TRY TO PREDICT GRID FAULTS: A DYNAMIC, SEMANTIC-BASED AND MULTI-DIMENSIONAL APPROACH

Michele RUTA
Politecnico di Bari - Italy
michele.ruta@poliba.it

Giuseppe LOSETO
Politecnico di Bari - Italy
giuseppe.loseto@poliba.it

Simone TEGAS
Enel Distribuzione S.p.A.- Italy
simone.tegas@enel.com

Gianpatrizio BIANCO
Enel Distribuzione S.p.A.- Italy
gianpatrizio.bianco@enel.com

Floriano SCIOSCIA
Politecnico di Bari - Italy
floriano.scioscia@poliba.it

Eugenio DI SCIASCIO
Politecnico di Bari - Italy
eugenio.disciascio@poliba.it

ABSTRACT

The analysis of big data volumes, produced by energy distributors and concerning the operation of the grid, requires several different techniques, particularly if one would like to use such large amount of information in possible grid failure assessment and even prevention. A preliminary data volumes reduction is strongly needed: currently grid data need huge storage and processing resources for long-term analysis. This paper presents a data aggregation statistical framework allowing to compact grid data sets in order to enable further semantic-based decision support (grounded on logic-oriented reasoning technologies) and devoted to try to predict failures, to suggest targeted maintenance and to possibly optimize grid assets.

1. INTRODUCTION AND BACKGROUND

Energy distributors continuously collect telemetry of several operational parameters of the grid in normal working conditions. Furthermore, in case of interruptions, data about the event are gathered in inspections executed during repair. Hence, large data amounts are produced every day with associated storage and management costs whose relevance convinces in more and more maximizing their value exploiting novel analysis methods able to suggest preventive maintenance and to possibly optimize grid assets.

Nevertheless, each possible information mining is restrained by limits in the various steps of data life cycle: (i) manual data *collection* procedures may lack completeness and accuracy; (ii) *storage* may not adopt state-of-the-art technologies and best practices to reduce data inconsistency as much as possible; (iii) *knowledge discovery* requires specialized approaches and trained professionals. As the pioneering experience on the New York City grid revealed [1], knowledge discovery from grid distribution data has many challenges:

- the state of the grid at the time of past failures is needed to train the predictive system, but taking accurate snapshots of past states is difficult due to variations in the database and/or the physical components;
- the grid infrastructure contains a large number of components, belonging to several types and varying in manufacturer, type, age; hence significant features must be computed by means of statistics;
- dynamic data –both grid-related (*e.g.*, sampled voltage and current values) and contextual (*e.g.*, weather, time of day)– must be aggregated over time, but acquisition rates vary significantly according to data type; time windows must be chosen carefully, possibly exploiting inherent system periodicity;
- grid components are susceptible to several types of failures, but available data is imbalanced and for many classes the training samples may be too scarce to extract regularities with robust generalization properties.

For the above reasons, the early data cleaning and aggregation steps are the most critical stages of the whole knowledge discovery process and the ones where close collaboration among analysts and domain experts is strongly needed. Data aggregation and cleaning helps (i) reduce data volumes in order to decrease storage capacity costs, (ii) build event features and labels for the subsequent classification and inference steps [1].

Data preprocessing may require several rounds of information refinement, and the final result may not even be satisfactory from a prediction effectiveness standpoint. Nevertheless, acknowledging issues in data gathering and storage, such an activity is yet relevant for the distributor, as it allows improving internal information management processes. This paper reports on an ongoing collaboration, supported by the funded national project RES NOVAE (PONREC 2007-2013), between *Enel Distribuzione S.p.A.* Distributor System Operator (DSO) and the *Information Systems laboratory* of Politecnico di Bari, aimed at defining a knowledge discovery framework that could allow to predict failures and outages in the grid. Such an approach aims to exploit novel and not evident correlations between grid information data already available from the DSO side. The proposed framework is being developed using a dynamic, semantic-based and multidimensional approach to characterize events and perform statistical and logical inferences, in order to derive implicit failures information and identify the grid components most prone to malfunction. The final goal is to allow distributors to switch from a reactive to a proactive grid maintenance approach, pursuing resource allocation rationalization and cost savings.

