GESCONDA: A Tool for Knowledge Discovery and Data Mining in Environmental Databases

Miquel Sànchez-Marrè[1], Karina Gibert[1,2] and Ignasi Rodríguez-Roda[3]

[1] Knowldege Engineering and Machine Learning group (KEML.G). Universitat Politècnica de Catalunya. Campus Nord-Edifici C5. Jordi Girona 1-3. 08034 Barcelona. Catalonia. EU
miquel@lsi.upc.es

[2] Dep. Statistics and Operation Research. Universitat Politècnica de Catalunya. Pau Gargallo 5. 08028 Barcelona. Catalonia. EU
karina@eio.upc.es

[3] Laboratori d'Enginyeria Química i Ambiental (LEQUIA). Universitat de Girona. Campus de Montilivi s/n. 17071 Girona. Catalonia. EU
ignasi@lequia.udg.es

Abstract. In this paper, a tool for *Knowledge Discovery* and *Data Mining* in environmental databases is presented. In the long term, the main goal of this research is to design and develop a tool, named GESCONDA, for intelligent data analysis and management of implicit knowledge from databases; it will provide support to *Knowledge Discovery* and *Data Mining* tasks that can guide the decision-making process, with special focus on environmental databases. The first stage of the project is to develop a prototype of the tool. Differing from the existing commercial systems, the more relevant aspects of this proposal are the possibility of interaction between the developed methods, the development of mixed techniques (combining tools from different fields, as AI or Statistics, that 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 its functionality and to illustrate some of the possibilities of supporting knowledge discovery and data mining on environmental real domains. Finally, the use of GESCONDA in the context of environmental systems is presented, as well as results obtained in a concrete case study.

1 Introduction

An Intelligent Environmental Decision Support System (IEDSS) 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 IEDSS is an ideal decision-oriented tool for suggesting recommendations in an environmental domain. The main outstanding feature of IEDSS 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 between environmental experts and knowledge engineers. However, when databases summarising the historical behaviour of the environmental system are available, a more interesting and promising approach is suitable: that of using several well known automated techniques from both Statistics and Machine Learning fields to get the knowledge. The use of all those techniques together for analysing data are usually named as *Data Mining* or *Knowledge Discovery* technologies.

It is remarkable the high quantity of information and implicit knowledge patterns 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, or about the performance of a wastewater treatment plant, or about characterizing environmental emergencies (toxic substances wasting, inflammable gas expansion), or about geomorphologycal description of seismic activity. 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.

This research is involved with building an Intelligent Data Analysis System (IDAS) 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, or modelling relationships between variables. Also, several machine learning techniques, coming from Artificial Intelligence, are integrated in the same tool, such as clustering, classification, rule induction, decision tree induction, case-based reasoning techniques, support vector machines, and dynamical analysis. The system is also provided with a higher level component, which allows information exchange between one method and another and also with a recommender that, given the problem, can suggest the more appropriate analysis.

In the literature, there are some related works and financed research projects on the active machine learning field in environmental *data mining*, such as in [19], [6], [9] and [2]. Also, some European research centres are involved in the project EDAM (INTAS 99-00099) [17], a research project on environmental *data mining*, learning algorithms and statistical tools for monitoring and forecasting. However, authors are not aware of the existence of a specific software for *knowledge discovery* and *data mining* of environmental databases, taking into account the special features of environmental domains, such as the temporal and dynamic component of data, or the problem of noisy data, and data filtering with no clear relevant or irrelevant features. In fact, these are major differences with other commercial or freeware software, such as WEKA.

The issue of our work aims at designing and building an *Intelligent Knowledge Data Discovery and Data Mining System*, especially suitable for environmental data analysis. The software tool, which is called GESCONDA [14], [16], [26], is, at present, partially built-up, and it will be growing with the addition of new functionalities in the near future. The development of GESCONDA is the main goal

of an ongoing research project financed by the Spanish Government between 2000 and the end of 2003. At present, a continuation of the project is still financed by the Spanish Government till the end of 2007.

The text is organised as follows. In section 2 the functionalities and architecture of GESCONDA are described. Section 3 shows the use of the tool to discover new knowledge in some common *scenario* showing the input data to the system, the software execution, and final outputs that GESCONDA can generate. The application of the tool to real environmental databases, and in a particular case study is detailed in section 4. Finally, in section 5, main conclusions and future developments of the tool are discussed.

