Jul 1st, 12:00 AM

Development of an Intelligent Data Analysis System for Knowledge Management in Environmental Data Bases

Miquel Sànchez-Marrè

Karina Gibert

Ignasi Rodríguez-Roda

Eva Bueno

Lidia Mozo

See next page for additional authors

Presenter/Author Information

Miquel Sànchez-Marrè, Karina Gibert, Ignasi Rodríguez-Roda, Eva Bueno, Lidia Mozo, Aleix Clavell, Mario Martína, and Philippe Rougé

Development of an Intelligent Data Analysis System for Knowledge Management in Environmental Data Bases

Miquel Sànchez-Marrè[a], <u>Karina Gibert</u>[b], Ignasi Rodríguez-Roda[c], Eva Bueno[d], Lidia Mozo[d], Aleix Clavell[e], Mario Martín[a] and Philippe Rougé[f]

[a]*Knowldege Engineering and Machine Learning group (KEML). Universitat Politècnica de Catalunya , Barcelona, Catalonia, EU ({mmartin, miquel}@lsi.upc.es)*

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

[c]*Laboratori d'Enginyeria Química i Ambiental (LEQUIA). Universitat de Girona, Girona, Catalonia, EU (ignasi@lequia.udg.es)*

[d]*Facultat de Matemàtiques i Estadística, Universitat Politècnica de Catalunya, Barcelona, Catalonia, EU*

[e]*Facultat d'Informàtica de Barcelona, Universitat Politècnica de Catalunya, Barcelona, Catalonia, EU*

[f]*Sociedad Regional de Abastecimiento de Aguas, S.A. (SOREA), Grupo AGBAR, Barcelona (prouge@agbar.es)*

Abstract: In this work, the goals of the ongoing research project "Development of an Intelligent Data Analysis System for Knowledge Management in Environmental Data Bases" are presented. This project started in 2000, and the Spanish Government will finance it until 2003. Its main goal is to design and develop a prototype of a tool for intelligent data analysis and implicit knowledge management of databases, with special focus on environmental databases. It is remarkable the high quantity of information and knowledge patterns implicit in large databases coming from the monitoring of any system or dynamical environmental process. For instance, historical data collected about meteorological phenomena in a certain area, about the performance of a wastewater treatment plant, about characterising environmental emergencies (toxic substances wasting, inflammable gas expansion), or about geomorphologycal description of seismic activity.

Differing from the existing commercial systems, the more relevant aspects of this proposal are: the interaction of the developed methods, the development of mixed techniques that can cooperate among them to extract the knowledge contained in data, the existence of dynamical data analysis, and the existence of a recommender agent, which will suggest the best method to be used depending on the target domain and on the goals specified by users.

The purpose of the paper is to present the architecture of the system as well as some of the methods it incorporates.

Keywords: Knowledge Acquisition and Management, Data Mining, Machine Learning, Environmental Databases.

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 obtaining the knowledge with 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. These joint techniques are usually named as data mining or knowledge discovery technologies.

All this information and knowledge is very important for prediction tasks, control, supervision and minimisation of environmental impact either in Nature and Human beings themselves. The project is involved with building an Intelligent Data Analysis (IDA) tool to provide the support to these kind of environmental systems. This tool is basically composed by several statistical data analysis methods, such as one-way and two-way descriptive statistics, missing data analysis, clustering, and relations between variables. Also, several machine learning techniques will be integrated, coming from Artificial Intelligence, such as clustering, classification rule induction, decision tree induction, case-based reasoning techniques, reinforcement learning, and dynamical analysis.

In this paper, a progress report of the project "Development of an Intelligent Data Analysis System for Knowledge Management in Environmental Data Bases" is presented.

2. PARTNERS

A multidisciplinary team of partners develops this project:

- The main partner is the KEML group (Knowledge Engineering and Machine Learning), from UPC (Universitat Politècnica de Catalunya), which since 1989 is researching on knowledge acquisition, knowledge engineering, machine learning, data mining and intelligent decision-support systems. Recently, KEML has been working in the development of Artificial Intelligence integrated architectures, to control and supervise environmental complex processes. The group is also interested in the design of a methodology for automatic knowledge acquisition and data analysis in ill-structured domains, in the development of an environment for developing case-based reasoning systems, and in the implementation of the needed tools for applying those methodologies to the automatic building of

intelligent systems, especially in environmental domains.

- The Environmental and Chemical Engineering experts in this project are the team from the Chemical and Environmental Engineering Laboratory from UdG (Universitat de Girona), which has a great experience on environmental systems, especially on Wastewater Treatment Plants. This group has worked on many plants spread over Catalonia.