The remainder of the paper is organized as follows. In Section 2 the proposed knowledge discovery framework is outlined. Section 3 describes a case study and provides early experimental results. Conclusion and future work close the paper.

2. PROPOSED FRAMEWORK

A semantic-based data mining approach has been adopted for the dynamic analysis of some Low- and Medium-Voltage (LV, MV) sections of the Italian grid. The infrastructure portions have been selected by taking into account the presence of a considerable number of elements useful to characterize a generic distribution network. A Java-based prototype tool has been so developed for information aggregation, processing, analysis and visualization. It allows distributor's grid operators to monitor data series and gain insight into conditions leading to failures. Grid numerical data are aggregated and analyzed to infer further knowledge by means of Machine Learning (ML) techniques. Design and evaluation of the proposed approach is being conducted on a dataset provided by Enel Distribuzione, consisting of 27 months of observations for 126 MV lines (3402 total tables). Have been also provided data about failures happened in 29 months (during 2011, 2012 up to May 2013; 27 of them are the same of the MV lines observations). The analyzed dataset comprises 2438 interruptions on LV lines and 659 on MV ones, most of them are transient.

Figure 1 sketches the general activity diagram of the proposed framework. At the first stage data related to current variations on MV lines have been aggregated to reduce the amount of raw observations by means of: (i) statistical indexes; (ii) extension of the sampling period; (iii) aggregation of similar observations.

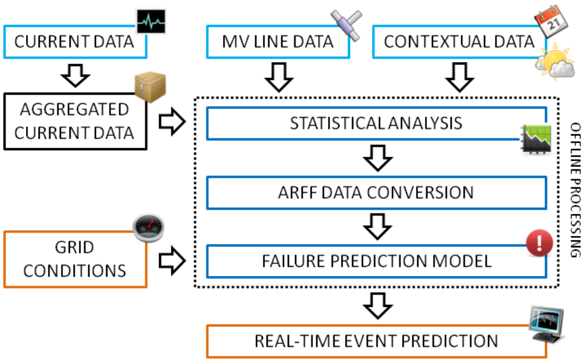


Figure 1. Activity diagram of the proposed framework

Absolute and relative indexes of position and dispersion (average, standard deviation, relative standard deviation, relative delta between values) have been computed exploiting a sliding window of 30 minutes (default period is 10 minutes). In this way, sudden changes in grid behavior (see for example current variations in blue circles in Figure 2) become particularly evident. Moreover, similar data (*i.e.*, data with same average and statistical indexes) are aggregated to further reduce the amount of needed space (yellow box in Figure 2). The aggregated data will replace redundant values in each of 3402 tables so that compacted files could be exploited in the long-term processing devoted to extract failure prediction models. The proposed tool includes a prognostic module able to possibly forecast five basic grid failures (structural failure, mechanical failure, breakage, water infiltration, generic failure). It starts from data about:

- structure of MV lines, related to composition of a power line in terms of constitutive properties such as length, nature and covering materials, number of connected LV and MV clients;
- contextual parameters, *e.g.* date, time and weather conditions referred to past observed events happened on the power grid;
- current variations in four different observation periods (one hour, one day, one week, one month) before the happened failure. Also in this case, by means of statistical analyses, 11 absolute and relative indexes are extracted for each period, as reported in Table I.

