iEMSs 2008: International Congress on Environmental Modelling and Software
Integrating Sciences and Information Technology for Environmental Assessment and Decision Making
4th Biennial Meeting of iEMSs, http://www.iemss.org/iemss2008/index.php?n=Main.Proceedings
M. Sànchez-Marrè, J. Béjar, J. Comas, A. Rizzoli and G. Guariso (Eds.)
International Environmental Modelling and Software Society (iEMSs), 2008

On the role of pre and post-processing in environmental data mining

Karina Gibert[1,2], Joaquín Izquierdo[3], Geoff Holmes[4],Ioannis Athanasiadis[5], Joaquim Comas[6], Miquel Sànchez-Marrè[2]

[1] *Department of Statistics and Operation Research, Technical University of Catalonia, Barcelona, Catalonia*
[2] *Knowledge Engineering and Machine Learning Group, Technical University of Catalonia, Barcelona, Catalonia*
[3] *Centro Multidisciplinar de Modelación de Fluidos, Universidad Politécnica de Valencia, Spain.*
[4] *Department of Computer Science, University of Waikato, New Zealand*
[5] *Istituto Dalle Molle di Studi sull'Intelligenza Artificiale, Lugano, Switzerland*
[6] *Laboratory of Chemical and Environmental Engineering (LEQUIA), University of Girona*

Abstract: The quality of discovered knowledge is highly depending on data quality. Unfortunately real data use to contain noise, uncertainty, errors, redundancies or even irrelevant information. The more complex is the reality to be analyzed, the higher the risk of getting low quality data. Knowledge Discovery from Databases (KDD) offers a global framework to prepare data in the right form to perform correct analyses. On the other hand, the quality of decisions taken upon KDD results, depend not only on the quality of the results themselves, but on the capacity of the system to communicate those results in an understandable form. Environmental systems are particularly complex and environmental users particularly require clarity in their results. In this paper some details about how this can be achieved are provided. The role of the pre and post processing in the whole process of Knowledge Discovery in environmental systems is discussed.

Keywords: preprocessing; postprocessing; data mining; knowledge discovery of data; statistics; artificial intelligence; environmental systems.

1. INTRODUCTION

Environmental systems (ES) typically contain many interrelated components and processes, which may be biological, physical, geological, climatic, chemical, or social. Whenever we attempt to analyze ES and associated problems, we are immediately confronted with complexity stemming from various sources. However, there is a great need for data analysis, modelling of ES and development of decision support systems in order to improve the understanding of ES behaviour and the management of these complex systems (specially under abnormal situations). As stated in [Gibert et al.2008], the special features of environmental processes demand a new paradigm to improve the analysis and consequently its management.

Knowledge Discovery of Data (KDD) appeared in 1989 referring to high level applications which include particular methods of *Data Mining (DM, see figure 1)*, oriented to extract useful and understandable knowledge from data. KDD processes and the application of DM techniques are specifically appealing for environmental data, since activities permitting

extraction of maximum useful information from data bases are *per se* very important although they use to be preparatory for an environmental software system development. Also the KDD approach facilitates the integration of different knowledge sources and fields of expertise and the involvement of end-user (domain expert) criteria and stakeholders' points of view in algorithm design and result interpretation. Finally, it facilitates the sharing and rapid re-use of data and extracted technical knowledge and experiences among domain experts.

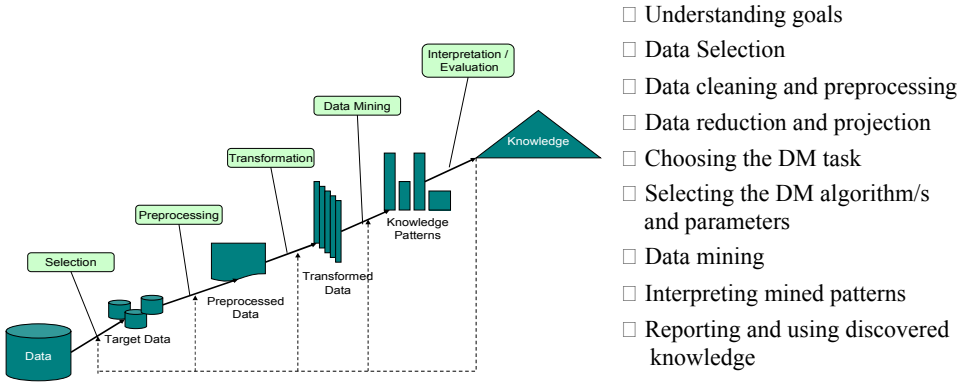

Figure 1 Outline of the Knowledge Discovery from Data process

Fayyad's proposal marked the beginning of a new paradigm in KDD research, considering prior and posterior analysis as important as the application of DM techniques itself:

> *"Most previous work on KDD has focused on [...] DM step. However, the other steps are of considerable importance for the successful application of KDD in practice".*

In fact, prior and posterior analysis requires great effort when dealing with real applications. Prior analysis is critical, mainly owing to two reasons:

- Real data sets tend to be imperfect, contains errors, outliers, missing data, extra noise and tools either for detecting or correcting it are required.

- Application of a certain data mining technique may require specific conditions for the data set (only binary variables, centered data, normality, only qualitative variables, etc). In this case, tools for verifying that those conditions hold as well as to transforming data in the appropriate way in order that they hold them, are required.

In environmental data, where errors of measuring (from automatic monitoring), uncertainty, imprecision, multi-scalarity, heterogeneity, non-linearities, non stacionariety, non-normality, are frequent, together with redundant variables, or irrelevant, or even contradictory, systematic and objective exploration as well as visualization and transformation of data is particularly critical for:

- better understanding the data set
- detecting imperfections in the data and managing them in the proper way
- correctly preparing data for the selected DM technique/s, if required assumptions do not hold

In (Witthen et al. 1993) a review of selected methods of machine learning with an emphasis on practical applications is presented, together with suggestions on how they might be used to address some important problems in the primary production industries, particularly agriculture.