2 Functionalities 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 a multi-layer architecture of 4 levels (see figure 1) connecting the user with the environmental system or process. These 4 levels are the following:

- Data Filtering: providing tools for data cleaning and data preparing
  - Data cleaning
  - Missing data analysis and management
  - Outlier data analysis and management
  - Statistical one-way analysis
  - Statistical two-way analysis
  - Graphical visualisation tools
  - Attribute/Variable transformation
  - Random number and random variable generators
  - Learning feature relevance techniques

- Recommendation and Meta-Knowledge Management: providing tools for recommending a proper way to face the analysis in order to extract the more useful knowledge regarding the concrete problem to be solved.
  - Problem goal definition
  - Method suggestion
  - Parameter setting
  - Attribute/Variable Meta-knowledge management
  - Example Meta-knowledge management
  - Domain theory knowledge elicitation

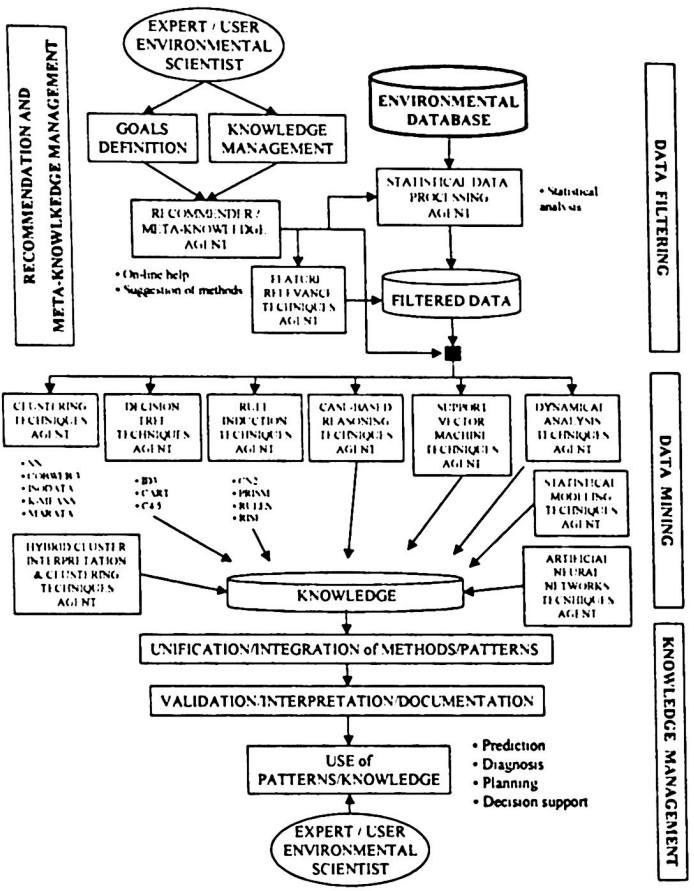

Fig. 1. Architecture of GESCONDA

- Data Mining: providing a set of different *data mining* techniques, coming either from Statistics or Artificial Intelligence.
  - Statistical and machine learning clustering: to discover groups of homogeneous objects in a given database.
  - Decision tree induction: To discover the relevant discriminant attributes regarding a given classification of the objects.
  - Classification rule induction: to discover classification rules that can predict the label of new unseen instances of the database.
  - Case-based reasoning: to predict the label of new unseen instances or cases by similarity to previously learned instances. Also, dynamic analyses of data are possible.

- Reinforcement learning: to learn the structure of a given database on the basis of positive and negative reinforcement signals.
- Support vector machine: to learn some numerical functions to classify instances among several classes.
- Statistical modelling: to extract predictive numerical models based on can relationships among variables and providing variability explanations.
- Dynamic analysis: to face the analysis of temporal and/or spatial evolution of certain environmental processes using several techniques.

- Knowledge Management
  - Integration of different knowledge patterns for a predictive task, or planning, or system supervision.
  - Validation of the knowledge pattern acquired by the *Data Mining* techniques. For the different techniques there are different suitable tools, such as cross-validation methods for estimating the accuracy of several inductive models, such as decision trees or classification rules or graphical residuals' analysis for statistical modelling and correctness analysis of several clustering methods for clustering analysis.
  - Knowledge utilisation by end-users
  - User interaction

GESCONDA provides 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 IEDSS. The portability of the software is provided by a common Java platform.

In next sections there is a more detailed description of the possibilities of the software regarding the knowledge discovery from environmental databases.

3 Running Data Analysis

GESCONDA is a standard Java application that, once installed, can be launched from the Start Menu. The environment is a main window embedding a Menu Bar and a small Toolbar.