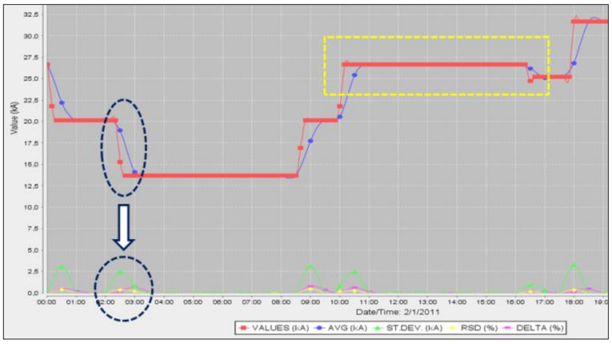
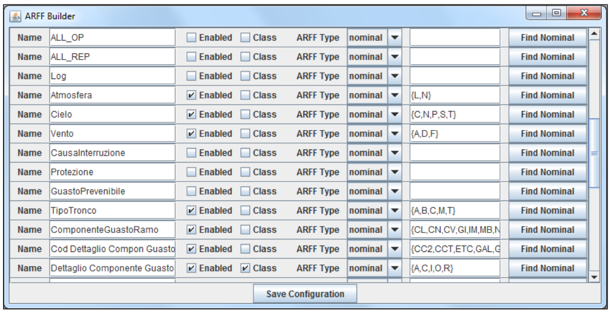


Figure 2. Data analysis tool – Visualization of current trend for a single MV line

#	Feature Description
1	Average
2	Absolute variance (on average) between two consecutive value
3	Relative variance (on average) between two consecutive value
4	Absolute Max value
5	Relative Max value (with respect to the average)
6	Absolute Min value
7	Relative Min value (with respect to the average)
8	Standard Deviation
9	Relative Standard Deviation
10	Num. of values out of IQR
11	Num. of values out of IQR / Total num. of values

Table I. Statistical indexes used to characterize current variations

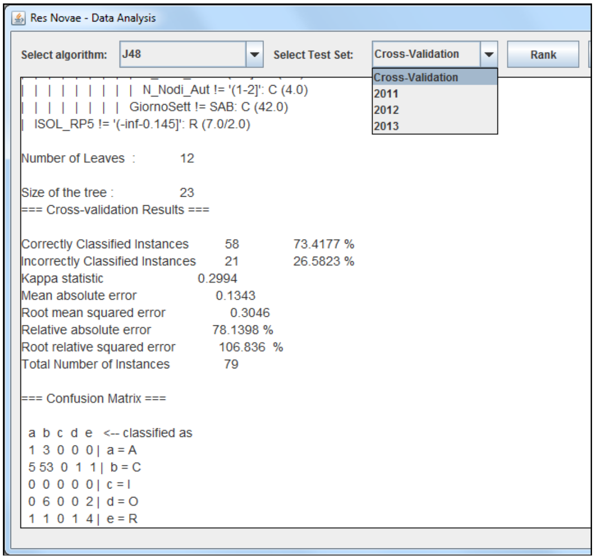
All the above parameters are exploited to learn associative rules and build a decision tree classifier to detect possible grid. The classifier has been implemented using *WEKA* machine learning toolkit [2] and requires data in ARFF (Attribute Relationship File Format) as input. DSO data (usually stored as comma-separated values) was converted to ARFF by means of a simple GUI (shown in Figure 3) enabling operators to export the features of interest from the reference dataset.



Name	ALL_OP	□ Enabled	□ Class	ARFF Type	nominal		Find Nominal
Name	ALL_REP	□ Enabled	□ Class	ARFF Type	nominal		Find Nominal
Name	Log	□ Enabled	□ Class	ARFF Type	nominal		Find Nominal
Name	Atmosfera	☑ Enabled	□ Class	ARFF Type	nominal	{L,N}	Find Nominal
Name	Cielo	☑ Enabled	□ Class	ARFF Type	nominal	{C,N,P,S,T}	Find Nominal
Name	Vento	☑ Enabled	□ Class	ARFF Type	nominal	{A,D,F}	Find Nominal
Name	CausaInterruzione	□ Enabled	□ Class	ARFF Type	nominal		Find Nominal
Name	Protezione	□ Enabled	□ Class	ARFF Type	nominal		Find Nominal
Name	GuastoPrevenibile	□ Enabled	□ Class	ARFF Type	nominal		Find Nominal
Name	TipoTronco	☑ Enabled	□ Class	ARFF Type	nominal	{A,B,C,M,T}	Find Nominal
Name	ComponenteGuastoRamo	☑ Enabled	□ Class	ARFF Type	nominal	{CL,CN,CV,GI,IM,MB,N	Find Nominal
Name	Cod Dettaglio Compon Guasto	☑ Enabled	□ Class	ARFF Type	nominal	{CC2,CCT,ETC,GAL,G	Find Nominal
Name	Dettaglio Componente Guasto	☑ Enabled	☑ Class	ARFF Type	nominal	{A,C,I,O,R}	Find Nominal

