Innovative Applications of Diagnosis, Forecasting, Pattern Recognition and Knowledge Discovery in Power Systems

Manuel Mejía-Lavalle Gustavo Arroyo-Figueroa Eduardo F. Morales[1]
Instituto de Investigaciones Eléctricas
Reforma 113, Palmira 62490 Cuernavaca, Morelos, México
[1]INAOE L.E. Erro 1, 72840, StMa. Tonantzintla, Puebla, México
{mlavalle, garroyo}@iie.org.mx, emorales@inaoep.mx

ABSTRACT
We present our experiences in five power systems domains where we applied diverse knowledge discovery - 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 next case approaches a pattern recognition problem to detect potential electric illicit users. The last case presents a fossil fuel power plant diagnosis system based on temporal probabilistic networks. We outline successful and bad practices, our contributions, and comment about possible solutions for future work that we think it has to be done to maximize the usefulness of knowledge discovery in the power industry.

Categories and Subject Descriptors
H.2.8 [**Database Applications**]: Data mining – *electric power industry*

General Terms
Algorithms

Keywords
Data mining, experiences, industrial applications, diagnosis, forecasting, fraud detection

1. INTRODUCTION
Knowledge discovery (KD) and data mining has been employed with success in various fields and in many real world problems [1][2]. Data mining is applied to huge volumes of historical data mainly with the expectation of finding knowledge, trends or behavior patterns that permit to 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 [3]. Some of the traditionally used computer techniques to accomplish data mining are: neural networks, induction of decision trees, production rules, feature selection and case-based methods.

In this paper we present our experiences in several Mexican power electric industry domains where we applied diverse knowledge discovery 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 next case approaches a pattern recognition problem to detect potential electric illicit users, where we proposed and realized a pre-processing feature selection method. The last case presents a novel fossil fuel power plant diagnosis system based on temporal probabilistic networks. We outline successful and bad practices, and comment about possible solutions for future work that we think it has to be done to maximize the use 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 expert knowledge is coded as production rules as the following:

978-1-4244-4241-6/09/$25.00 ©2009 IEEE

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's 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 of the 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 [4]. 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) 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) 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) 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 plan, as future work, to incorporate more human and data knowledge to this system. To facilitate this phase we will investigate about automatic elicitation tools.

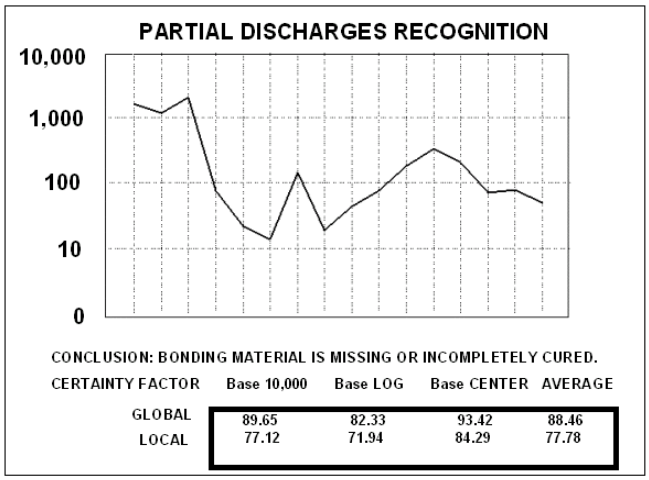

Figure 1. EGFD final display.

3. FLASHOVER FORECASTING

To approach the flashover on high-voltage insulators forecasting, we developed and integrated four knowledge discovery - data mining tools that combine the ID3 algorithm [5] and the nearest neighbor case-based reasoning method [6]. The first tool builds a decision tree from historic data, the second generates production rules, the third uses 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. 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 some comparative results. In the future we will work on the simplification and reduction of the number of production rules generated 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. Test results

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

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	3.68	143	10.7	74	643

4. ELECTRIC UTILITIES KD

This case is about extracting knowledge from data. Some well-known data mining tools (C4.5, CN2, FOIL and PEBLS) were applied and evaluated on 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, the size of the generated knowledge 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 [7]. 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 8 years of thermo and hydroelectric information.

The 32 variables included: power plant identifier, date, installed 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. The PF was evaluated using the following formula:

PF = (Gross Generation / HP * Net Generation) * 100

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 have more influence in 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, it became evident 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). So we proposed the creation of the following knowledge discovery modules:

Table 2. Comparison of the number of Errors*

Plant Factor	C4.5	C4.5*	CN2	FOIL	FOIL*
VeryLow	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
VeryGood	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 (test 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.