Input data files can be analysed by GESCONDA. They follow the standard format of instance disposal in rows, and attribute disposal in columns. Prior to the data file loading, the user should introduce the meta-information associated with each variable into the system. Thus, the user must specify if the variable is qualitative or quantitative, and in the later option, she/he should list its modalities, as well as their ordering, if existent. Also, the weight of the variable can be modified. In addition, the variables can be declared as active for the analyses or not, depending on the issues of the user.

After loading the data file, all changes can be saved into a GESCONDA database file format (GSP file), in order to retrieve the work in future working sessions with the tool. Retrieving the work is done through the opening of a previously created

database (GSP file). Once data are loaded, the first thing to do in order to extract knowledge patterns from data is the descriptive statistical analysis and the data filtering task. This lets the user to check whether there are errors, outliers, bad codified data, missing data, as well as to summarise main data features, such as the minimum and maximum values, the mean, standard deviation, variance, and so on. Additionally, if variable transformations are needed previous to the analysis, such as linear transformation, variable re-coding or variable standardisation, tools for arranging the variables according to the user needs are provided. Other facilities such as random variable generation following several distributions such as Bernoulli, Binomial, Gaussian, Exponential, Uniform, Discrete or specific probability value computation are also available, providing, among others, the possibility of selecting random samples, from the data base, to be considered, for instance in a re-sampling process. Feature relevance techniques are available too.

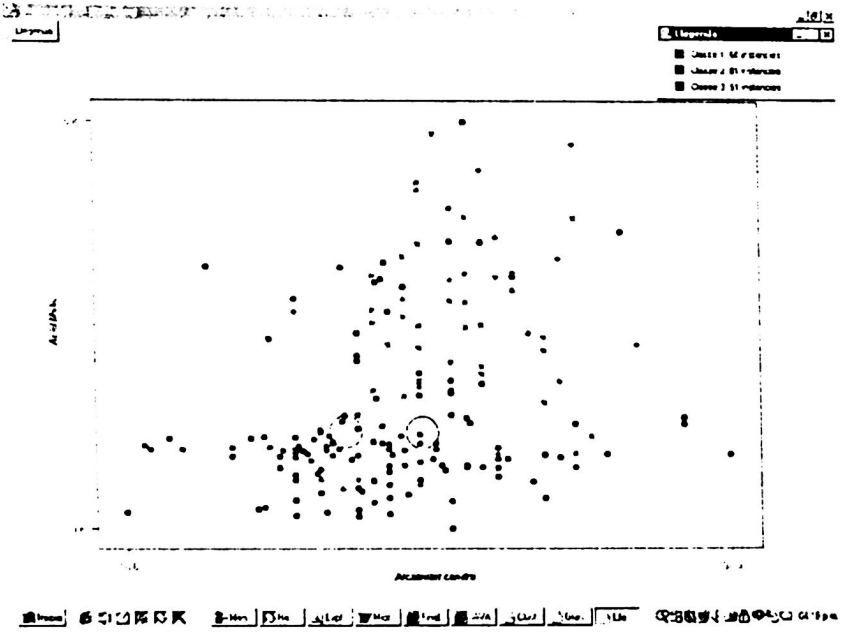

Fig. 2. Results from a clustering process

Different data mining techniques can be used, according to the user's goal, which could be to discover some concepts hidden in the data (clustering or grouping), or to discover some discriminant knowledge (decision trees, classification rules) or to induce some quantitative model (statistical modelling) to be used for later prediction of the concept (class or cluster) corresponding to new instances. Furthermore, they could be combined to make a more accurate data analysis. Different *scenarios* are possible.

One common *scenario* could be when one is facing an unknown environmental database, with a huge amount of instances and/or features. One possibility is to start by using a clustering technique to identify typical situations in the target environmental process. Several methods are available in GESCONDA, such as K-means [11], Isodata [1], Nearest-Neighbour classifier [8], [12], Marata and COBWEB/3 [18].

After experimenting with the different techniques, and trying several parameter values of the methods, GESCONDA provides the user with a sensitivity analysis giving information about coincidences in the classes discovered by different methods, which can be used for finding the stable set of classes [14]. Some graphical tools for visualizing the obtained classes and prototypes can be used, like that in the figure 2. The resulting class of every instance is recorded as a new attribute or variable, which could be later involved as the response variable in further analysis. For instance, an inductive decision tree technique can be used discover a predictive knowledge model, such a decision tree, to find the best set of attributes to predict the class label for new instances of the environmental database. In GESCONDA, the user can select and test ID3 [22], CART [3], C4.5 [21] methods, with optional pruning techniques. Another complementary action is to directly induce classification rules to predict the class label for a new instance. Several methods exist in the machine learning field, but within GESCONDA, RULES [20], PRISM [4], CN2 [5], and RISE [10] methods are implemented. The user can test several methods' parameters and validate the obtained classification rules. A snapshot of one window with the induced classification rules is depicted in figure 3. Some validation techniques (simple validation, cross-validation) are also available to test the quality of the induced models.