Also, particular efforts in post processing the results directly provided by a DM technique are important in this context, in order to make these results directly understandable by an environmental scientist, who has to make real decisions upon them, which will surely have a real impact to the evironmental system behaviour. In fact, it can be said that the quality of the decisions will depend, not only on the quality of the data mining results themselves, but also on the capacity of the system to communicate the relevant results to the decision-maker as understandably as possible. The software tools used for applying DM techniques to a data set usually produce long listings plenty of results that may be useful in whatever particular application is performed. However, given a particular real case, not all this information is useful. Thus, it is important to:

- Identify the relevant information from the software outputs, depending on the aims of every particular analysis.
- Find the best way to present the selected results to the user.

It can be said that pre and post processing are one of the most important and critical parts of the whole KDD process. On the one hand, because *data cleaning* [Moore et alt, 1993], transformation, selection of DM techniques and optimization of parameters (if required) are often time consuming and difficult, mainly because the approaches taken should be tailored to each specific application, and human interaction is required; on the other hand, because the correctness of the DM itself is critically depending on the quality of the data and wrong or poor preprocessing may lead to incorrect results. Data miners should become conscious of the importance of performing very careful and rigorous preprocessing, and allocate sufficient time to this activity. In fact, once those tasks have been accomplished, the application of DM algorithms becomes trivial and can be automated, requiring only a small proportion of the time devoted to the whole KDD process. Regarding the post processing, the results provided by the software implementing the DM techniques, select of relevant information from the automatic outputs produced by the different software applications, and choose the proper way of transforming or synthesizing it to make directly understandable to the user is also critical to provide good decision support. This task is also difficult to standardize and time consuming and has to be designed ad-hoc for every particular application, requiring much human guidance. In real applications, the time devoted to both pre and post-processing is rarely below 70% of the time for the whole KDD process.

In this paper issues related to pre and post processing in environmental applications will be addressed, together with the most popular solutions for the different cases. This paper does not pretend to be exhaustive, but to provide tools to environmental scientists for addressing the most common problems arisen in pre and post processing, as no clear methodology for tackling these two steps of KDD process has no been established yet.

2. PREPROCESSING IN ENVIRONMENTAL SYSTEMS

As previously stated, environmental data often includes measurement errors (from automatic monitoring, sensors, etc), uncertainty, imprecision, multi-scalarity, heterogeneity, non-linearities, non stacionariety, non-normality, and tools are required to:

- better understand the data set
- detect imperfections in data sets and manage them in the proper way
- correctly prepare data for the selected DM technique/s, if required assumptions do not hold

This section is devoted to providing elements for pre-processing environmental data for KDD. Pre-processing ranges from the simplest descriptive techniques to the more sophisticated data analysis methods, depending on the nature of data and the goals of the

analysis itself. Authors think that most of the operations performed in a pre-processing step can be reduced to two main families of techniques:

- Detection techniques: Those oriented to detect imperfections in data sets or to verify the accomplishment of required assumptions for a particular analysis:
    - Outlier detection
    - Missing data detection
    - Influent observations detection
    - Normality assessment
    - Linearity assessment
    - Independence assessment
- Transforming techniques: Those oriented to perform transformations in the data set in order to correct the imperfections detected before, or to achieve the technical conditions to apply a certain analysis technique.
    - Outlier treatment
    - Missing data imputation
    - Dimensionality reduction techniques or data projection techniques
    - Creation of new transformed variables
        - Standardization
        - Aggregation
        - Transformation (logarithmic, quadratic)
        - Discretization
        - Recodification
    - Filtering
    - Resampling

Classically, statistics provided a wide set of possibilities to preprocess data. The global process of transforming a raw data set to a correct one ready for analysis is called *data cleaning* [Moore et alt, 1993]. Besides the classical statistical techniques for *data cleaning*, inductive techniques are an alternative for several activities of the environmental scientist, when analytical/traditional methods fail, are too slow, or simply do not exist. Finally, visualization techniques also play an important role in the correct preprocessing of data. In the following sections different situations and possibilities are addressed.

2.1. Visualization

Visualization is a powerful strategy for leveraging the visual orientation of sighted human beings. Sighted humans are extraordinarily good at recognizing visual patterns, trends and anomalies; these skills are valuable at all stages of the KD Miller (in press, to appear in 2007). For example, the presence of outliers, missing values, or errors are typical pre- and post-processing KDD tasks where visualization techniques can be valuable.

Graphs commonly used for classical exploratory visualization, like boxplots, histograms, time series plots or two-dimensional scatter plots, perform poorly considering the great number of variables involved in environmental datasets, along with their complex interrelations, and spatial-temporal references. Thus, more sophisticated visualization methods are required, as for example:

- Distributional plots,
- Three, four, and five dimensional plots (color and symbols may be used to represent the higher dimensions),
- Dimension scaling, for example log scales,
- Rotatable frames,
- Animation with time and interactive graphs,
- Geo-referenced visualizations and maps.

Most DM packages, as Weka, include visualization tools, while more advanced features are provided with wide-spread tools such as Matlab or a dedicated data language such as IDL

or the CommonGIS tool (Andrienko and Andrienko 2004). Reader is also pointed to dedicated visualization tools such as XGobi (Swayne et al. 1998). Visual representations are extremely effective, and may convey knowledge far better than numerical or analytical forms. They should be always considered in environmental KDD.