- 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.
- 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 tools to automatically prepare the data to the format required by the mining application.
- Mining tools: the user selects from several data mining tools, one to be applied to the preprocessed data. Usually, it is necessary to try different algorithms since there is no algorithm that performs better than other algorithms in all the domains.
- Post-Processing: through this module, the user may request the conversion of the extracted knowledge by the mining tool in a representation that is easier 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, characteristic tables (prime relation tables and feature tables), Horn clauses, and prototypes.

We proposed these ideas before tools such as Weka, Orange, Elvira, and others, arrived to the KD 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, among others.

5. ILLICIT PATTERN RECOGNITION

CFE faces the problem to accurately detect customers that illicitly use energy. There exist a large volume of historical information (millions of records) stored in the Commercial System (SICOM), an electric billing database with several years of operation. SICOM was created mainly to register the users contract information and invoicing data. In this case, due to the size of the database, we decided to perform a feature selection pre-processing stage.

To make feasible the mining of this large database, in an effective and efficient way, we first evaluate 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 it also considered the discovered knowledge size, assuming that the smaller is easier to interpret.

Also the boundary selection criteria 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 a simple extension that allows unifying the criterion for the attributes boundary in the different evaluated ranking methods [8].

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) [9] or numeric attributes (*dG* method) [10]. Some results and comparisons against other feature selection methods in Weka [11] 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.
ReliefF	4	93.89	14.3 mins.
Kullback-Leibler 1	4	90.10	6 secs.
Mutual Information	4	90.10	4 secs.
buBF	5	97.50	1.5 sec
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

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*	1 2	2	*68*	1 4	5	*95*	2 6	9	*75*	1 4	3	*88*	1 8	***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*	9 5	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 Informa.	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***
Matusit	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***
Euclide	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.)

Furthermore, these ideas were successfully applied to other well-known databases [13], 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 of feature selection methods for mixed data, this is to say, for nominal and numeric attributes at the same time.

6. POWER PLANT DIAGNOSIS

This section presents a fossil fuel power plant diagnosis system based on temporal probabilistic networks. The diagnosis system uses a cause-consequence analysis based on a model to deal uncertainty and time. The model is called as Temporal Event Bayesian Network (TEBN) [14]. A TEBN is a Directed Acyclic Graph (DAG) when a node represents an event or disturbance and each edge represents a causal-temporal relationship between temporal nodes. A TEBN is a formal representation of causal and temporal relationships between events. In other words, a TEBN is a model of a system or plant behavior disturbances that can be compared to actual plant response. A TEBN is defined as follows:

> Definition 1. A Temporal Event Bayesian Network is defined as TEBN=(V, E), where V is the set of temporal nodes and E is the set of edges. Each temporal node is defined by an ordered pair (ς, τ) and the conditional probability matrix that specifies the probability of each ordered pair given its parents.

The values of each temporal node can be seen as the "cross product" between the set of values (Σ) and the set of time intervals (T), except for the default state, which is associated to only one interval.

The cause-consequence analysis is used to indicate that the primary emphasis is on determining causes of events that have occurred, as well as predicting future consequences. The TEBN formalism includes dealing with uncertainty and time for the diagnosis of disturbances and prediction of events, using the signals as important pieces of information. The analysis stars when the event detection module detects an event. This module takes the readings of the sensor and compares its value with the low and high limits. A signal whose values are out of its specified limits is said to be an event.

The inference mechanism of a TEBN model is based on the detection of events and the propagation of evidence. The inference mechanism updates the marginal posterior probabilities of each node (variable) of the network given the occurrence of an event or events. We define ***tc*** as the time when an event is detected and α as the *real time occurrence function*, this function is define as the absolute value of the difference between the time of occurrence of a pair of connected events. As the net does not have any temporal reference, the time of occurrence of the first event fixes temporally the network. The value of α is used to determinate the time interval of the "effect" node considering the "cause" node as initial event. Afterwards, the evidence is propagated through the network to update the probabilities of the other nodes. These probabilities show the potentially occurrence of the past and future events. The stop condition is when a terminal or leaf node is reached.

We can identify three main steps in the inference mechanism procedure:

1. Detection of the event or events occurrence and definition of the time interval of occurrence of the event(s);
2. Propagation of the evidence occurrence through the net and update of the probabilities of the variables;
3. Determination of the potential past and future events.

A TEBN is built based on the human expertise and a process knowledge database. The human expert gives the causal relationships between the events when a disturbance occurred. The distributed control system gives the historical relevant information about the evolution of the disturbances.