Both from the decision tree model or from the directly induced classification rules, a predictive knowledge model, implemented as a knowledge base, can be directly built with the final classification rules. This knowledge pattern can be used, for instance, to set-up an IDSS for predictive tasks in an environmental domain.

Another possibility, suitable for databases containing only numerical variables, is to use the statistical modelling component, currently with multiple linear regression or analysis of variance models (ANOVA). In that case, quantitative models (linear models or correlation models) are found and followed by a validation process with several charts, graphs and model parameter estimation.

Finally, the case based reasoning techniques included in GESCONDA provide a reasoning about data based on analogy analysis, which is able to propose control or supervisory actions or plans to safely guide the environmental process to normal situations.

GESCONDA has been used with different databases. One of the fields in which it has shown best results is in the field of Wastewater Treatment Plants, as it will be detailed in next section.

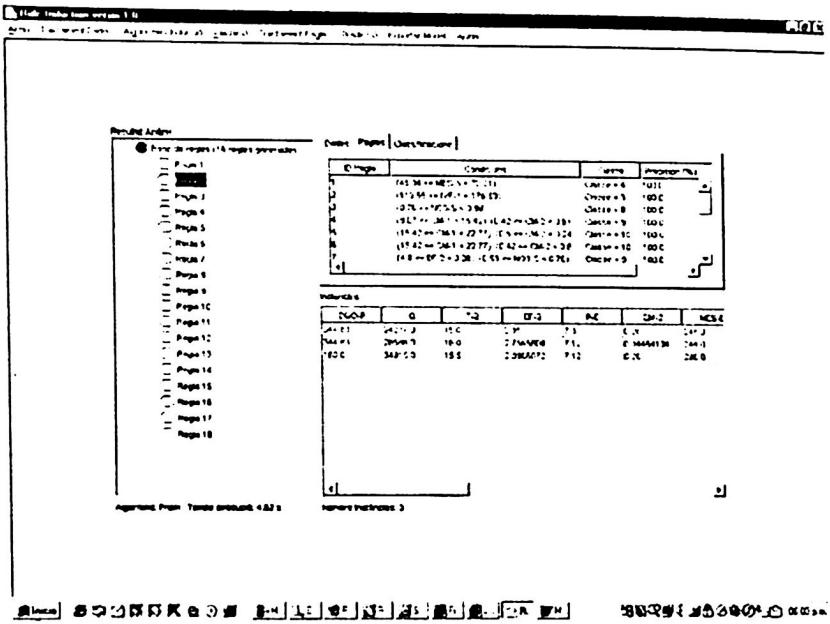

Fig. 3. Induced classification rules from data

4 Application

GESCONDA has been used to face different aspects of real Environmental Systems. In particular, it has been successfully used to analyse data coming from Wastewater Treatment Plants (WWTP), or to discover and manage the hidden knowledge in other environmental domains, such as air pollution from México D.F. However, in this paper, an overview of the research developed in the field of WWTP is detailed.

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 industrial and waste or domestic Waste Water. The biological process used to achieve this goal is really complex and delicate; on the one hand, because of the intrinsic features of the input, the wastewater; on the other hand, because of the serious consequences of an incorrect management of the plant.

A tool like GESCONDA allows extraction of relevant knowledge on the WasteWater Treatment process, which is really difficult to model using classical mechanistic and physical models. Thus, in such dangerous environmental systems, a tool like GESCONDA provides a high valuable support for catching the knowledge and experience about the process from databases, and provides useful guides and

prognosis to improve the supervision, control and management of a plant. This enables the development of better IEDSS for governmental agencies or companies in charge of wastewater treatment plants management.

In fact, GESCONDA has been used with several Catalan WWTP, such as Manresa [27], Girona, Cassà de la Selva-Llagostera, Lloret de Mar, Granollers and Montornés. Although the flow diagram and the technology of these facilities present some slight differences, the amount, type and frequency of measured variables is almost the same in all the plants.