2.2. Outlier Detection

Outliers are objects with very extreme values in one or more variables (Barnett and Lewis 1978). Graphical techniques were once the most common method for identifying them, but increases in database sizes and dimensions have led to a variety of automated techniques. The use of standard deviations is possible when and only when considering a single variable that has a symmetric distribution, but outliers may also take the form of unusual combinations of two or more variables. The data point should be analyzed as a whole to understand the nature of the outlier and multivariate approach is required.

The treatment will depend on the nature of the outlier (error, member of another population, intrinsic extreme value, etc). The influence of outliers can dramatically affect the results or certain methods, a concern which should feature the choice of tools used throughout the rest of the process. See Moore and McCabe (1993) for an interesting discussion on the dangers of simply eliminating rows with outliers:

> *"In 1985 British scientists reported a hole in the ozone layer of the Earth's atmosphere over the South Pole. [...] The British report was at first disregarded, since it was based on ground instruments looking up. More comprehensive observations from satellite instruments looking down had shown nothing unusual. Then, examination of the satellite data revealed that the South Pole ozone readings were so low that the computer software [...] had automatically suppressed these values as erroneous outliers! Readings dating back to 1979 were reanalyzed and showed a large and growing hole in the ozone layer [...] suppressing an outlier without investigating it can keep valuable out of sight."* (Moore and McCabe 1993).

2.3. Missing Value Treatment

Sometimes, a number of cells are missing from the data matrix. These cells may be marked as a *, ?, NaN (Not a Number), blank space or other special character or special numeric code such as 99999. The latter can produce grave mistakes in calculations if not properly treated. It is also important to distinguish between random and non-random missing values (Allison 2002; Little and Rubin 1987). Non-random missing values are produced by identifiable causes that will determine the proper treatment, also influenced by the goals of the task. Imputation (see Rubin 1987) is a complex process for converting missing data into useful data using estimation techniques. It is important to avoid false assumptions when considering imputation methods, which may have a significant effect on the results extracted. All the methods have pros and cons, and the choice must be made with care. In particular, removing rows with missing cells from a dataset may cause serious problems if the missing values are not randomly distributed. It is of utmost importance to report any elimination performed.

However, many data-mining algorithms can treat missing values. Such algorithms should be preferred, in cases of scarce datasets.

A method proposed in Gibert et al.(2008b) is to perform a previous clustering with complete variables and to impute missing values locally to the discovered clusters. Such an approach may even help to identify errors in the data set (if some observation places in a wrong cluster and the reasons for that are investigated). Another method proposed in Athanasiadis and Mitkas (2007) was to use qualitative indicators for substituting missing or erroneous air quality measurements. In operational systems, instead of making numerical estimates of a missing records, data-mining techniques may be used for making qualitative

estimates (i.e by assigning quality labels instead of predicting values). Similar work has been done for substituting missing or erroneous values with fuzzy sets, probability distributions or confidence intervals.

2.4. Pre-processing imbalanced data sets

Imbalanced sets appear in many real applications areas, where the number of instances of one class (usually the normal class) is much bigger than the number of instances of the other classes (usually abnormal classes), which frequently are the most important ones, since they are the object of interest in the specific problem. Some applications in which imbalanced data sets typically appear are the identification of anomalies in certain systems designed to work under steady conditions, such as Water Distribution Systems [Izquierdo et al., 2007], the detection of oil spills from satellite images [Kubat et al., 1998], the identification of power distribution fault causes [Xu et al., 2007] and the prediction of pre-term births [Grzymala-Busse et al., 2003]). This issue is growing in importance since it appears more and more in most real domains of classification, especially in systems where normal data are abundant while abnormal ones are scarce.

Most classical machine learning algorithms generally perform poorly on imbalanced data-sets because they are designed to minimize the global error rate [Japkowicz and Stephen, 2002] and in this manner they are biased toward the majority class, that is to say, they tend to classify almost all instances within the majority class, and thus poorly classify the minority class examples.

Two kinds of solutions can be used to cope with imbalanced sets: pre-processing solutions, trying to balance data by over-sampling the minority classes, under-sampling the majority one or a combination of both (see [Batista et al., 2004; Chawla et al., 2002; Guo and Viktor, 2004]), and solutions at the algorithmic level, modifying the cost per class [Provost and Fawcett, 2001], adjusting the probability estimation by establishing a bias towards the minority classes [Weiss and Provost, 2003], etc.

In [Fernández et al., 2008] different pre-processing mechanisms are used in conjunction with a Fuzzy Rule Classification System to deal with imbalanced data sets.

As an example, in [Izquierdo et al., 2007] a complex hybrid model, which uses a calibrated classical model of a Water Supply System and a neuro-fuzzy technique, is used to obtain diagnosis of leaks and other anomalies in the system. To train the neural network both normal and abnormal data are needed. While normal data are abundant, abnormal ones are scarce since only a reduced number or records from reported anomalies are usually available. The calibrated classical model of the hydraulic network is used to simulate abnormal data, thus producing a data set with enough fuzzy examples to correctly train the neural network.

2.5. Uncertainty in environmental data

Most of the approaches used in KDD assume precise data. They assume that we deal with exact measurements. But in most, probably in all real-world scenarios, always imprecise measurement is usually obtained. There is always a degree of uncertainty. Even being able to measure a magnitude, its exact value will be never known. It can only be known that the measurement is somewhere in a certain range, bounded by the precision of the measurement instrument or measurement procedure itself.

The multiplicity of factors involved in environmental processes provokes uncertainty in data. A number of issues contribute to this complexity. Among others (i) a huge number of sensors are involved, (ii) knowledge related to the spatial structure of these sensors is not well known, (iii) different types of variables can affect the same sensor, (iv) one specific

variable can exhibit simultaneously different states. In addition, subjectivity is often influential and fuzziness becomes important.