- Finally, it was considered that involving a company into the project would be an excellent opportunity for applying the results of the project to a real environmental system. So, the third participant in this project is SOREA (Sociedad Regional de Abastecimiento de Aguas, S.A.), which is interested in the project results; it belongs to the multinational group AGBAR, and one of its registered firms is NETAIGUA, S.A. It focuses on Water management in Europe and Latin America, as well as other tasks of environmental engineering. Its participation will allow the evaluation of applicability of the final tool, and experts of the company would be able to take part in the decision making, for determining and guiding the advance of the project.

3. ARCHITECTURE OF THE SYSTEM

The objective of the project is to design and implement an Intelligent Data Analysis System (IDAS). GESCONDA is the name given to the IDAS developed within the project. On the basis of previous experiences, it was decided that GESCONDA would have multi-layer architecture of 4 levels connecting the user with the environmental system or process. These 4 levels are the following:

- Data Filtering
  - o Data cleaning
  - o Missing data analysis and management
  - o Outlier data analysis and management
  - o Statistical one-way analysis
  - o Statistical two-way analysis
  - o Graphical visualisation tools
  - o Atribute/Variable transformation
- Recommendation and Meta-Knowledge Management
  - o Problem goal definition
  - o Method suggestion
  - o Parameter setting

o Atribute/Variable Meta-knowledge management
o Example Meta-knowledge management

o Domain theory knowledge elicitation
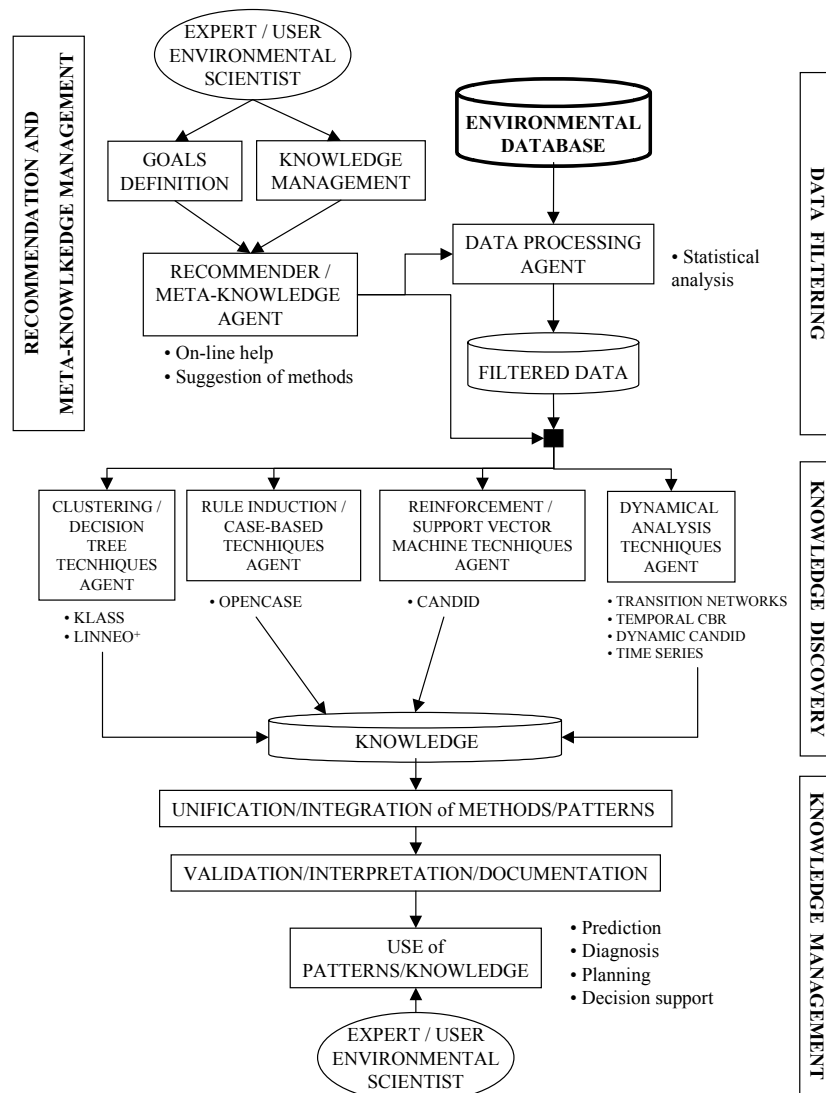
Figure 1. Architecture of GESCONDA

- Knowledge Discovery

  o Statistical and machine learning clustering methods
  o Decision tree induction methods
  o Classification rule induction methods
  o Case-based reasoning methods
  o Reinforcement learning methods
  o Support vector machine methods
  o Dynamic analysis methods

- Knowledge Management

  o Integration of different knowledge patterns for a

predictive task, or planning, or system supervision.
  o Validation of the knowledge pattern acquired
  o Knowledge utilisation by end-users
  o User interaction

In figure 1, the architecture of the system is depicted in detail.