Among others, identification of characteristic situations in a given WWTP was faced using clustering or *clustering based on rules* [15]. Those methods have been successfully applied to all the WWTP mentioned above. Some typical situations concerning the inflow have been clearly identified in all the *scenarios* such as storm, overloading, underloading or industrial waste, while some specific situations have only been identified from those databases that include qualitative variables (e.g. rising sludge, foaming, filamentous bulking, etc.).

Moreover, case-based reasoning techniques have been developed to diagnose and control the activated sludge process of Girona WWTP [25], as Granollers WWTP and La Llagosta WWTP. Between twelve and seventeen variables have been chosen to define the case structure, most of them quantitative (e.g. temperature, flow, suspended solids, ammonia concentration, etc.), but including some significant qualitative variables (predominant protozoa, sludge aspect, etc). A hierarchical discretized tree was used as indexing structure in the case libraries, which were fed with an initial seed.

From those results, validated Knowledge Bases codified by means of heuristic rules [7] could be developed, covering operational faults due to mechanical equipment or electrical failures (e.g. clogged pumps or air system faults), primary and secondary treatment operational problems, and transition states to intermediate alarms.

In fact, using the different techniques provided by GESCONDA, a knowledge model of the operation of the plant can be induced. This knowledge is suitable for a knowledge-based system that can assist the control of the plant. The two-phase methodology used to acquire and tune the knowledge is detailed in [24]. For the plant of Granollers, an IEDSS to supervise the plant operation was built-up as described in [23]. It was implemented and performing real-time support to the process management of the wastewater treatment plant. Results obtained for the plant of Girona are not installed at the facility yet, but those from La Llagosta WWTP, are being implemented right now.

However, it is important to remark tat some of the variables that clearly characterise identified situations cannot be used on-line, mainly because of the analytical delay, which in some cases is up to five days. This should to be taken into account when transferring the initially extracted knowledge (with clustering tools) to the knowledge-based system.

4.1 A Case Study

As a concrete illustration, the results obtained for the plant of Lloret de Mar are presented here with more detail.

Data comes from the water line of the WWTP. Here is a brief description of the process (see fig. 4), where the wastewater flows sequentially through two main stages:

- The *pre-treatment* consists in a screening. The principal role of screening is to remove big solids and gravel from the flow stream, which could damage subsequent process equipment and reduce overall process reliability.
- The *biological treatment* where a (biomass) population of microorganisms (biomass) degrades the organic matter dissolved in the wastewater, this occurs inside the biological reactors. The biological treatment of the target plant consists of a double stage activated sludge configuration, also known as *two-sludge process*. The two-sludge process is the following:
  - a system using a *high-rate activated sludge* for organic matter removal
  - followed by a second stage for *nitrification*

A portion of wastewater influent can be by-passed around the first stage to provide organic matter and biomass for the nitrification process and promote the flocculation and solids capture in the secondary clarification. The main reason to separate the organic matter stage from the nitrogen removal is to treat toxic substances in the first stage, and thus protect the sensitive nitrifying bacteria. Water leaving the biological reactors is left in settler for some hours, where the biomass is separated from the water by gravity and then the clean water is discharged to the environment.

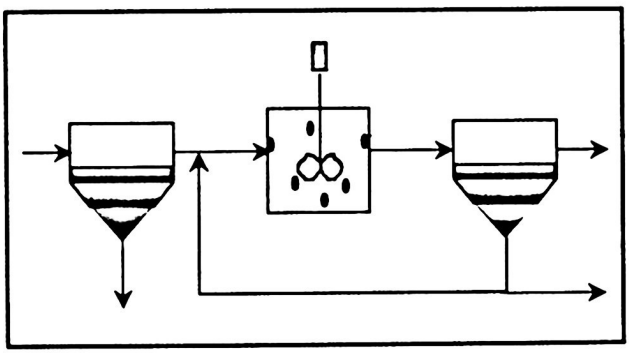

Fig. 4. Flowchart of the water line of a wastewater treatment plant

Database is a sample of 149 observations taken between January and May 2002 from the plant referred above. 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. 4):

