

Knowledge Discovery in Environmental Data Bases using GESCONDA

Karina Gibert¹, Xavier Flores², Ignasi Rodríguez-Roda², Miquel Sànchez-Marrè³

¹*Dep. Estadística i Investigació Operativa, Universitat Politècnica de Catalunya, Barcelona, Catalonia, EU
(karina@eio.upc.es)*

²*Laboratori d'Enginyeria Química i Ambiental (LEQUIA). Universitat de Girona, Girona, Catalonia, EU
({xavi | ignasi}@lequia.udg.es)*

³*Knowledge Engineering and Machine Learning Group (KEMLG). Universitat Politècnica de Catalunya ,
Barcelona, Catalonia, EU (miquel@lsi.upc.es)*

Abstract: In this work, last results of the research project “Development of an Intelligent Data Analysis System for Knowledge Management in Environmental Data Bases (DB)” are presented. The project is focussed on the design and development of a prototype for Knowledge Discovery (KD) and intelligent data analysis, and specially oriented to environmental DB. It is remarkable the high quantity of information and knowledge patterns that are implicit in large DB coming from environmental domains. In this project, several environmental DB such as meteorological phenomena, wastewater treatment plants (WWTP), or environmental emergencies were used for testing. KD is a prior and mandatory step to get reliable Intelligent Environmental Decision Support Systems. Although in the literature other KD tools exists (WEKA, Intelligent Miner...) none of them integrate, like GESCONDA, statistical and AI methods, the possibility of explicit management of the produced knowledge in Knowledge Bases (KB) (in the classical AI sense), mixed techniques that can cooperate among them to discover and extract the knowledge contained in data, dynamical data analysis... in a single tool, allowing interaction among all the methods. The purpose of the paper is to present the final architecture of GESCONDA, as well as some of the methods incorporated in last phases. Later, an application to discover knowledge patterns from an environmental DB (a WWTP) is detailed. The DB has been mined using several methods available in GESCONDA. First of all, statistical filtering approaches were applied for data preparation. Afterwards, a hybrid clustering technique (clustering based on rules) was used to discover the structure of the target phenomenon. Finally, clustering results were used as input for rule induction making new knowledge explicit. Results and feedback from validation steps show that the tool seems to be useful and efficient for KD.

Keywords: Knowledge Acquisition and Management, Data Mining, Machine Learning, Environmental Databases, Statistical Modelling, Rules Induction, WasteWater Treatment Plant.

1. INTRODUCTION

An Environmental Decision Support System (EDSS) can be defined as an intelligent information system for decreasing the decision-making time and improving consistency and quality of decisions in Environmental Systems.

An EDSS is an ideal decision-oriented tool for suggesting recommendations in an environmental domain. The main outstanding feature of EDSS is the knowledge embodied, which provides the system with enhanced abilities to reason about the

environmental system in a more reliable way. A common problem in their development is how to obtain that knowledge. Classic approaches are based on getting knowledge by manual interactive sessions with environmental experts. But when there are available databases summarising the behaviour of the environmental system in the past, there is a more interesting and promising approach: using several common automated techniques from both Statistics and Machine Learning fields. The conjoint use of those techniques is usually named as data mining (Gibert et al. 1998).

All this information and knowledge is very important for prediction, control, supervision and minimisation of environmental impact either in Nature and Human beings themselves. The project is involved with building an Intelligent Data Analysis System (IDAS) to provide the support to these kind of environmental decision-making. In this paper, last methods incorporated in system are introduced. A real application is presented in

order to illustrate how the system supports KD in a real-world environmental database.

3. FINAL ARCHITECTURE

GESCONDA is the name given to the IDAS developed within the project. On the basis of previous experiences, it was decided that it would have multi-layer architecture of 4 levels connecting the user with the environmental system or process.