The diagnosis system has been applied for fault diagnosis and prediction in a steam generator of a fossil power plant. We consider the drum level control system with three potential disturbances: a power load increase (LI); a feed water pump failure (FWPF); and a feed water valve failure (FWVF). The drum is a subsystem of a fossil power plant that provides steam to the super-heater and water to the water wall of a steam generator. The drum system is composed by three systems: the feed water, the water steam generator and the super-heater steam. One of the main problems in the drum is to maintain its level in safe operation. Fig. 2 shows a TEBN that represents the events of the drum system of a steam generator power plant. The network structure was defined based on the knowledge of an expert operator. The definition of the time intervals for each temporal node was obtained based on knowledge about the process dynamics. Once the structure and time intervals were defined, the required parameters were estimated from process data. A full-scale simulator of a 350 MW thermal power plant generated the process data.

In the process, a signal exceeding its specified limit of normal functioning is called an "*event*", and sequences of events that have the same underlying cause are considered as a "*disturbance*". To determine which of the disturbances is present is a complicated task, because there are similar sequences of events for the four main disturbances. We need additional information in order to determine what the real cause is. In particular, the temporal information about the occurrence of each event is important for an accurate diagnosis. For example, the feed water flow increase (FWF) can be caused by two different events: the feed water pump augmentation (FWP) and a feed water valve opening increase (FWV). We can use the time difference between the occurrence of each event, FWV-FWF and FWP-FWF, for selecting the "cause" of the increase of the FWF.

According to the process data, the time interval between a pump current augmentation and an increase of the flow (FWP->FWF) is between 25 to 114 seconds. The time interval between the valve opening increase and an increase of the flow (FWV->FWF) is between 114 to 248 seconds. Hence, if the flow increase occurs in the first time interval, the probable cause is an augmentation of the pump; but if the flow increase occurs in the second time interval, the probable cause is a valve opening increase.

For evaluation purposes, we selected 80% of this database (800 records) for parameter learning and 20% (200 records) for evaluation. The network was evaluated using two scores: % of accuracy and % of Relative Brier Score (RBS) total square error. The % of accuracy was evaluated by number of correct predictions of unknown variables of the network. The Brier Score was defined as: $BS = \Sigma^{n}_{i=1} (1 - P_i)^2$. Where P_i is the marginal posterior probability of the correct value of each node given the evidence. The maximum Brier Score is: $BS_{MAX} = \Sigma^{n} (1)^2$. The % of RBS was defined as: $RBS\ (\text{in } \%) = \{1 - (BS / BS_{MAX})\} \times 100$.

The test methodology includes three basic steps: (i) Assign a value to a subset of nodes, (ii) propagate the evidence and (iii) compare the posterior probabilities of the nodes with the actual values. The assigned nodes were selected for 3 sets of tests: (1) *Prediction*: root nodes are observed (LI, FWPF, FWVF and SWVF); (2) *Diagnosis*: leaf nodes are observed (STT, STF and SWF); and (3) *Prediction and diagnosis*: intermediate nodes are observed (STV, FWP, FWV and SWV).

Table 5 shows the results of the evaluation for the three sets of tests in terms of the mean (μ) and the standard deviation (σ) for both scores. These results show the prediction and diagnosis capacity of the temporal model in a real process. Both scores are between 80 and 97% for all the set of tests, with better results when intermediate nodes are observed, and slightly better results for prediction compared to diagnosis. We consider that these differences have to do with the "distance" between assigned and unknown nodes, and with the way that the temporal intervals were defined. We are encouraged by the fact that the model can produce a reasonable accuracy in times that are compatible with real time decision-making.

Table 5. Empirical evaluation results

Parameter	μ	σ
Prediction		
% of RBS	87.37	9.19
% of Accuracy	84.48	14.98
Diagnosis		
% of RBS	84.25	8.09
% of Accuracy	80.00	11.85
Diagnosis and Prediction		
% of RBS	95.85	4.71
% of Accuracy	94.92	8.59