Tracking and reporting of uncertainties related to measurement and other sources of noise is an area that is sometimes not treated rigorously, despite the implications. Therefore, the minimum theoretically achievable error of any model built on the data cannot be less than the error contained in the original data. Models with reported fit greater than this are overfitted and their performance measures do not reflect true predictive capacity. In general, as much data as possible should be available, there is less uncertainty, or at least that uncertainty can be better quantified.

To be exact, all recorded data involves uncertainty in some neighborhood, and the real measured values are really somewhere inside a certain interval, with a width depending on the accuracy of the measurement procedure. It has to be noted that we are not speaking about a probability, but about some kind of likelihood that certain crisp value is being obtained [Zadeh, 1995]. Thus, the use of single values (means, for example, or maximum and minimum values), although it implies a quick and sometimes convenient way of calculation or completion of registers with missing values, it is often simplistic and it should be corrected with appropriate safety values. The statistical approach of subjectively attributing probabilities to the inaccuracies many times gives the problem an artificial character, involving a high degree of randomness and runs the clear risk of inventing information about unknown distributions.

An appropriate conceptual tool for this type of data is the theory of fuzzy sets [Zadeh, 1965]. Many environmental research works make use of fuzzy logic to model uncertainty (see [Bazartseren et al., 2003; Juang, 2003; Oh and Pedrycz, 2004; Kuncheva et al., 2000; Faye et al., 2003; Izquierdo et al., 2006], among others).

2.6. Transformation and Creation of Variables

Sometimes transformation of variables may assist analysis. For example, normality may be forced when using ANOVA or, for ease of interpretation, variables with a large number of categorical labels can be grouped according to expert knowledge. Under some circumstances, discretization of continuous variables is appropriate (eg *Age* into *Child* under 18 years, *Adult* between 18 and 65 years, *Elderly* over 65 years). In fact, in real applications it is quite common to globally discretize any numeric attributes before applying learning algorithms to datasets, since a number of them cannot handle numeric attributes directly. In these cases prior discretization is essential. Even if it can, prior discretization often accelerates induction, and may produce simpler and more accurate classifiers. As it is generally done, global discretization denies the learning algorithm any chance of taking advantage of the ordering information implicit in numeric attributes. However, a simple transformation of discretized data preserves this information in a form that learners can use. In (Frank et alt, 99) it is shown that, compared to using the discretized data directly, this transformation significantly increases the accuracy of decision trees built by C4.5, decision lists built by PART, and decision tables built using the wrapper method, on several benchmark datasets. Moreover, it can significantly reduce the size of the resulting classifiers. This simple technique makes global discretization an even more useful tool for data preprocessing.

Noise is often a critical issue, and especially with environmental data some bias may exist that can be removed with a filter. Transformations should always be justified and documented, and the biases that may be introduced noted (Gibert and Sonicki 1999). Interpretability of transformed variables should be kept.

Creation of additional variables is also used in KDD. Here, expert knowledge is usually the guide. Exploratory variable creation without such assistance is almost always prohibitively time consuming, and as noted, may obfuscate physical interpretation and exacerbate noise. Efficient techniques for data reduction, however, do exist and are well used.

However, avoidance of unnecessary transformations is recommended, especially if the transformation decreases interpretability (for example Y = log(streamflow), although Y is normal). If transformations are definitely required, some bias may be introduced into the results; thus, it is convenient to minimize arbitrariness of the transformation as much as possible (in recoding *Age*, Adult may be defined from 18 to 65 or from 15 to 70), and this implies that the goals of the analysis must also be taken into account. For arithmetic transformations, imputation of missing data before the transformation is thought to be better. Note that where data is numerical and the scales changes between variables, normalization may be necessary.

Finally, a good practice is to select a data mining technique that fits well on the target phenomenon and the kind of available data, instead of adopting the approach of strongly transform the data till it fits on the technical assumptions of the preferred data mining technique, in spite of loosing interpretability, or introducing arbitrariness. This means that if data is non-normal, it is better to see, first of all, if there exist a data mining technique suitable for this purpose which does not requires the normality to be applied; or this means that an alternative to ID3 should be used if data is numeric, instead of forcing discretization.

2.7. Data Reduction and Projection

When the number of variables is too high to deal with in a reasonable way, which is not unusual in data mining context, a data reduction method can be applied. Either Data Projection or Feature Selection methods are suitable possibilities in these cases.

2.7.1. *Data projection*

This may be accomplished by eliminating some variables wholesale, or projecting the feature space of the original problem into a reduced fictitious space, with fewer dimensions. Principal Components Analysis (PCA) (Dillon and Goldstein 1984) is one of the best known techniques used for the latter purpose. Each principal component is a linear combination of the original variables, and the aim is to work with a reduced set of these, such that the loss of information is not relevant. Thus, Principal Component Analysis is suitable for synthesizing an original set of numerical variables into a small number of fictitious variables conserving as much information as possible from original dataset. Equivalent techniques are available for qualitative data, like multiple correspondence analysis (Lebart et al. 1984; Dillon and Goldstein 1984). However, it has to be taken into account that in most cases, interpretation of the new variables (or factors) may not be clear, and if this is the case, there will be implications for understandability of the final results.

2.7.2. Feature Selection and Feature Weighting

Datasets may contain irrelevant or redundant variables (Gibert et al. 2008). As previously stated, the quality of discovered knowledge is usually dependant on the quality of the data that they operate on. Also, the success of some learning schemes, in their attempts to construct models of data, hinges on the reliable identification of a small set of highly predictive attributes. The inclusion of irrelevant, redundant and noisy attributes in the model building process phase can result in poor predictive performance and increased computation.