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 the next section there is a more detailed description of the statistical data-filtering agent and of the clustering agent.

4. STATISTICAL DATA-FILTERING AND CLUSTERING AGENTS

The statistical data filtering agent is in charge of database management, statistical descriptive analysis, and graphical representations. These tasks provide the whole IDA system with powerful data filtering techniques to prepare the data for later knowledge discovery step.

Database management allows adding a new variable to the database, deleting one variable from the database, and modifying the characteristics of a variable such as its relevance or the range of values. Also, the management of the examples is supported. Some variable transformations such as re-coding and standardisation are provided. Different random number distribution generation and different probability distribution generation are supported too. Figure 2 depicts a new variable addition in the database.
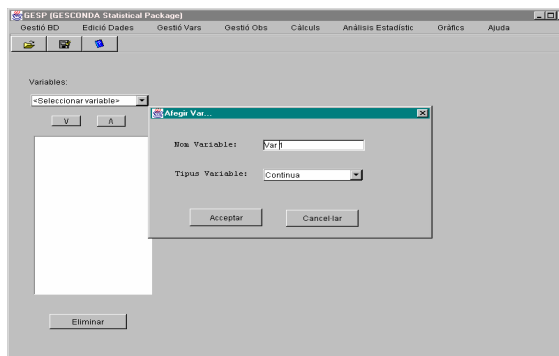
Figure 2. New variable addition in a database.

Descriptive statistical analysis is composed by basic statistical analysis such as computation of mean, standard deviation, median value or correlation coefficient. One-way and two-way analysis of both variables and classes are also provided. Missing value management is also supported. Figure 3 depicts a screen with several possible descriptive statistical analysis.

Graphical representations of analysis results are implemented through both one-way plots and two-way plots, as well as histograms or letterplots for class distribution visualisation.
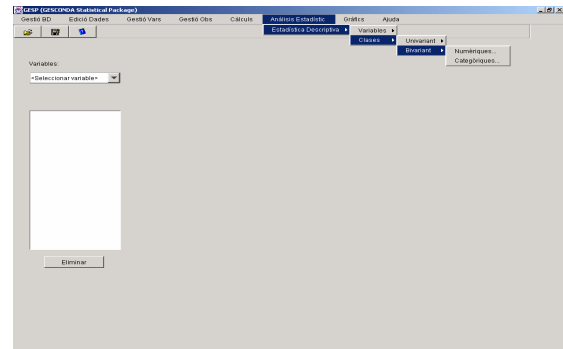
Figure 3. Possible descriptive statistical analysis

The clustering agent is the responsible to induce possible clusters of data from unsupervised databases. The goal is to determine the number of clusters or classes and/or setting a class/cluster label for each example of the database. Several algorithms such as K-means algorithm, a nearest-neighbour classifier, Cobweb/3 algorithm and Isodata algorithm are already implemented, and others are foreseen in next future, such as Autoclass, some fuzzy clustering algorithm, and bagging techniques will be applied to derive better classifiers . Several similarity/distance measures have been implemented such as Clark, Canberra, Manhattan, Euclidean, *L'Eixample* to provide a more customisable system.

K-means algorithm has been already implemented. The purpose of K-means is to classify a set of data in such a way that those belonging to a given group are as similar as possible. To this end, the matrix of similarity must be established. Each element of the matrix has a number, which is the measure of similarity between each pair of objects. This method requires an a priori definition of a certain number of groups. Analysis rearranges the objects on the basis of the variables selected, so that at the end of the process they are as similar as possible.

A nearest-neighbour classifier is implemented. The nearest-neighbour scheme proceeds in an incremental way adding examples from the database to the more similar/less distant prototype of previously formed classes.