Figure 3. ARFF conversion tool – Panel for attributes selection and conversion

Different classification algorithms could be used for the prediction model: J48 (a Java implementation of C4.5 decision tree [3]); Random tree [4]; Best-First decision tree [5]; Functional Tree [6]. The prognostic process allows to select the most useful one (with related parameters) and then to identify the training and validation data (see Figure 4). Validation is related to the failures collected during a period subsequent to the one used as training.



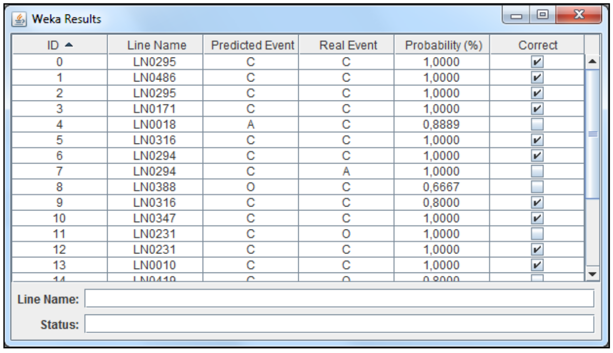
Select algorithm: J48
Select Test Set: Cross-Validation

```
| | | | | | | | |  N_Nodi_Aut != '(1-2]': C (4.0)
| | | | | | | |  GiornoSett != SAB: C (42.0)
|  ISOL_RP5 != '(-inf-0.145]': R (7.0/2.0)

Number of Leaves  :       12

Size of the tree :        23
=== Cross-validation Results ===

Correctly Classified Instances       58          73.4177 %
Incorrectly Classified Instances     21          26.5823 %
Kappa statistic                    0.2994
Mean absolute error                  0.1343
Root mean squared error                0.3046
Relative absolute error             78.1398 %
Root relative squared error          106.836  %
Total Number of Instances            79

=== Confusion Matrix ===

 a  b  c  d  e   <-- classified as
 1  3  0  0  0 |  a = A
 5 53  0  1  1 |  b = C
 0  0  0  0  0 |  c = I
 0  6  0  0  2 |  d = O
 1  1  0  1  4 |  e = R
```

Figure 4. Data Mining tool – Visualization of classification results

To prove the accuracy of the classifier, it is also possible to start a 10-fold cross-validation. In this case, the classifier is used to predict failures on the same year. The reference dataset is divided in ten pieces, nine of them are used for training whereas the last set is used for testing. The whole process is repeated ten times, using a different segment for testing each time and finally the average in terms of accuracy of the ten results is given.

As reported in Figure 4, for each classification process, the tool returns the following results: (i) accuracy of the prediction model; (ii) confusion matrix; (iii) tree-based prediction model in textual or graphical form. Finally, through the panel in Figure 5, it is also possible to see the list of classified instances with the probability of the predicted value. In this way, operators can filter the output to highlight prediction for specific MV lines.



