere (34/351/2)			Avg. Acc
	TF	Ac	Pt	TF	Ac	Pt	TF	Ac	Pt	TF	Ac	Pt	TF	Ac	Pt	
All atts	25	*82*	0	27	*66*	0	29	*99*	0	60	*74*	0	34	*91*	0	***82.4***
buBF	9	*77*	0.2	4	*72*	0.22	5	*97*	0.31	10	*74*	0.6	4	*90*	0.9	***82.0***
vG	8	*75*	0.01	3	*69*	0.02	4	*95*	0.2	11	*73*	0.03	4	*91*	0.3	***80.6***
dG	7	*75*	12	2	*68*	14	5	*95*	26	9	*75*	14	3	*88*	18	***80.2***
CFS	6	*74*	0.05	2	*66*	0.04	2	*96*	0.3	18	*74*	0.09	8	*90*	3	***80.0***
ReliefF	11	*74*	0.4	3	*66*	0.9	6	*93*	95	4	*70*	0.9	6	*93*	4	***79.2***
SOAP	3	*73*	0.01	3	*66*	0.02	2	*95*	0.2	3	*70*	0.02	31	*90*	0.01	***78.8***
Mutual I	3	*72*	0.9	4	*68*	1	2	*90*	1.4	18	*73*	1	3	*86*	1	***77.8***
OneR	5	*70*	0.8	3	*67*	1	3	*88*	1.3	12	*72*	1	4	*85*	1	***76.4***
KL-1	3	*71*	0.9	4	*61*	1.2	3	*92*	1.7	16	*70*	1	2	*86*	1	***76.0***
KL-2	4	*68*	0.9	4	*62*	1.1	2	*89*	1.5	11	*68*	1	3	*83*	1	***74.0***
Matusita	3	*66*	1.7	3	*61*	2.3	2	*91*	3.3	17	*68*	2.5	2	*83*	2	***73.8***
Bhattac	3	*67*	0.8	3	*60*	1	1	*90*	1.4	9	*68*	1	2	*83*	1	***73.6***
Euclidea	2	*66*	1	3	*62*	1.4	2	*90*	1.2	10	*67*	1.1	2	*82*	1	***73.4***
ChiSqua	3	*67*	1	2	*60*	1.6	3	*88*	1.3	11	*65*	1.2	2	*80*	1	***72.0***
Shannon	4	*66*	0.9	4	*61*	1.3	2	*87*	1.6	9	*66*	1	2	*80*	1	***72.0***

"(25/205/7)" means (attributes/ instances/ classes) for Autos dataset, and so on.
TF=Total features selected Ac=Accuracy (%) Pt=Pre-processing time (secs.)

In our opinion, a great variety of Mexican power industry applications are still waiting to be tackled with data mining techniques, but we need develop more and sophisticated tools to accomplish the challenges. Some future work includes problems with real and very large power system databases such as the national power generation performance database, the national transmission energy control databases, the de-regulated energy market database, and the Mexican electric energy distribution database. Also, we need apply statistical tests to observe if the differences in accuracies, processing time or another parameter of the proposed methods are really significant.

Table 5. Our experiences and recommendations.

Approach used	Advantage	Drawback	Possible solution
Expert System	Representation of human-expert knowledge in a natural way.	Complex elicitation process.	Develop more sophisticated and computer aided elicitation tools.
Neural Network	Captures knowledge from numeric data.	It needs manual tuning. Discovered knowledge is in a black box.	Develop tools for dynamical tuning and to extract knowledge from neural inter-connections.
Induction Tree	Captures and shows knowledge from nominal data in an explicit way.	It needs previous data discretization. Obtained results are not very precise.	Develop tools for automatic and efficient data discretization. Improve output thru post-processing-visualization tools.
Data Mining	Discovers and shows hidden knowledge from data.	It needs an integration of the pre-processing, processing and post-processing phases.	Construct a integrated system with: data quality process, final user easy of interpret knowledge representation and visualization tools.
Feature Selection	Detects relevant attributes and reduces problem size.	There are no infallible method.	Research for metrics that evaluate attribute relevance (numeric and nominal data at once) in an effective way.

References

1. The 24th Annual International Conference on Machine Learning (ICML-2007) June 20-24, Oregon State University, USA, 2007.
2. Piatetsky-Shapiro, G. et al, Knowledge Discovery in Databases: An Overview, In Knowledge Discovery in Databases, Piatetsky-Shapiro, G. eds., Cambridge, MA, AAAI/MIT, 1991, pp 1-27.
3. Mejía, M., Rodríguez, G., A New Neural Network Paradigm for Power Systems Applications, Proceedings of the IASTED International Conference on Power Systems and Engineering, Vancouver, Canada 1992, pp. 41-48.
4. Quinlan, J., Discovering Rules by Induction from Large Collections of Examples, Expert Systems in the Micro-Electronic Age, Michie, D., (ed), Edinburgo, Escocia, Edinburgh University Press, 1979.

5. Mejía, M., Rodríguez, G., Montoya, G., Knowledge discovery in high-voltage insulators data, Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, Proceedings of the Tenth International Conference, Atlanta, Georgia, USA, June 1997, pp. 223-230.
6. Mejía, M., Rodríguez, G., Obtaining expert systems rules using data mining tools from a power generation database, Expert Systems with Applications, J.Liebowitz (ed), 14(1/2) Pergamon, 1998, pp. 37-42.
7. Mejía, M., Rodríguez, G., Arroyo, G., Morales, Feature selection-ranking methods in a very large electric database. *MICAI 2004: Advances in Artificial Intelligence, 3rd Mexican Int. Conf. on Artificial Intelligence*, Springer Berlin, April, pp. 292-301.
8. Mejía, M., Morales, E. 2006. Feature Selection in an Electric Billing Database Considering Attribute Inter-dependencies. In Petra Perner (ed) *Advances in Data Mining. Applications in Medicine, Web Mining, Marketing, Image and Signal Mining: 6th Industrial Conference on Data Mining*, ISBN: 3-540-36036-0, ISSN: 0302-9743, LNCS 4065, Springer Berlin / Heidelberg, Leipzig, Germany, pp. 284-296.
9. Mejía, M., Morales, E. 2007. Two Two Simple and Effective Feature Selection Methods for Continuous Attributes with Discrete Multi-Class. *MICAI 2007 6th Mexican Int. Conf. on Artificial Intelligence*, LNAI 4827, Springer Berlin, November, pp. 452-461.
10. www. cs.waikato.ac.nz/ml/weka, 2004.
11. www. ia.uned.es/~elvira/ , 2004.
12. Newman, D.J. & Hettich, S. & Blake, C.L. & Merz, C.J. UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science, 1998.