- Input (measures taken at the entrance of the plant)
  - Q-E: Inflow waste water (daily cubic meters of water)
  - DQO-E: Chemical oxygen demand (mg/l)
  - MES-E: Suspended Solids (mg/l)
  - P-E: Phosphate (mg/l)
- First stage process variables (measures taken in the first biological reactor):
  - MLSS-1: Mixed liquor suspended solids at the biological reactor (mg/l)
  - IVF-1: Sludge volumetric index (ml/g)
  - CM-1: Organic load (kg of organic matter/ Kg MLSS)
- First stage output variables (measures taken when the wastewater comes out the from the first stage settler):
  - DQO-P: Chemical oxygen demand (mg/l)
  - MES-P: Suspended Solids (mg/l)
- Second stage process variables (measures taken in the second stage biological reactor):
  - MLSS-2: Mixed liquor suspended solids (mg/l)
  - IVF-2: Sludge volumetric index (ml/g)
  - CM-2: Organic load (kg of organic matter/Kg MLSS)
  - T-2: Temperature (Centigrade)
  - EF-2: Sludge Retention Time (days)
- Output (measures taken when the water is meeting the river):
  - DQO-S: Chemical oxygen demand (mg/l)
  - MES-S: Suspended Solids (mg/l)
  - NO3-S: Nitrate (mg/l)
  - P-S: Phosphate (mg/l).

First of all, descriptive statistics was used for data cleaning. Outliers identification and missing data treatment was very important, since in this specific context, some missing values are not at random. Indeed, some analytical tests are only performed in front of certain problems in the plant, so, missing values as the result of a given test is already an indicator of the lack of the corresponding problem. Also, feature weighting techniques were used to provide a relevance degree of each feature of the database to get more accurate similarity computations.

Then, reciprocal neighbours method was used as the clustering method to identify typical situations in the plant, providing a 12-classes partition P_{RN} (including details on the meaning of those classes). Then, the experts, taking as a validation criteria that they could understand the *meaning* of the identified classes, manually validated the results.

Once validated, P_{RN} could be considered as a reference partition for quality assessment of the clusters obtained with other clustering methods provided by GESCONDA. For this particular plant, other partition of the same data set were found by using the following methods (between brackets the identifier of every partition):

- Reciprocal neighbours, using the criteria of Ward. (P_{RN}).
- K-means, asking for 12 classes and using bagging. (P_K).

- Marata, asking for 12 classes, (P_M).
- Nearest-neighbour, with similarity threshold 0.48, producing 12 classes (P_{NN}).
- Isodata with the following set of parameters: 12 initial and final classes, threshold 0.05 for merges, 25 iterations and by default values for the remaining parameters, (P_I).

Table 1. Quality coefficients regarding PRN

Partition	% of well classified objects
P_K	54.6
P_M	38.77
P_{NN}	50.015
P_I	44.52

For every partition, the percentage of *well-classified* objects regarding PRN can be calculated (see table 1) and it can be seen which clustering method is providing closer results to those accepted by the experts (in this particular application K-means with bagging seems to be the best). It has to be taken into account that this value, even being close to 50%, is far from random assignment, since the number of classes to be recognized is much bigger than 2.

As a matter of fact, this quality coefficient can be used as some kind of *similarity* between partitions, as the higher it is, more similarly the objects are classified in the pair of compared partitions. This allows building table 2, which is suitable for a more global analysis of the results. It shows that the more similar results are obtained by P_{RN} and P_K, with a 54.6% of objects classified in the same way, while the more different pair are P_K and P_M which have a 73.3% of objects classified in different ways.

Table 2. Similarity between partitions

Partition	PRN	PK	PM	PNN
PK	54.6			
PM	50.015	22.18		
PNN	38.77	46.42	41.81	
PI	44.52	39.81	28.62	47.41

Even more, given a reference partition, GESCONDA provides, for each other partition, a *validation report* that, besides the quality coefficient, establishes the correspondence between the class labels of both partitions (upon bigger intersections) and provides some information about the intersections in both partitions. Taking advantage of this information, table 3 shows the size of the classes of every partition. It can be seen that Marata is concentrating majority of objects in two single classes, while the rest contains isolated objects in general; this can be the reason why, for this particular application, Marata shows low similarity with other partitions. From table 3, Isodata seems the method that produces a more uniform distribution of objects in classes of similar sizes. It can also be appreciated that some classes keep their

behaviour independently of the method. For instance, classes 1 and 7 (both referring to common situation, see [14]) are big for all partitions. On the contrary, classes:

- c3 (abnormal situations due to proliferation of filamentous bacteria (bulking) in the first stage that difficult the sludge settleability)
- c4 (days with higher loading rates, organic overloading)
- c8 (periods of partial nitrification due to the growth of autotrophic biomass)
- c10 (viscous bulking)

use to be small classes in all the methods (excepting Isodata).