ID ▲	Line Name	Predicted Event	Real Event	Probability (%)	Correct
0	LN0295	C	C	1,0000	☑
1	LN0486	C	C	1,0000	☑
2	LN0295	C	C	1,0000	☑
3	LN0171	C	C	1,0000	☑
4	LN0018	A	C	0,8889	□
5	LN0316	C	C	1,0000	☑
6	LN0294	C	C	1,0000	☑
7	LN0294	C	A	1,0000	□
8	LN0388	O	C	0,6667	□
9	LN0316	C	C	0,8000	☑
10	LN0347	C	C	1,0000	☑
11	LN0231	C	O	1,0000	□
12	LN0231	C	C	1,0000	☑
13	LN0010	C	C	1,0000	☑
14	LN0410	C	O	0,8000	□

Line Name:
Status:

Figure 5. Data Mining tool – Visualization of classified instances

Note that the obtained prediction model outputs data in terms of labeled features. Hence, it is difficult to understand by users without experience in data mining. To overcome this limit, an ontology-based approach for a semantic characterization of decision tree model rules has been proposed. Each failure prediction rule has been annotated in a logic-based formalism. Each condition composing a rule was described through a machine-understandable and user-friendly conjunct referred to a domain ontology developed for the power grid domain and not reported here for the sake of brevity. This will enable a multi-dimensional and semantic-based characterization of relevant grid parameters and conditions for decision support to grid management. Such annotations could be further exploited in different knowledge-based applications as in general-purpose Decision Support Systems (DSSs), where a semantic-based approach allows to perform implicit deductions about events also in case of real-world non-exact matching situations [7].

3. CASE STUDY

In order to clarify the proposed approach and show its benefits, a case study about failure prediction on MV lines is reported hereafter. The following real working scenario is considered as an example. *Operators of an electricity distribution company need to analyze grid data to prevent possible failures on MV lines. They would like to identify the most probable causes and conditions of grid malfunctioning in order to plan preventive maintenance aiming to reduce grid outages and improve the quality of service*.

The operator selects the features of interest used as input along with the reference years used as training set and test set. In the proposed example the first nine months of 2012 were used to predict grid failures in the following eight months. The mining tool extracts the data from the whole database, also rebuilding the aggregated current values if needed. Hence, one of the four decision tree algorithm cited in Section 2 is applied. In particular a binary and unpruned J48 classifier is selected. In order to reduce the time needed to train the algorithm and create the reference model, the early 55 input features of grid lines are ranked according to the *Gain Ratio (GR)*. GR enables to select most important characteristics of a line evaluating the worth of them with respect to the predicted attribute. Attributes with GR=0 will be discarded. Most relevant features are shown in Table II. The classifier was trained using only the 42 attributes with a meaningful GR.

#	GR	Feature Description
1	0,4783	Length of line section with rigid RP5-type isolation
2	0,4222	Num. of transformation points on utility pole
3	0,3661	Section type of a MV line
4	0,3075	Sub-type of component belonging to a MV line
5	0,2417	Available power (kW) for MV clients in middle concentration areas
6	0,2163	Type of component belonging to a MV line
7	0,1823	Length of line section with S-type isolation
8	0,1742	Absolute variance (on average) between two consecutive current value
9	0,1675	Type of installation
10	0,1590	Length of naked aerial line section

Table II. Top 10 features ranked by Gain Ratio

Running the classification algorithm with the above settings, a possible rule (R_1) is extracted: *a power line is particularly prone to mechanical failure if it consists of: (i) an aerial section; (ii) a section with a rigid isolation of type RP5 with a length less than about 150 meters; (iii) between 6 and 12 secondary power grid substations; (iv) lacking of motorized switch-disconnectors and conductor junctions*. Notice that the rule is based on structural features of the MV line, due to their high GR value. However, it could be also useful to obtain further rules related to different attributes. In this case the classifier can be re-trained on the same dataset but selecting only a subset of initial attributes, *e.g.*, the ones derived from contextual conditions and current variations. A novel rule (R_2) can be so derived: *from January to September and in particular during Saturday, power lines are prone to structural failure if they present: (i) more than two automated secondary power grid substations; (ii) an available power less than 9000 kW on the line for MV clients in high concentration area; (iii) current values during the last week with a standard deviation less than 26 A or greater than 39 A and an absolute variance (on average) greater than about 1.40% between two consecutive current values.*