Feature subset selectors are algorithms that attempt to identify and remove as much irrelevant and redundant information as possible prior to learning or knowledge discovery. Feature subset selection can result in enhanced performance, a reduced hypothesis search space, and, in some cases, reduced storage requirement. Automated techniques for identifying and removing unhelpful or redundant variables usually take one of two forms: direct examination of the relevance of candidate variables, or searching the best combination of attributes in terms of model performance and feedback. The former are called *filters* and the latter *wrappers* (see Hall 1999 for details). For a survey of common

feature selection techniques, see Molina et al. (2002). Usually, analyzing the feature subset selection provides better results than analyzing the complete set of variables (Andrew 1998).

However, attribute selection generally involves a combination of search and attribute utility estimation plus evaluation with respect to specific learning schemes. This leads to a large number of possible permutations where very few benchmark studies have been conducted. In (Andrew et al. 2003) a benchmark comparison of several attribute selection methods for supervised classification is presented. All the methods produce an attribute ranking, a useful devise for isolating the individual merit of an attribute. Attribute selection is achieved by cross-validating the attribute rankings with respect to a classification learner to find the best attributes. Results are reported for a selection of standard data sets and two diverse learning schemes C4.5 and naive Bayes.

Other techniques are based on *feature weighting* (see for example Aha 1998 and Núñez et al. 2003), which is a more general and flexible approach than feature selection. The aim is to assign a degree of relevance (a weight) to each attribute. Similarities (or dissimilarities) become emphasized according to the relevance of the attribute, and irrelevant attributes will not influence the results, so quality of inductive learning improves.

Feature weight assignment is frequently used to denote the relevance of attributes in similarity computations. When some attributes are irrelevant for the prediction task, the appropriate weight learning could improve the data mining process [Aha, 1998]. Empiric works [Wettschereck et al., 1997] and theoretical ones [Langley and Iba, 1993], suggest that the learning complexity is exponential regarding the number of irrelevant attributes. Therefore, the failures in the data mining process could be related to a similarity model, and in particular, with an incorrect weight assignment methodology.

In recent years, a great deal of research research works has been done in feature weight assignment. The main goal is to assign high weights to attributes that are identified as relevant, and at the same time, to assign low weights to those that are irrelevant. Most of the methods of weight assignment use a global scheme, that is, they associate a weight to the whole space of the attribute. In [Wettschereck et al., 1997] a conceptual framework for the weight assignment methods classification is presented, considering bias from the performance algorithm (wrapper) [Kohavi and John, 1998] or not (filter), previous transformations of data before analyzing relevance, if the relevance of an attribute is invariant or not all along the domain, or if they require domain specific knowledge. A comprehensive review of feature weighting algorithms can be found in [Núñez, 2004].

Although the weight assignment improves the accuracy in classification and retrieval tasks, feature selection is vital to reduce the dimensionality in learning tasks, completely eliminating irrelevant attributes (Martorell et al.2007). In general, feature weighting is more appropriate for tasks where features vary in their relevance, but such methods search larger spaces of weight assignments. On the other hand, feature selection algorithms perform better when the features used to describe instances are either highly correlated with the class label or completely irrelevant.

2.8. Hybrid approach

Frequently, data-driven (either stochastic or connectionist, among others) models are accused of lack of understandability (black boxes) because of the substantial drawback they exhibit: they are synthesized ONLY on the available data, with no detailed information of the underlying process. As a consequence, in prediction and extrapolation tasks and especially when data are noisy or sparse, they sometimes are inadequate and inaccurate. It is thus reasonable to argue that DM techniques should be used not only to replace

knowledge-based models but also to complement them and produce so-called hybrid models. Integration of both deterministic (numerical) and stochastic (data-driven), or logic (qualitative) models presumably should provide more accurate predictions and cope better with uncertainty, non-normality, non-linearity, or even with the original nature of data. It appears clearly more sensible to use the already available deterministic information given by the very well-known theory-driven models than stubbornly starting from the scratch, throwing away all knowledge, trying to use a data-driven model alone. Also, it seems better to take advantage of the precision of a numerical variable than discretizing all numerical ones to use a machine learning method for qualitative variables. This approach should also appear more appealing to scientists that have been working for years with theory-driven models who will be reluctant to sink all their *precious* knowledge into oblivion. From a conceptual point of view, this approach takes into account that (environmental) complexity comes not only from the system but also from outer perturbations producing stochastic influences (noise). A number of papers using hybrid modeling in different areas, in particular environmental modeling, can be quoted (Espert et al. (1999), Vojinovic et al. (2003), Krasnopolsky et al. (2006), Chercassky et al. (2006), Izquierdo et al. (2006), Izquierdo et al. (2008)). It can be foreseen that computational intelligence (machine learning) will be used not only for building data-driven models, but also for building optimal adaptive model structures of such hybrid models. Combining Artificial Intelligence techniques with Statistical ones can also improve the quality of discovered knowledge (Gibert et al.2005a, Pérez-Bonilla et al.2007, Gibert et al.2007).

3. THE ROLE OF POST PROCESSING

Apart from the important role of preprocessing, together with the correct selection of the data mining technique which will really answer the target questions, there is an important job to be done between getting the results of the data mining techniques and using them to support decision-making: to *understand* the results.

Indeed, the quality of decisions taken upon KDD results, depend not only on the quality of the results itself, but on the capacity of the system to communicate those results in an understandable form to the decision maker.

The software tools used for applying the DM techniques to a data set use to produce long listings plenty of results that may be useful in any particular application. The closer to statistical packages the software, the longer and more complex the output with more numerical information displayed. However, given a particular real case, not all this information is useful. Moreover, the major part is irrelevant. Thus, it is important to:

- Identify the relevant information from the software outputs, depending on the aims of every particular analysis. For example, from a regression analysis, it may be irrelevant to the environmental expert to know the exact value of h_i indexes, but from this information, the set of influential observations that have to be carefully analyzed should be reported.
- Find the best way to present the selected results to the user in such a way that it becomes directly understandable, given that the final user does not know the technical details of the Data Mining method used. So, probably, from the results of a logistic regression, it is more interesting to provide the interpretation of the estimated coefficients rather than the logistic equation itself.