Table 3. Class sizes

Partition	PRN	PK	PM	PNN	PI
c1	28	26	81	52	22
c2	20	25	2	2	17
c3	4	4	1	1	10
c4	2	0	1	2	10
c5	12	1	3	1	9
c6	14	24	1	30	16
c7	25	27	50	35	11
c8	4	1	3	2	20
c9	8	8	1	1	11
c10	6	6	3	5	3
c11	11	13	2	17	15
c12	9	14	1	1	5

This is consistent with the fact that common situation is the most frequent, and small classes are identifying some problems that, fortunately, do not appear very often in the plant.

The other classes have variable size depending on the method. Taking into account the form of the 12 classes in a graphical way (using graphs like figure 2), for the five partitions, it can be seen that in general, there are some strong groups that become well identified by all the clustering methods used. Some other parts of the data set have a class assignment more sensible to the method.

It is interesting to note here that this stable groups are always recognised, even when the very nature of the method is really different, like in this case, where originally mathematical methods, like reciprocal neighbours, born in statistical context and based in algebraic concepts are confronted with others originally conceived in the context of Artificial Intelligence, which use a more logical-like paradigm and generating prototypes [13].

On the other hand, with the same database, several rule induction and decision tree techniques were also used to discover knowledge patterns [6] comparing the results in terms of predictive accuracy, the number of attributes and examples used, and the meaningfulness to experts. With the different methods could be ensured again that major groups or classes were correctly predicted with most of the methods.

Doing this kind of work with commercial data mining packages available is not easy, at present. GESCONDA offers an integrated environment especially oriented to complex dynamical environmental databases making easier the extraction of knowledge, the knowledge consolidation and its later use.

5 Conclusions and Future Work

An integrated tool including clustering agents, inductive learning and other statistical tools, makes easier sharing results of different knowledge acquisition methods, in order to improve and validate the later induced knowledge models. This is one of the outstanding features of the GESCONDA tool, in comparison against other commercial tools for data mining and knowledge discovery purposes.

In addition, the meta-knowledge management and recommendation agent, which will be implemented in a near future, will be other important characteristic of the tool. Also, temporal analysis techniques which will be offered in next versions of GESCONDA will clearly point its suitability for environmental databases data mining and knowledge discovery, against other commercial tools.

The main conclusion is that the construction of an integrated Intelligent Data Analysis system, which offers a common interface to the user for applying a set of different tools for helping his/her knowledge acquisition and decision-making processes, is very promising. Previous partial experiences with the current prototype on this line suggested great benefits of making it. Currently, the statistical data filtering with many data and statistical management functionalities is completed, and a first version of clustering, inductive decision tree, statistical modelling, classification rule induction agents are working in a beta-version of GESCONDA. The development of clustering based on rules, which was originally in KLASS, now is being built in Java, to be included in the clustering agent when finished. Dynamical analysis of spatial/temporal trends is being studied and developed, such as case-based reasoning techniques or temporal statistical series models or qualitative transition networks analysis.

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 depicted in the figure 1 will also be built, and finally, the validation of the system with real databases, with the collaboration of some Companies' staff and the environmental engineers of the LEQUIA group, will continue guaranteeing the usefulness of the system.

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). Also, they want to acknowledge Philippe Rougé, and SOREA Company for their kind collaboration.

and to Eva Bueno, Lídia Mozo, Aleix Clavell, Guillaume Larré, Francesc Coll and Eugènia Bòrdes for implementing parts of the system.