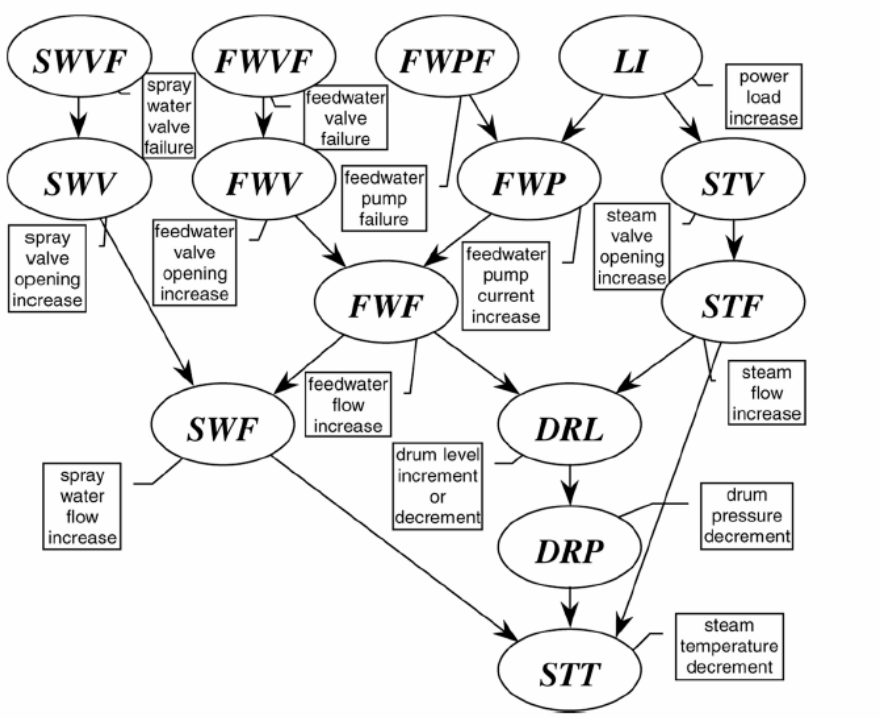

Figure 2. TEBN for the steam-drum system.

We development a tool based on TEBN approach [15]. The selected disturbance was first simulated on a fossil power plant training simulator (MICROTERM 300). During these simulations, the system is actively performing its analysis and the relevant real-time performance is recorded. At the same time, all the data from the simulator that the diagnosis system needs for its analysis are sent. The analysis begins when an event is detected. For all the modeled failures, the general arrangement of the messages is as follows:

- The first message contains a description of the particular event that was detected. It shows the plant status through the relevant variables, indicating for each one its value, status and tendency.
- The second message is activated by operator request or when the steam generator has reached a critical alarm condition. This message informs the probable disturbance, the suggested recovery and the predicted consequences.

Fig. 3 shows the user interface for the detection of an increase in the feedwater flow.

- The first screen shows the event detection and its related variables. For this case, a *feedwater flow increase* (FWF) is detected at 4:38:00. The related variables are *feedwater valve* (FWV), *feedwater pump* (FWO) *and drum level condition* (DHL).
- The second screen shows the probabilistic analysis: the probability of the cause (diagnosis) and the probability of the occurrence of the future events (prediction). According with the process data the FW flow occurs in the first time interval. Hence, the most probable disturbance is the FW pump failure (the velocity of the pump tends to maximum) with a probability value (certainty) of 0.75. The recommended actions are (1) Run the stand bye FW pump and (2) Repair the FW pump. The probable past events are a FW pump increase, 25 to 114 seconds before the FW flow event, with a probability of 0.61; and a FW valve increase, 114-248 seconds before the FW flow event, with a probability of 0.32. The probable future event is a high level condition 10 to 27 seconds after the FW flow event detection with a probability of 0.95.

7. CONCLUSIONS

We have presented five knowledge discovery 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. Table 6 summarizes our experiences: we outline advantages, contributions, drawbacks and possible solutions that we think it have to be done in the near future to maximize the usefulness of the data mining techniques.

In our opinion, a great variety of Mexican power industry applications are still waiting to be tackled with the knowledge discovery approaches, but we need to develop more 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.