Cobweb/3 is a modification of original Cobweb algorithm for incremental clustering managing both continuous and discrete attributes. It generates an incremental hierarchical classification of examples based on the merging and splitting operations, using the category utility criterion to measure the quality of a partition. The output from a Cobweb clustering result is shown in figure 4.
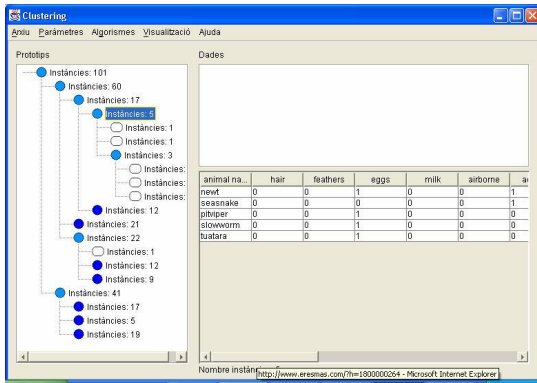
Figure 4. Cobweb hierarchical clustering results.

Isodata is a partitional clustering algorithm where new clusters can be created or merged from existing clusters if certain conditions are met. A cluster is split if it has too many examples and an unusually large variance along the feature with largest spread. Two clusters are merged if their cluster centres are sufficiently close, based on a parameter supplied by the user.

The clustering agent implements the different algorithms and has some visualisation tools, such as histograms, two-way plots of clusters, and tables with all the elements of a cluster. See figure 5.
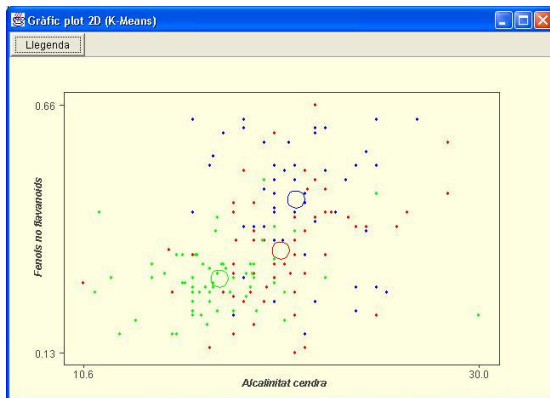

Figure 5. Two-way plot of a clustering process

5. PREVIOUS WORK AND APPLICATIONS

The previous research lines of the participants originated this project. In fact, since much time ago, the partners were working in related fields, and several autonomous tools were already designed. These tools will be integrated altogether in the kernel of the GESCONDA system, in order to get a complete IDAS which could provide different capabilities to the Environmental System managers.

Among the previous results related with the project goals, the following tools should be mentioned:

- Opencase, Sànchez-Marrè *et al*. [1999]

- Linneo+, Béjar [1995]

- KLASS+, Gibert [1995] and Gibert [1998]

- CANDID, Martín [1995]

As mentioned above, all this tools will be absorbed in GESCONDA and will constitute a part of the system kernel.

In fact, there have also been some attempts of building integrated tools for supporting decision making in the specific context of WasteWater Treatment Plants. DEPUR Serra *et al*. [1994] , Sànchez-Marrè [1991] was a first prototype for an knowledge-based system oriented to the diagnose of WWTP operating situation. Afterwards, and integrated architecture for control and supervision of WWTP, called DAI-DEPUR, Sànchez-*Marrè et al*. [1996] was built, including Opencase. Recently, a knowledge-based Hybrid Supervisory System to support the operation of a real Wastewater Treatment Plant has been developed and implemented, which has been successfully performing real-time support to the operation of the Granollers facility since September 1999 Rodríguez-Roda *et al*. [2002].

The KEML group is also fostering the BESAI research group, which is an international group, mainly European, to put in contact AI researchers and Environmental Science researchers, with the aim of building a corpus of multidisciplinary knowledge, and making co-operative efforts for environmental problem solving by means of AI techniques.

In the literature, there are some related works on the machine learning active role in environmental data mining such as in Morabito [2001], Comas *et al*. [2001], Demyanov [2000] and Bratko *et al*. [2000]. Also, the project EDAM (INTAS 99-00099), which is a project on environmental data mining, learning algorithms and statistical tools for monitoring and forecasting, is being carried out by some European research centres. See Kanevsky *et al*. [2000].

6. CONLUSIONS AND FUTURE WORK

The main conclusion is that the construction of an Intelligent Data Analysis system, which offers a common interface to the user for using a set of different tools for helping his/her decision-making

processes, is very promising and the previous partial experiences on this line suggested great benefits making it. Currently, the statistical data-filtering agent is near completed with many data and statistical management functionalities, and the clustering technique agent is very advanced, with K-means algorithm, a nearest-neighbour clustering algorithm, Isodata algorithm and COBWEB/3 algorithm already implemented.