Consider that the above rules have been obtained via a fully automated approach, without the domain expert support. They could evidence obvious or already known situations, but also highlight new and original correlations between data. The experts should select and validate derived rules using them in redeploying the company maintenance procedures. In addition, as shown in Figure 5, for each line the classifier predicts an event with the related probability. Hence, it is also possible to refine the prediction by considering as reliable only the events with a confidence level greater than a threshold value. An empirical evaluation was executed to assign this value (T_P = 0.75) granting the highest accuracy of the prediction algorithm.

Detected rules can be finally converted in a semantic-based annotation. For example, the R_2 expressed in Description Logic [8] notation *w.r.t.* the reference ontology is:

$$R_2 \equiv \forall\, \text{during.}(\neg\, \text{October} \sqcap \neg\, \text{November} \sqcap \neg\, \text{December}) \sqcap \geq 2\ \text{hasAutomatedSecondarySubstation} \sqcap \forall\ \text{hasAvailablePower.}(\text{HighConcentrationArea} \sqcap \leq 9000\ \text{kWatt}) \sqcap \forall\ \text{hasStandardDeviation.}(\leq 26\ \text{Ampere} \sqcap \geq 39\ \text{Ampere}) \sqcap \forall\ \text{hasAbsoluteVariance.}(\geq 14\ \text{perMil})$$

In order to assess the feasibility of the proposed approach, an early performance evaluation was carried out for both data aggregation and failure prediction using the dataset about MV line data described in Section 2. For data aggregation, experiments basically aimed to: (i) measure the aggregation rate for the values of current; (ii) assess possible benefits w.r.t. a general-purpose compression algorithm; (iii) evaluate the information loss during the aggregation process.

	Worst Case	**Best Case**
Num. of observations	4370	
Data Size (kB)	135.85	135.69
Num. of aggregated observations	1176	44
Aggregated Size (kB)	51.65	1.99
GZIP Size (kB)	14.49	10.83
Aggregated + GZIP Size (kB)	9.67	0.32
Absolute information loss (A)	0.4	< 0.001
Relative information loss (%)	11.42	< 0.1

Table III. Data aggregation results

Table III shows the results obtained evaluating values of current for all 126 MV lines during the 27 months observation. Reported data are the average of single month results for a single line. In particular worst and best case represent a MV line with many and few current fluctuations, respectively. Notice that performance of the aggregation algorithm strongly depend on the current trend, in fact long sections of similar data can be easily aggregated whereas many variations need more storage space. However in both cases aggregation allows to considerably reduce the amount of data and in the best case it is also more efficient than GZIP[1] (used as reference compression algorithm). The combined usage of aggregation and general-purpose compression allows to obtain a

[1] RFC 1952, GZIP file format specification, version 4.3, May 1996, *http://www.ietf.org/rfc/rfc1952.txt*

remarkable reduction up to about 8.70% of the initial data size (considering the whole dataset), from 471 MB to 41 MB of storage memory.

Furthermore, to evaluate the information loss during the aggregation process, a comparison was made between early values and data rebuilt after aggregation. Table III also reports both absolute (Ampere) and relative (percentage error w.r.t. the average in a month) information loss. Also in this case, numerous variations lead to worst performance in terms of accuracy of data reconstruction. Anyway, the average relative approximation due to aggregation is of about 4% with respect to the initial values. Figure 6 shows a comparison between real (in red) and aggregated values (in blue) highlighting possible difference in the current trends.