As an example, consider an application where clustering is applied as the most suitable data mining technique, regarding the goals of the analysis. Most of the software implementing clustering algorithms provides information about the number of clusters discovered, and the set of objects belonging to every one of the clusters. Upon these results, the experts need to understand the underlying clustering criteria as well as the meaning of the classes themselves. In data mining contexts where the number of classes increases, and the number

of variables is high, tools that help the user to postprocess the clustering results till the conceptualization of the classes are very useful. In (Gibert et al.2005b) the *Class Panel Graph (CPG)* is presented as an integrated graphical tool that provides a perspective of the whole data set regarding the previously discovered classes, in such a way that identification of variables with particular behaviours in every class is easy. Figure 2 shows the aspect of a CPG which can display either histograms conditioned to the classes, or boxplots (and bar charts for qualitative variables).

The CPG supports the interpretation of the results as well as the process of conceptualizing the classes. The authors are not aware that commercial software offer facilities to show in a single integrated graph the behaviour of as many variables as possible. Also, in Pérez-Bonilla (2007) a methodology that automatically induces concepts from classes is presented (CCCS) in such a way that the interpretation of classes performed by the expert becomes easier on the basis of some preliminary concepts suggested by the system. Either the CPG as the CCCS methodology are implemented in the KLASS software, which will be presented below. Also, in a clustering context, it may be useful to display the prototypes of the classes in a 2-D or 3-D scatterplot of numerical variables (Gibert et al.2006), as implemented in the GESCONDA software, also described below.

As another case, let us consider here the classification tasks. Among the algorithms solving those taks, those that inductively construct decision trees are found. Other approaches using neural networks, evolutionary algorithms or Bayesian networks are also applied to solve the classification problem. Most of these algorithms work as black box approaches. As a result, it is often difficult to understand the decision model.

Visualization techniques can help to overcome these problems. One of them, for example used by SGIs MineSet sytem (Brunk et al., 1997), shows a scheme of the decision tree and allows the user to select among important parameters of the model. The user interactively can select different attributes and understand the model. In (Ankerst et al., 2000) a more sophisticated approach is used: each attribute value is shown by a colored pixel and all of them are arranged in bars; then the pixels corresponding to each attribute bar are sorted separately and the attribute exhibiting the purest value distribution is chosen to split the decision tree; the process is iteratively repeated until all leaves correspond to pure classes. In addition, information like number of training patterns corresponding to one node, purity of partitions, etc. can complete the picture.

In general, many tools are available to post process the results of many different data mining techniques, such as decision trees, neural nets, statistical modelling, etc. This paper do not pretends to be exhautive, but to provide an overview of the importance of inserting this kind of tools inside the methodology to make knowledge discovery more fruitful.

4.1. Weka

The Weka workbench (Witten and Frank 2005) contains a collection of visualization tools and algorithms for data analysis and predictive modelling, together with graphical user interfaces for easy access to this functionality. A command line interface is also included, for larger scale processing. It was originally designed as a tool for analyzing data from agricultural domains but is now used in many different application areas, largely for educational purposes and research. The main strengths of Weka are that it is (a) freely available under the GNU General Public License, (b) very portable because it is fully implemented in the Java programming language and thus runs on almost any computing platform, (c) contains a comprehensive collection of data preprocessing and modelling techniques, and (d) is easy to use by a novice due to the graphical user interfaces it contains.

Figure 2: Example of CPG where it can be seen that group T7 contains elements with lower values on variable Q-E (inflow of a WWTP), group T15 contains elements with greater values of DQO-E (chemical organic matter).

Pre-processing in Weka is mainly achieved with filters. The idea is to pass data through the filter in order to affect some form of transformation. Two general types of filter are provided, supervised and unsupervised. Supervised filters use class information to affect the transformation while unsupervised filters are class-blind. Beyond that filters work on data either at the instance (ie example level) or the attribute level. Thus there are four types of filter:

Supervised attribute: providing operations for selecting the attributes more correlated with a class (Attribute Selection) or some coercions (Discretize, NominalToBinary) or reordering (ClassOrder). Also PLSfilter is a specialised filter particularly useful in spectral applications where attribute co-linearity is expected.

Supervised instance: These filters mainly support experimentation or down-sampling of a problem (Resample, SpreadSubsample or StratifiesRemoveFolds). They work on the entire set of instances in a dataset.

Unsupervised attribute: The bulk of the available filters in Weka are of this type, including operations to add columns to the data set, either with a new attribute (Add, Copy, ClusterMembership, FirstOrder), an identifier (AddID) or a new partition (AddCluster) or combining existing attributes (AddExpression), or transforming them (Normalize, Standardize, Discretize, PKDiscretize, MergeTwoValues, NumericToBinary, NumericToNominal, NominalToBinary, NumericTransform, StringtoNominal, StringToWordVector, Center, Wavelet, ChangeDateFormat, SwapValues, MathExpression, Obfuscate), to change some of the rows of a given column (NumericCleaner –for extreme values, AddNoise, AddValues, ReplaceMissingValues) or to mark some rows of a given column (Interquartile Range). Also, operations to eliminate some columns of the data set (Remove, RemoveType, RemoveUseless) or operations for creating derivative datasets from the original one (KernelFilter, MakeIndicator, PartitioneMultifilter, RandomProjection, TimeSeriesDelta, TimeSeriesTranslate,)

Unsupervised instance: Including filters to Normalize variables, transform to a sparse format NonSparseToSparse, produce random subsamples with or without replacement Resample, randomly reorder the set of instances (Randomize), or to remove subsets of instances under different criteria (RemoveFrequentValues of a nominal attribute, RemovePercentage, RemoveRange, RemoveWithValues, RemoveFolds)

4.2. Preprocessing with SPSS

Pre-processing techniques in SPSS range from plain detection and correction of data errors made at the introduction stage, to sophisticated transformations, including re-codification performed on one or several variables or the creation of new variables from other already existing ones. SPSS provides the nice possibility of apply transformations and re-codifications either to the whole set of data or to a predefined subset (by using the **select cases** option) Also, sometimes it is necessary to change the order, to merge different files or to select specific cases for analysis. Another need, which is frequent in the case of working with real data bases, comes from the fact that files are not suitably organized. SPSS provides tools for all these tasks.