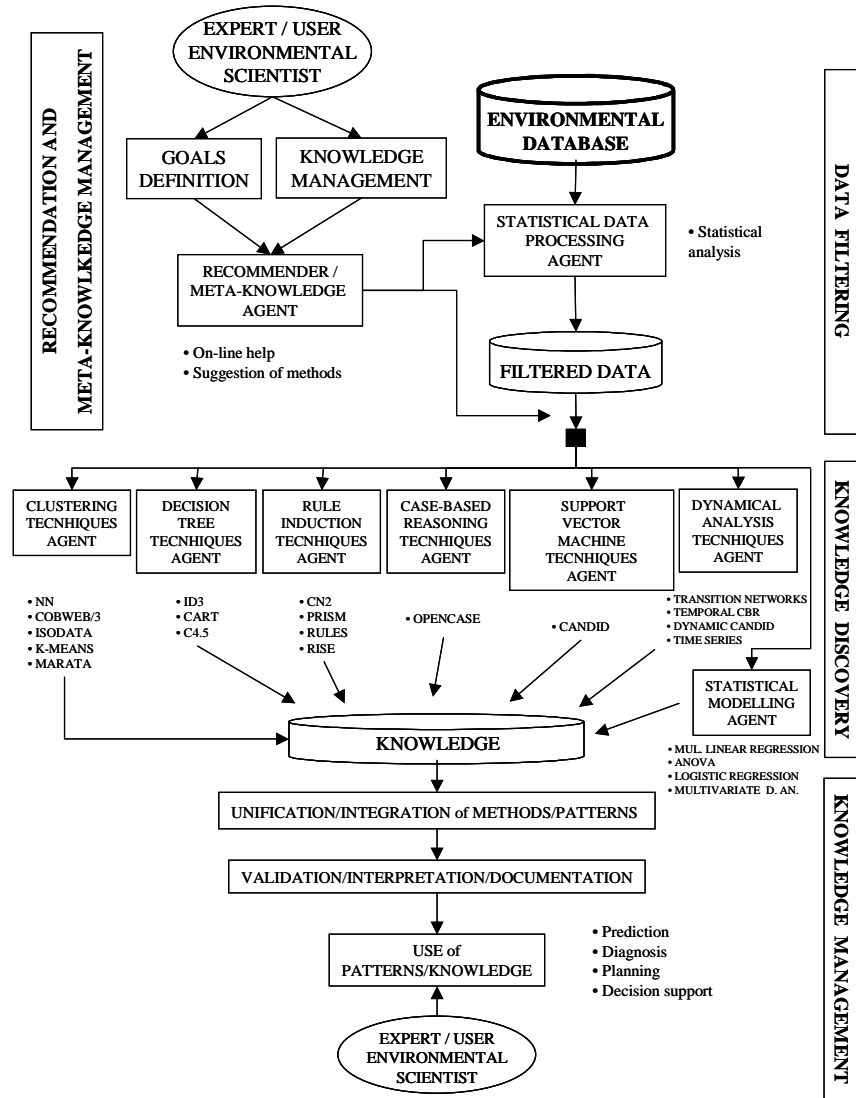


Figure 1. Architecture of GESCONDA

These 4 levels are the following:

- Data Filtering: a) Data cleaning; b) Missing data analysis and management; c) Outlier data analysis and management; d) Statistical one-way analysis; e) Statistical two-way analysis; o Graphical visualisation tools; f) Attribute or Variable transformation
- Recommendation and Meta-Knowledge Management: a) Problem goal definition; b) Method suggestion; c) Parameter setting; d) Attribute or

Variable Meta-Knowledge management; e) Example Meta-knowledge management; f) Domain theory knowledge elicitation.

- Knowledge Discovery: a) Clustering (ML and Statistical); b) Decision tree induction; c) Classification rule induction; d) Case-based reasoning; e) Support vector machine; f) Statistical Modelling; g) Dynamic analysis.
- Knowledge Management: a) Integration of different knowledge patterns for a predictive task,

or planning, or system supervision; b) Validation of the acquired Knowledge pattern; c) Knowledge utilisation by end-users; d) User interaction. Fig. 1, depicts the architecture of the system.