As mentioned above, some of the agents are already being built and the schedule of the project is correctly followed. In the future, the other agents will be also built, and finally the validation of the system with real databases, with the collaboration of the SOREA people and the environmental engineers of the LEQUIA group, will guarantee the usefulness of the system.

7. ACKNOWLEDGEMENTS

8. REFERENCES

Béjar J. Knowledge Acquisition in ill-structured domains. Ph.D. Thesis. Dept. de Llenguatges i Sistemes Informàtics. Universitat Politècnica de Catalunya. (In Spanish), 1995.

Bratko I., Dzeroski S., Kompare B. And Urbancic T. *Analysis of environmental data with machine learning methods*. Jozef Stefan Institute, Center for Knowledge Transfer in Information Technologies, 2000.

Comas J., Dzeroski S. Gibert, K., Rodríguez-Roda I. and Sànchez-Marrè M. Knowledge Discovery by means of inductive methods in wastewater treatment plant data. *AI Communications* 14(1):45-62, January 2001.

Demyanov V., M. Kanevski, E. Savelieva, V. Timonin, S. Chernov, V. Polishuk. Neural Network Residual Stochastic Cosimulation for Environmental Data Analysis, Proc. of *2nd ICSC Symposium on Neural Computation (NC'2000)*, May 2000, Berlin, Germany, pp. 647-653.

Gibert K., T. Aluja and U. Cortés, Knowledge Discovery with Clustering Based on Rules. Interpreting Results,In *Principles of Data Mining and Knowledge Discovery*, J. M. \.{Z}ytkow, M. Quafafou Eds., LNAI 510, p 83-92, Springer-Verlag, 1998.

Gibert K. and Cortés U. Combining a knowledge based system with a clustering method for an inductive construction of models}. In P. Cheeseman et al. (Eds.), *Selecting Models from Data: AI and Statistics IV*, LNS no 89,

pp 351--360, New York, Springer-Verlag, 1994.

Gimeno J.M., J. Béjar, M. Sànchez-Marrè, U. Cortés and I. R.-Roda. Discovering and Modelling Process Change: an Application to Industrial Processes. Proc. of *2nd Int. Conference on the Practical Application of Knowledge Discovery and Data Mining (PADD 98)*, pp. 143-153. London, U.K. March, 1998.

Kanevski M., M. Maignan, A. Pozdnukhov, S. Canu. *Environmental Data Mining with Machine Learning and Geostatistics,* RR-00-10, May 2000.

Martín, M. Reinforcement learning for embedded agents facing complex tasks. PhD thesis. Departament de Llenguatges i Sistemes Informàtics. Universitat Politècnica de Catalunya, 1998.

Morabito F. C. *Environmental data interpretation: the next challenge for intelligent systems.* NATO Advanced Research Workshop on Systematic Organization of Information in Fuzzy Systems. 2001, Vila Real, Portugal.

Rodríguez-Roda I., Comas J., Colprim J., Poch M., Sànchez-Marrè M., Cortés U., Baeza, J. & Lafuente J. *Water Science & Technology*, 45(4-5), pp. 289-297 (2002). ISSN 0273-1223.

Sànchez-Marrè M., U. Cortés, I. R.-Roda, M. Poch. Sustainable Case Learning for continuous domains. *Environmental Modelling & Software* 14:349-357, 1999a.

Sànchez-Marrè M., U. Cortés, J. Béjar, I. R.-Roda and M. Poch. Reflective Reasoning in a CBR Agent. Chapter in *Collaboration Between Human and Artificial Societies* (J. A. Padget, ed.). Lectures Notes in Artifical Intelligence, LNAI-1624, Springer-Verlag. December, 1999b.

Sànchez-Marrè M., U. Cortés, J. Lafuente, I. R.-Roda y M. Poch. DAI-DEPUR: an integrated and distributed architecture for wastewater treatment plants supervision. *Artificial Intelligence in Engineering* 10(3):275-285. Elsevier Science Ltd., 1996.

Sànchez-Marrè M., I. R-Roda, J. Lafuente, U. Cortés y M. Poch. DAI-DEPUR Architecture: Distributed Agents for Real-Time WWTP Supervision and Control. *2nd IFAC/IFIP/IMACS Symposium on Artificial Intelligence in Real Time Control (AIRTC'94)*, València, pp. 179-184, 1994.

Serra P., M. Sànchez-Marrè, J. Lafuente, U. Cortés y M. Poch. DEPUR: a knowledge based tool for wastewater treatment plants. *Engineering Applications of Artificial Intelligence* 7(1):23-30, 1994.