The menu **Transform**, in the SPSS's main menu bar, includes a number of options allowing a number of transformations:

- **Computing variables** allows calculating values for a variable based on numeric transformations of other variables.
- **Count Occurrences** creates a variable that counts the occurrences of the same value(s) in a list of variables.

Recoding values allows data values modifications by recoding them. It is particularly useful for collapsing or combining categories. Values can be recoded into existing variables or, alternatively, new variables can be created to recode values of existing variables.
Rank cases allows creating new variables containing ranks, normal scores and percentile values for numeric variables. The **Automatic Recode** option allows converting string and numeric values into consecutive integer numbers.
Missing values in functions make it possible to treat missing values in different ways.

Data files are not always organized in the ideal form for some specific needs. A wide range of file handling and transformation capabilities is available. Among them:
Sort cases that sorts rows based on the value or one or more variables.
Transpose cases and variables creates a new data file with rows into columns and vice versa.
Aggregate data aggregates groups of cases in the active dataset into single cases and creates a new file or new variables in the active data set.
Select subsets of cases allows the analysis to be restricted to a subset of cases or perform simultaneous analyses on different subsets.
Weight data weights cases for analyses based on the value of a weight variable. It is particularly useful to build contingency matrices.
Restructure data can create a single case from multiple cases or create multiple cases from a single one.

Some other additional data preparation features can be found in the **Data Preparation** (*Data Validation* in older versions) module. It allows creating reports, charts and analyses without additional preliminary work. Among other tasks, one can determine how certain values should be treated by assigning variable properties that describe the data, identify anomalies such as missing values, outliers, duplicate information, or create new variables with a lower number of categories to represent ranges of values from variables with larger number of possible values.

4.3. Preprocessing and postprocessing in GESCONDA

GESCONDA (Gibert et al.2006) is the name given to an Intelligent Data Analysis System developed with the aim of facilitating Knowledge Discovery (KD) and especially oriented to environmental databases. On the basis of previous experiences, it was designed to include extensive preprocessing tools: data cleaning, missing data management, outlier analysis and treatment, statistical univariate analysis, statistical bivariate analysis, visualization tools, attribute or variable transformation facility, including discretizations, recodifications and creations of new variables, feature weighting for supervised and unsupervised data sets (Gibert et al.2006b). Portability of the software between platforms is provided by a common Java platform.

GESCONDA contains a postprocessing module including tools for validation of the results of different data mining techniques, such as results from clustering (scatter plot of pairs of numerical variables, marked by class and superimposing the position of class prototypes, rates of missclassification if a reference partition is available), rule induction (eliminating low precision rules, evaluate an inducted rules base over a test set and estimate the quality of the rule base), etc.

integration of different knowledge patterns for a predictive task, or planning, or system supervision, together with AI and statistics mixed techniques, consideration of knowledge use by end-users.

4.4. Pre and post processing in KLASS

KLASS (Gibert et al.2005b) is a software package originally conceived for Knowledge Discovery (KD) in real domains with complex structure (Gibert et al.1999). It provides a mixture of statistical and artificial intelligence tools to support KD, including basic statistics and providing an integrated system to support the whole process of KDD including pre and post processing, provided that the main data mining technique to be used is related with clustering or rule induction.

Regarding preprocessing, KLASS offers functionalities for basic statistics (simple or by groups), histograms, boxplots (side-by-side), (letter)plots, cross-tables. The performance of the system is quite high, since the user has control over many parameters of the graphics (like the number of classes of a histogram, or the limits of the axis of a plot), providing a very flexible tool. It also offers a complete module of data management, including missing data treatment, creation of transformed variables either using mathematical expressions or via recodification or discretization (here the Boxplot based discretization is provided, which discretizes the numerical variable in such a way that the resulting qualitative variable maximizes association degree with a previously discovered class variable, Pérez-Bonilla et al.2007). Construction of a prior expert knowledge base (which can be non-complete) is also available and it can be used to bias a posterior clustering process, by means of the Clustering Based on Rules option (Gibert et al.1999), in such a way that the final classes hold the semantic constraints expressed by the rules.

Regarding the postprocessing, Klass offers some interesting tools to support the interpretation of a clustering results, apart from the classical representation of the dendrogram; It also provides the Class Panel Graph (Gibert et al.2005), which is a very interesting possibility in clustering contexts to understand better the meaning of the classes. It also implements the CCCS methodology (Pérez-Bonilla 2007) for assigning concepts to every class, improving even more the support to the understanding of the results. There is also a function for visualizing knowledge bases, containing probabilities or not, and selecting the rules with degrees of certainty over a certain threshold.

One of the particularities of the system is that it is designed in such a way that the outputs, either graphical or numerical or textual, are produced in LaTeX font files, which are directly processed by the kernel of KLASS and automatically sent to the LaTeX viewer and displayed on the screen. From the final user point of view, this makes no difference with other systems, since graphical representations are directly displayed on the screen as well as other results. However, as reporting the results of the KDD process is always involved with the elaboration of technical papers, KLASS also includes a reporting facility in such a way that the user can specify a set of steps to be performed sequentially and a single big LaTeX document including all the results is produced. The user only needs to edit this document and add personal comments on it to get a complete report of the analysis. KLASS provides either standard or personalized reports. It is a flexible possibility since it is possible to automatically transform every result of the single steps into PostScript or PDF documents, which can be managed as usual, for example, pasting it into a Word document.