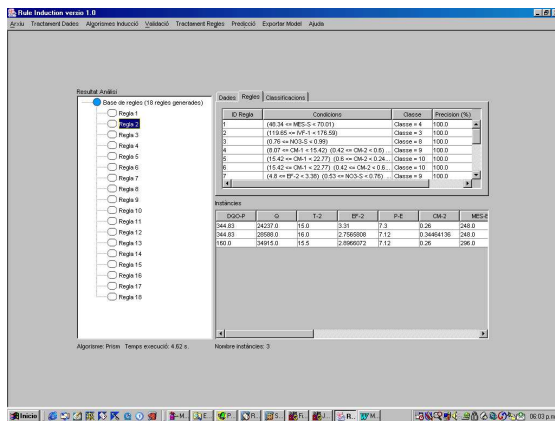


Figure 2 A set of induced rules

The GESCONDA system will provide a set of mixed techniques that will be useful to acquire relevant knowledge from environmental systems, through available databases. This knowledge will be used afterwards in the implementation of reliable EDSS. The portability of the software will be provided by a common Java platform.

In next section there is a more detailed description of the rules induction agent and the statistical modelling one.

4. AGENTS

In previous works (Sánchez-Marrè et al 2002) data-filtering agent, clustering agent and decision-tree induction agent, which were developed the first were already presented. In this paper, details on agents added to GESCONDA in a second development period are presented.

4.1 Rules induction

The rule induction agent is the responsible to induce classification rules directly from supervised databases. Resulting rule-bases can be the input for building a Knowledge-Based classifier system.

Induced rules have two components: the left-hand side is a boolean expression built as a conjunction of *selectors* or conditions on the attributes ($Length = high, Diameter \in [3.4, 5.8] \dots$); the right-hand side is the label of the class to which the instances satisfying the left-hand side belong ($Class4, \dots$).

These techniques induce a set of rules covering as much as possible all the instances within the training set. In general, not all the instances can be classified by the induced rule set. Resulting rules are provided with a predictive accuracy, estimation

rate and with coverage estimation rate. Some validation techniques, such as simple validation or cross-validation, have been implemented to test the reliability of induced rules. After validation, the rules set can be used as a model to predict the class of new unseen instances. The knowledge base can be exported to a text file or a CLIPS file, for a later use within a knowledge-based system.

Fig. 2 depicts a set of rules induced from a certain database by GESCONDA. Several different algorithms such as RULES (Pham&Aksoy, 1995), PRISM (Cendrowska, 1987), CN2 (Clark & Niblett, 1989), and RISE (Domingos, 1996) have been selected and implemented in the system. The agent also supports a tuning of the resulting rules, as well as the possibility of removing very low accurate rules either manually or automatically.

First three algorithms are selector-based. RULES computes induced rules starting from an initial rule with empty left-hand side. Step by step, it adds one selector each time until it obtains a 100% accurate rule, where all covering instances are correctly assigned to one class. If some instances are not classified yet, it builds a new possible rule.

PRISM is based on a similar principle as RULES. Main difference is that induced rules are computed separately for each class. First, only instances labelled with the first class are considered for the inductive process, and so on until the last class.

The very popular CN2 is based on a heuristic search for the best combination of selectors, which are known as *complex* in the algorithm terminology. Only k complex are maintained and explored. The best complex is selected as the basis for new rules. The right-hand side of the rule is set to the more frequent class from instances covered by the complex. When all instances are covered or no more complex can be formed it stops.

RISE is very different from the previous ones. It is an instance-based algorithm, which starts considering each instance within the training set as a possible rule. Iteratively, it generalizes the most similar instances (i.e., rules) making new rules, always more general. A similarity measure is needed. Several similarity measures have been implemented. It progresses until no more accuracy gain is obtained with a generalization step.

4.2 Statistical Modelling agent

The multivariate descriptive techniques allow studying the structure of a given domain. Further, it is convenient to properly formalize the model in order to use it in the future, either with descriptive or predictive purposes. When those models are formalized under a mathematical paradigm and

taking into account the uncertainty, statistical modelling is on. It establishes algebraic relations between a response variable and a set of explicative variables (attributes, regressors) in such a way that knowing the values of the explicative variables for a certain instance, the response value can be determined with a known precision.

The statistical model produces quantitative and formal results about relationships between variables and it complements the qualitative and non-formal results from descriptive analysis.

This agent is in charge of building different statistical models, depending on the cases:

- Multiple linear regression, which allows to relate a numerical response with a set of numerical or categorical (qualitative) regressors;
- ANOVA, involved with explaining a categorical response on the basis of a set of regressors;
- Logistic regression, explaining a dicotomic variable by a set of regressors.