References

1. Ball, G.H., Hall, D.J.: ISODATA, a novel method of data analysis and pattern classification. Technical Report. Stanford Research Institute. (1965).
2. Bratko, I., Dzeroski ,S., Kompare, B., Urbancic, T.: *Analysis of environmental data with machine learning methods*. Jozef Stefan Institute, Center for Knowledge Transfer in Information Technologies, (2000).
3. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: *Classification and Regression Trees*, Belmont: Wadsworth, Belmont, (1984).
4. Cendrowska, J.: PRISM: an algorithm for inducing module rules. *International Journal of Man-Machine Studies* 27 (4). (1987) 349-370.
5. Clark, P., Niblett, T: The CN2 induction algorithm. *Machine Learning* 3. (1989) 261-283.
6. Comas, J., Dzeroski, Gibert, K., Rodriguez-Roda I. and Sànchez-Marrè M.: Knowledge Discovery by means of inductive methods in wastewater treatment plant data. *AI Communications* 14 (1). (2001) 45-62.
7. Comas, J., Rodriguez-Roda, I., Sànchez-Marrè, M., Cortés, U., Freixó, A., Arráez, J., Poch, M. A knowledge-based approach to the deflocculation problem: integrating on-line, off-line, and heuristic information. *Water Research*, 37, (2003) 2377-2387.
8. Cover, T.M., Hart, P.E.: Nearest neighbour pattern classification. *IEEE Transactions on Information Theory*, 13 (1968) 21-27.
9. Demyanov, V., Kanevski, M., Savelieva, E. ,Timonin, V., Chernov, S., Polishuk, V.: Neural Network Residual Stochastic Cosimulation for Environmental Data Analysis. Proc. of *2nd ICSC Symposium on Neural Computation (NC'2000)*, May 2000, Berlin, Germany, (2000) 647-653.
10. Domingos, P.: Unifying Instance-Based and Rule-Based Induction, *Machine Learning* 24 (2) (1996) 141-168.
11. Dubes, R., Kain, A.: Algorithms for Clustering Data. Prentice-Hall, Englewood Cliffs, USA, (1988).
12. Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis. John Wiley & Sons, (1973).
13. Gubert, K.: AI and Statistics techniques for Knowledge Discovery and Data Mining. In *Tendencias de la Mineria de Datos en España*. (Eds.) Raúl Giráldez, José C. Riquelme, Jesús S. Aguilar-Ruiz. In press. (2004).
14. Gubert, K., Flores, X., Rodriguez-Roda, I., Sànchez-Marrè, M.: Comparison of Classifications in Environmental Databases using GESCONDA. 4th ECAI Workshop on Binding Environmental Sciences and Artificial Intelligence (BESAI'2004). *BESAI'2004 Workshop Notes, pp. 13:1-13:10*. València, Comunitat Valenciana, (2004).
15. Gibert, K., Cortés, U. Clustering based on rules and Knowledge Discovery in ill-structured domains. *Computación y Sistemas* 1(4): CIC, Instituto Politécnico Nacional, (1998). 213-227.
16. Gibert, K, Flores, X., Rodriguez-Roda, I., Sànchez-Marrè, M.: Knowledge Discovery in environmental databases using GESCONDA. In *Proc. of 2nd International Environmental Modelling & Software Society Conference (iEMSs'2004)*, pp 50-55. Manno, Switzerland, (2004).
17. Kanevski, M., Maignan, M., Pozdnukhov, A., Canu, S.: *Environmental Data Mining with Machine Learning and Geostatistics*. RR-00-10, (2000).

18. McKusick, K. B., Thompson, K.: COBWEB/3: A portable implementation (Tech. Rep. No. FIA-90-6-18-2). Moffett Field, CA: NASA Ames Research Center, Artificial Intelligence Research Branch (1990).
19. Morabito, F. C.: *Environmental data interpretation: the next challenge for intelligent systems*. NATO Advanced Research Workshop on Systematic Organization of Information in Fuzzy Systems. Vila Real, Portugal (2001).
20. Pham, Aksoy:. RULES: a simple ruler extraction system. *Expert Systems with Applications* 8 (1) (1995).
1.Quinlan, J.R.: *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA, (1993).
22. Quinlan, J.R.: Induction of decision trees, *Machine Learning* 1(1) (1986), 81- 106.
23. Rodriguez-Roda, I., Comas, J., Colprim J., Poch, M., Sànchez-Marrè, M., Cortés, U., Baeza, J., Lafuente, J.: A hybrid supervisory system to support wastewater treatment plant operation: implementation and validation. *Water Science & Technology*, 45 (4-5), (2002) 289-297.
24. R-Roda, I. Comas, J., Poch, M., Sànchez-Marrè M. Cortés, U.: Automatic Knowledge Acquisition from Complex Processes for the development of Knowledge-Based Systems. *Industrial & Engineering Chemistry Research*, 40, (2001) 3353-3360.
25. R.-Roda I., Poch M., Sànchez-Marrè M., Cortés U., and Lafuente J. Consider a Case-Based System for Control of Complex Processes. *Chemical Engineering Progress*, 95(6), 39-45 (1999).
26. Sànchez-Marrè, M., Gubert, K., Rodriguez-Roda, R., Bueno, E., Mozo, L., Clavell, A., Martin, M. Rougé, P.: Development of an Intelligent Data Analysis System for Knowledge Management in Environmental Data Bases. *Procc. of I'' International Environmental Modelling & Software Society Conference (iEMSs'2002)*, 3, Lugano, Switzerland, (2002). 420-425.
27. Sànchez-Marrè, M., Béjar, J., Cortés, U., Gràcia, J., Lafuente, J., Poch, M. Concept formation in WWTP by means of classification techniques: a compared study. *Applied Intelligence*, 7(2), .(1997). 147-166.

Part IV

Remote Sensing