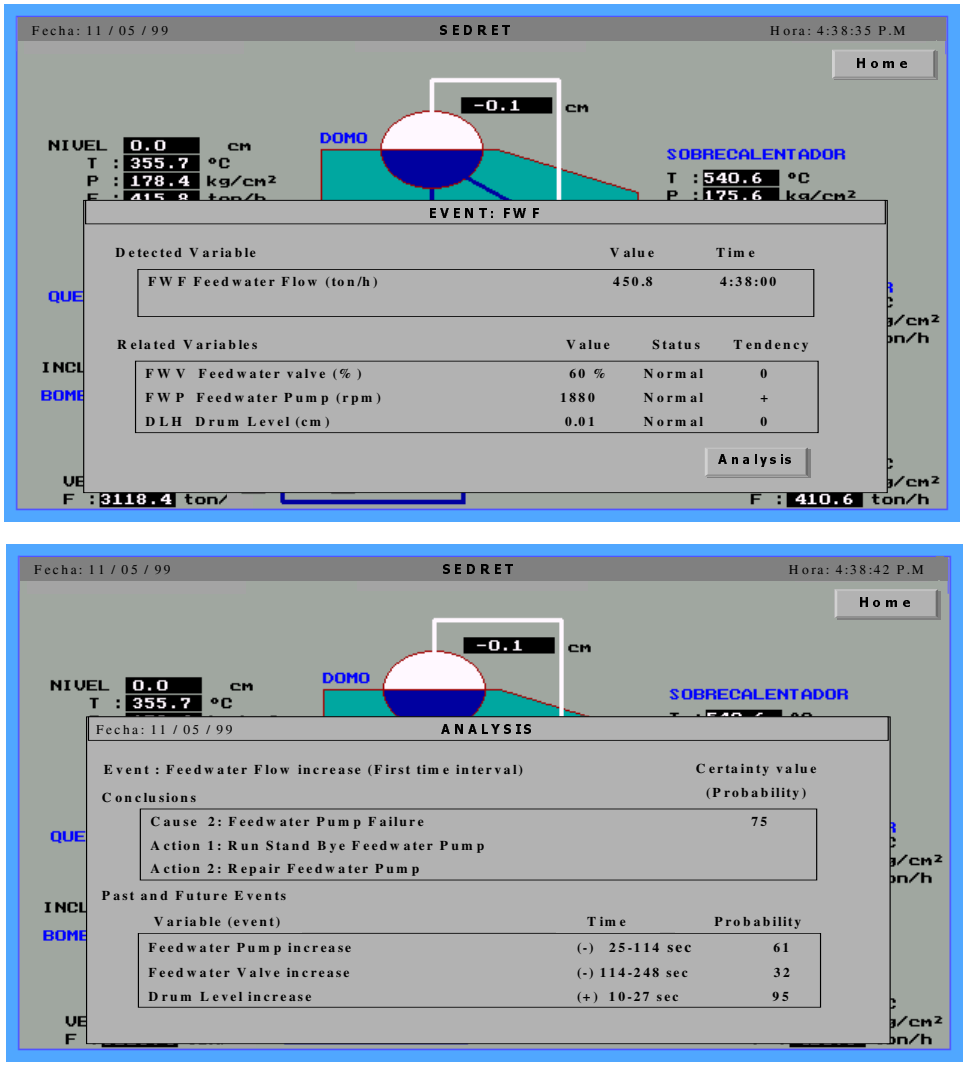

Figure 3. Interfaces of the system: (a) Event detected and its related variables, (b) Diagnosis of the disturbance and prediction of future events.

8. REFERENCES

[1] Rayid Ghani, Carlos Soares: Data mining for business applications: KDD-2006 workshop. SIGKDD Explorations 8(2): 79-81 (2006).

[2] Proceedings of the IEEE Int.Conf. on Data Minining, Wu, X. (Ed), Pisa, Italy, 2008.

[3] 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.

[4] 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

[5] 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.

[6] 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.

[7] 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.

[8] Mejía, M., Rodríguez, G., Arroyo, G., Morales, E. 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.

Table 6. Our experiences and recommendations

Approach used	Advantage	Our contribution	Drawback	Possible solution
Expert System	Representation of human-expert knowledge in a natural way.	Electric Generator Failure Diagnosis Expert System.	Complex elicitation process.	Develop more sophisticated and computer aided elicitation tools.
Neural Network	Captures knowledge from numeric data.	PHAF II Paradigm.	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.	Tools that combine the ID3 algorithm and the Nearest Neighbor Case-Based Reasoning method .	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.	Tool conceptualization that combine and integrate a user interface, pre-processing, mining and post-processing facilities.	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.	Metrics and algorithms considering inter-dependencies among attributes.	There is no infallible method.	Research for metrics that evaluate attribute relevance (numeric and nominal data at once) in an effective way.
Bayesian Network	Models that deal with uncertainty and time.	A temporal event bayesian network model (TEBN).	Difficult to scale it for real fossil fuel power plants.	Developing a network structure automatic learning mechanism based on process data.

[9] 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.

[10] 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.

[11] www. cs.waikato.ac.nz/ml/weka, 2004.

[12] www. ia.uned.es/~elvira/ , 2004.

[13] 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.Ding, W. and Marchionini, G. 1997 A Study on Video Browsing Strategies. Technical Report. University of Maryland at College Park.

[14] Arroyo-Figueroa. G., Sucar L. E., Temporal Bayesian Network for diagnosis and prediction. *In Laskey K., Prade H.: Proc. 15th Conference on Uncertainty on Artificial Intelligence*, 1999, pp. 13-20.

[15] Arroyo-Figueroa G., Alvarez Y., Sucar L. E., "SEDRET – an intelligent system for the diagnosis and prediction of events in power plants", *Expert Systems with Applications*, vol. 18, No. 2, 2000, pp. 75-86.