For each method, the system is implementing the following steps:

- Parameters estimation for the model
- Providing goodness of fitting coefficients
- Providing tools for a graphical residuals analysis, in order to validate the model.

5. AN APPLICATION

5.1 The data

The main goal of wastewater treatment plants is to guarantee the outflow water quality (referred to certain legal requirements), in order to restore the natural environmental balance, which is disturbed by industry waste or domestic wastewater.

The process used to achieve this goal is really complex and delicate; on the one hand, because of the intrinsic features of wastewater; on the other hand, because of the bad consequences of an incorrect management of the plant (Gime98).

Data analyzed in this paper comes from a Waste Water Treatment Plant in Catalonia (in Spain). Here is a brief description of the plant performing (see fig. 3): the water flows sequentially through several processes; in the *pretreatment*, an initial separation of solids from wastewater is performed; *primary treatment* consists of leaving the wastewater in a settler for some hours; solids will deposit down the settler and could be sent out; *secondary treatment* occurs inside a biological reactor: a (biomass) population of microorganisms degrades the organic matter dissolved in the wastewater; in the studied particular plant there are two separate biological reactors with a second settler between them (double stage activated sludge plant or AB process); in the *advanced treatment* another settler is used to separate the water from the biomass and water is clean and ready to exit the plant. The settler output (solids or biomass) produces a kind of mud which is the input of another process called *sludge line*.

Database is a sample of 149 observations taken between January and May 2002. Each observation refers to a daily mean. The state of the Plant is described through a set of 18 variables (or attributes) which can be grouped as (see fig. 3):

- Input (measures taken at the plant entrance): *Q-E*: Inflow wastewater (daily m^3 of water); *DQO-E*: Oxigen chemical demand (mg/l); *MES-E*: Suspended Solids (mg/l); *P-E*: Phosphates (mg/l).
- After Settler (measures taken when the wastewater comes out of the first settler): *DQO-P*: Oxigen chemical demand (mg/l); *MES-P*: Suspended Solids (mg/l);
- Biological treatment 1 (in 1st biological reactor): *MLSS-1*: Mixed liquor suspended solids (mg/l); *IVF-1*: Volumetric index (ml/g); *CM-1*: Organic load (Kg DBO/ Kg MLSS).
- Biological treatment 2 (in 2nd biological reactor): *MLSS-2*: Mixed liquor suspended solids (mg/l); *IVF-2*: Volumetric index (ml/g); *CM-2*: Organic load (Kg DBO/Kg MLSS); *T-2*: Temperature (C°);

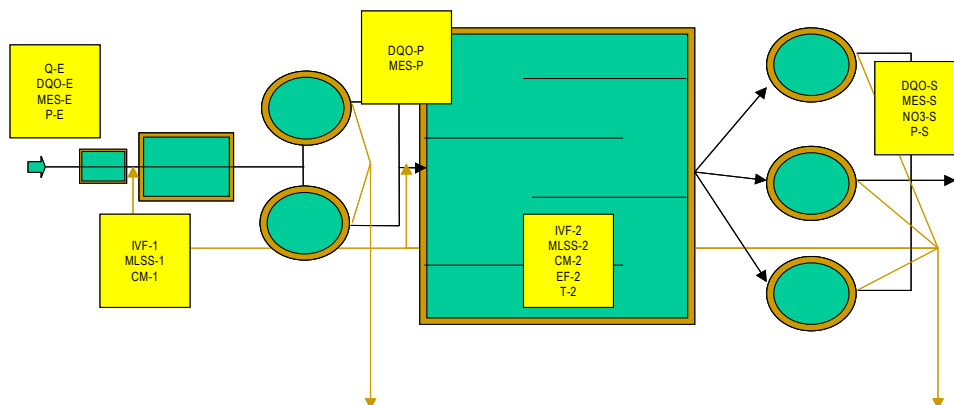


Figure 3. A wasetwater treatment plant chart

EF-2: Sludge Residence time (days).

- Output (when the water is meeting the river):
- DQO-S: Oxigen chemical demand (mg/l); MES-S: Suspended Solids (mg/l); NO3-S: Nitrates (mg/l); P-S: Phosphates (mg/l).