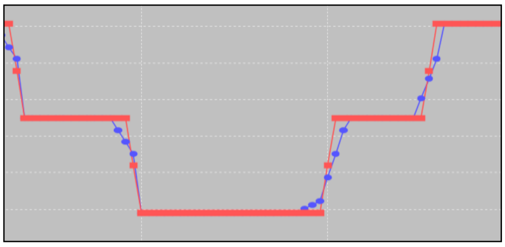


Figure 6. Data Analysis tool – Visualization of original (in blue) and rebuilt (in red) values

Finally an evaluation of the event prediction module was conducted to measure the accuracy of the proposed classifier. J48 was chosen as reference algorithm after a preliminary campaign: all algorithms were used to build a forecasting model, but higher accuracy was obtained with J48. The confusion matrix shown in Table IV reports on the weighted precision of the classifier and on single precision and recall values for each event.

Event	A	B	C	D	E	Recall (%)
(A) Mechanical Failure	**1**	2	0	0	0	33.3
(B) Structural Failure	3	**27**	0	1	0	87.1
(C) Water Infiltration	0	0	**0**	0	0	N.D.
(D) Breakage	0	1	0	**4**	0	80.0
(E) Generic Failure	0	3	0	0	**3**	50.0
Precision (%)	25.0	81.8	N.D	80.0	100	**77.8**

Table IV. Confusion Matrix

A prediction becomes true if the model allows prognosticating that a specific event occurs on a line within the test period. Observe that the classifier precision and recall are very high in case of structural failures, whereas they are lower for mechanical malfunctions. This is due to the short number of cases in both the training and validation datasets. However the precision can be further improved discarding, as said above, instances with a prediction confidence less than $T_{P.}$ In this case the algorithm reaches an overall precision of about 83%.

4. CONCLUSION AND FUTURE WORK

Early results were presented of an ongoing work on a dynamic, semantic-based and multidimensional knowledge discovery framework for grid failure prediction. Grid data cleaning (needed to provide an adequate input to the system for prognoses) was a lot more challenging than expected, requiring joint effort of analysts and domain experts to understand the peculiarities of the source database gathering stage. Furthermore, the scarcity of failure event data conditioned the effectiveness of the employed machine learning and inference algorithms. Nevertheless, the knowledge discovery process was able to identify non-obvious patterns characterizing grid interruption events. Performance of data aggregation has been satisfactory.

Further work is being done to improve the information cleaning and the machine learning tasks. Once the data management process is consolidated, the expansion of the data set with historical data and information extracted from other archives will allow to increase the effectiveness of the framework. Finally, a user interface exploiting the semantic characterization of information is being developed to support distributor managers in decision-making about grid maintenance.

REFERENCES

[1] C. Rudin, D. Waltz, R. Anderson, A. Boulanger, A. Salleb-Aouissi, M. Chow, H. Dutta, P. Gross, B. Huang, S. Ierome, D. Isaac, A. Kressner, R. Passonneau, A. Radeva e W. L., «Machine learning for the New York City power grid,» *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 34, n. 2, pp. 328-345, 2012.

[2] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann e I. Witten, «The WEKA Data Mining Software: An Update,» *SIGKDD Explorations,* vol. 11, n. 1, 2009.

[3] J. Quinlan, C4.5: programs for machine learning, Morgan Kaufmann Publishers Inc., 1993.

[4] B. Pfahringer, «Random model trees: an effective and scalable regression method,» 2010.

[5] J. Friedman, T. Hastie e R. Tibshirani, «Additive logistic regresssion: a statistical view of boosting,» *Annals of Statistics,* vol. 28, n. 2, pp. 337-407, 2000.

[6] N. Landwehr, M. Hall e E. Frank, «Logistic Model Trees,» *Machine Learning,* vol. 95, n. 1-2, pp. 161-205, 2005.

[7] M. Ruta, E. Di Sciascio e F. Scioscia, «Concept Abduction and Contraction in Semantic-based P2P Environments,» *Web Intelligence and Agent Systems,* vol. 9, n. 3, pp. 179-207, 2011.

[8] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi e P. F. Patel-schneider, The Description Logic Handbook, Cambridge University Press, 2002.