5.2 Descriptive analysis and data filtering

Facilities of Statistical Data Processing Agent were used for identifying outliers and missing data, and properly manage them. Descriptive analysis was also useful to identify some probabilistic models for some variables, which will determine the later statistical model construction.

5.3 Clustering

In order to identify the characteristic situations presented in the plant a clustering process was performed using Clustering based on rules (CIBR) (Gibert 1996), which is an ascendant hierarchical method that permits, among others, conjoint considering of numerical and categorical variables to identify the clusters. As usual in hierarchical clusters, the number of classes is determined a posteriori upon the hierarchy built by the method. As a result 12 classes were identified.

With the descriptive analysis agent, conditional distributions of different variables through classes could be studied. Fig. 4 shows first that of MLSS-1 (solids concentration in the reactor) along the classes: Classes were grouped in two main blocks, that where first step is operating as a properly biological reactor (with values of MLSS-1, greater than 3000 mg/l, classes 1 to 6), and that where it is working as preaeration (with low values of MLSS-1, classes 7 to 12), according to the special performing conditions of the studied plant, which indeed has a first stage that can operate in those two ways. Going further with such graphics for the rest of the variables (for instance column CM-1, fig. 4), differences among classes 1 to 6, on the one hand, and/or 7 to 12 on the other can be studied, what makes possible a first interpretation of classes. However, it is interesting to see how an induction rule method can also discover the meaning of the classes in an automatic way, confirming what it can already be seen here using only descriptive techniques (see 5.4).

5.4 Rule induction

After determining and evaluating the reliability of the clustering process, a new step to induce classification rules was performed. The rule induction agent used the class identifier obtained by the clustering agent as the input label of the

instances. In this application several inductive methods were ran in order to induce rules that could be later used for recognizing the class of a new instance. Also, by analysing the meaning of obtained rules, interpretation of classes will be clearer. Some parameters were tested and tuned, such as the number of intervals in the discretization of the continuous attributes, done by trial and error.

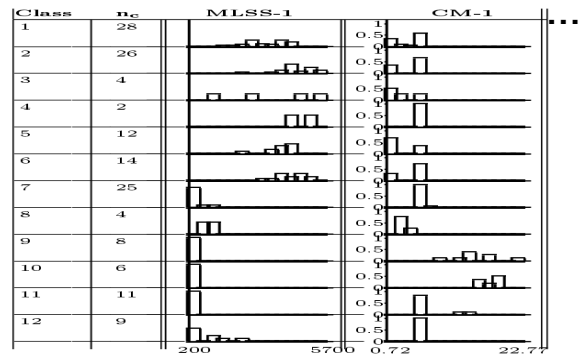


Figure 4 Multiple boxplot

Upon the experts criteria, the best set of rules, in terms of making explicit the knowledge about plant performing, is that produced by Prism. So, a final set of 18 rules is proposed as the best. Fig. 2 shows GESCONDA's rule induction process. It is worth noticing that some rules addressed concrete classes: r2 identified class 3, which corresponds to an initial bulking situation, and r3 identified cluster 8, which corresponds to nitrifying situation, although there are classes predicted by more than one rule. Thus, experts concluded that the inductive rules and the clusters previously obtained were a really useful and coherent knowledge discovered from the available data.

Rules generated by PRISM are totally accurate, but the coverage is not so good, and some instances could not be classified with the rule set. However, experts found the rules representative enough.

5.5 Final interpretation of the classes

As said before, in the first six classes the first stage of the plant works as biological reactor, while in the others as preaeration. Here is the final interpretation of the classes, which was built together with the experts, by combining the conditional distributions of all the variables along classes (fig. 4) with the rules induced by Prism:

- First stage operating as biological reactor:
 - Class 1: it is the common situation. This class is labelled as normal or correct plant operation.
 - Class 2: it is quite similar to class 1 but with optimal operation (better effluent quality).
 - Class 3: abnormal situation owing proliferation

of filamentous bacteria (bulking) in first stage, difficulting sludge settleability.

- *Class 4*: refers to those days with higher loading rates (organic overloading).
- *Classes 5, 6*: bulking (filamentous organisms proliferation) episodes in the second stage
- First stage operating as preaeration:
 - *Classes 7 and 12*: common situation.
 - *Class 8*: periods of partial nitrification due to the growth of autotrophic biomass.
 - *Class 9*: rainy and stormy days, with a nutrient disequilibrium suitable for viscous bulking .
 - *Classes 10, 11*: viscous bulking, not associated to exceeding proliferation of filamentous organisms. Mainly related to *Zooglea Ramigera*. This commonly occurs after rainy periods, associated with nutrient disequilibrium

6. CONCLUSIONS

The main conclusion is that GESCONDA is an Intelligent Data Analysis System, which offers a common interface to the user for using a set of different tools that helps his/her decision-making processes. From different real applications, it has been seen that this is a very promising approach and the previous partial experiences on this line suggested great benefits making it. Currently, the statistical data-filtering agent, the clustering agent, the decision trees induction agent and the rules induction agent are completed; the statistical modelling agent is including multiple linear regression and ANOVA (on and two way), and logistic regression is in progress.

Main agents are already built and the schedule of the project was correctly followed. At present, renovation for the next three years is pending, in order to face the remaining agents development (support vector machines or dynamical analysis). Validation of the current version of the system using real databases is on. The close assessment of the environmental engineers of LEQUIA, and SOREA people guarantee usefulness of the system.

From the presented application, it can be said that WasteWater Treatment Plants constitutes a complex domain which requires complex analysis using different approaches in order to extract useful knowledge. CIBR appeared to be a good method for identifying typical situations in that domain (like bulking or storming days). Exploratory techniques, such as displaying conditional distributions of the variables vs classes are of great help to understand the meaning of the classes. However, the possibility offered by GESCONDA of combining the results of a clustering process with rules induction, made much

more easier the interpretation and allowed consolidation of the discovered knowledge. An integrated tool like the proposed one allows facing the analysis of phenomena, like WWTP, where knowledge is not well-established yet and permits knowledge discovery in a friendly way.

From this results, an initial case base has already been built to be included in the supervisory system of a real plant which is using case based reasoning. As a complementary study, statistical models for predicting the class of a new observation on the basis of the variables identified as relevant in the rules induction process will provide a quantitative model useful for bounding the prediction error rate.

6. ACKNOWLEDGEMENTS

The authors wish to thank the partial support provided by the Spanish CICYT project TIC2000-1011, and by the EU project A-TEAM (IST 1999-10176). Thanks to Francesc Coll and M. Eugenia Garcia for partially developing a piece of software.

7. REFERENCES

- Cendrowska, J 1987. PRISM: an algorithm for inducing module rules. *Int'l Journal of Man-Machine Studies* 27(4):349-370.
- Clark & Niblett, 1989. P. Clark & T. Niblett. The CN" induction algorithm. *ML* 3:261-283.
- Comas J., Dzeroski S. Gibert, K., Rodríguez-Roda I. and Sánchez-Marrè M. 2001 Knowledge Discovery by means of inductive methods in WWTP data. *AI Comm* 14(1):45-62.
- Domingos, P 1996 Unifying Instance-Based and Rule-Based Induction, *ML* 24(2):141-168.
- Gibert K. 1996 The use of symbolic information in automation of statistical treatment of ISD. *AI Communications* 9(1) 36—37 IOS
- Gibert K., T. Aluja and U. Cortés, 1998 Knowledge Discovery with Clustering Based on Rules. *Interpr... LNAI* 510:83-92, Springer
- Pham & Aksoy, 1995 D.T. Pham & M.S. Aksoy. RULES: a simple ruler extraction system. *Expert Systems with Applications* 8(1).
- Rodríguez-Roda I., Comas J., Colprim J., Poch M., Sánchez-Marrè M., Cortés U., Baeza, J. & Lafuente (2002) *J. Water Science & Technology*, 45(4-5), pp. 289-297
- Sánchez-Marrè M., Gibert, K., Rodríguez-Roda, I., et al. (2002) Development of an Intelligent Data Analysis System for Knowledge Management in Environmental Data Bases. First Int'l Conf. of Int'l EMS Society. Vol III 420-425. Lugano, Suiza.