If the document to be produced is long, with a complex structure and contains hard mathematical notation, LaTeX offers nice advantages and the LaTeX results provided by KLASS are really useful. In this case, LaTeX is a widely used text processor, owing to the excellent support it provides to the generation of high quality mathematical formulae and scientific notation. However, including graphical representations from commercial statistical packages in a LaTeX document requires the use of special LaTeX packages to deal with graphical formats and makes a little bit more complicate the elaboration of the document, which requires transformation to PostScript or PDF to be completely visualized. Since the results of KLASS are produced in native LaTex code, inclusion of those graphics in the final report becomes trivial.

On the other hand, making the native LaTeX code accessible to the user permits the user to adjust labels or size of the titles of graphical representations. In this way, the quality of the image is maintained to its final use.

ACKNOWLEDGEMENTS

The project TIN2004-01368 has partially financed the development of GESCONDA.

REFERENCES

Aha, D. 1998. Feature weighting for lazy learning algorithms. In: Liu, H., Motoda, H. (Eds.), *Feature Extraction, Construction and Selection: A Data Mining Perspective*. Kluwer.

Allison, P. 2002. *Missing Data*. Sage, Thousand Oaks, CA, USA.

Almasri, M. and Kaluarachchi, J. 2005. Modular neural networks to predict the nitrate distribution in ground water using the on-ground nitrogen loading and recharge data. *Environmental Modelling and Software*, 20(7): 851-871.

Mark Andrew Hall and Geoffrey Holmes. Benchmarking attribute selection techniques for discrete class data mining. IEEE Transactions on Knowledge and Data Engineering, 15(3):1437-1447, November/December 2003.

Mark Andrew Hall and Lloyd Smith (1998): Practical feature subset selection for machine learning. In Proc 21st Australian Computer Science Conference, pages 181-191, Perth, Australia. Springer.

Andrienko, G. and Andrienko, A. 2004. Research on visual analysis of spatio-temporal data at fraunhofer ais: an overview of history and functionality of commonGIS. In: *Proceedings of the Knowledge-Based Services for the Public Services Symposium, Workshop III: Knowledge Discovery for Environmental Managment*. KDnet, pp. 26-31.

Ankerst, M., Ester, M., and Kriegel, H. (2000). Towards an effective cooperation of the computer and the user for classification. In *SIGKDD Int. Conf. on Knowledge Discovery & Data Mining (KDD 2000)*, Boston, MA, 179-188.

Athanasiadis, I., Kaburlasos, V., Mitkas, P. and Petridis, V. 2003. Applying machine learning techniques on air quality for real-time descision support. In: *Information technologies in Environmental Engineering*.

Athanasiadis, I., Karatzas, K. and Mitkas, P. 2005. Contemporary air quality forecasting methods: A comparative analysis between statistical methods and classification algorithms. In: *Proceedings of the 5th International Conference on Urban Air Quality*.

Athanasiadis, I. and Mitkas, P. 2004. Supporting the decision-making process in environmental monitoring systems with knowledge discovery techniques. In: *Proceedings of the Knowledge-Based Services for the Public Services Symposium, Workshop III: Knowledge Discovery for Environmental Managment*. KDnet, pp. 1-12.

Athanasiadis, I.N. & Mitkas, P.A. (2007). Knowledge discovery for operational decision support in air quality management. Journal of Environmental Informatics. 9:2, 100-107

Babovic, V. 2005. Data mining in hydrology. *Hydrological Processes*, 19:1511-1515.

Barnett, V. and Lewis, T. 1978. *Outliers in Statistical Data*. Wiley.

Batista, G., Prati, R. and Monard, M. A study of the behaviour of several methods for balancing machine learning training data, *SIGKDD Explorations* 6 (1), 20-29, 2004.

Bazartseren, B. Hildebrandt, G. and Holz, K. P. Short-term water level prediction using neural networks and neuro-fuzzy approach. *Neurocomputing*, 55, 439-450, 2003.

Belanche, L., Valdés, J., Comas, J., Rodríguez-Roda, I. and Poch, M. 2001. Towards a model of input-output behaviour of wastewater treatment plants using soft computing techniques. *Environmental Modelling and Software,* 5(14): 409-419.

Brunk, C., Kelly, J., and Kohavi, R. (1997). Mineset: An integrated system for data mining. In *SIGKDD Int. Conf. on Knowledge Discovery & Data Mining (KDD 1997)*, Newport Beach, CA, 135-138.

Cadez, I. and Smyth, P. 1999. Modelling of inhomogeneous markov random fields with applications to cloud screening. Tech. Rep. UCI-ICS 98-21.

Camargo, S., Robertson, A., Gaffney, S. and Smyth, P. 2004. Cluster analysis of western north pacific tropical cyclone tracks. In: *Proceedings of the 26th Conference on Hurricanes and tropical Meteorology*. pp. 250-251.

Cendrowska, J. 1998. Prism: An algorithm for inducing modular rules. *International Journal of Man-Machine Studies,* 27(4): 349-370.

Chawla, N., Bowyer, K., Hall, L. and Kegelmeyer, W. Smote: Synthetic minority over-sampling technique, *Journal of Artificial Intelligent Research* 16, 321-357, 2002.

Chercassky, V., Krasnopolsky, V., Solomatine, D. P., Valdes, J. 2006. Computational intelligence in earth sciences and environmental applications: Issues and challenges. Neural Networks 19, 113-121.

Cohen, W. 1995. Fast effective rule induction. In: *Proceedings of the Twelfth International Conference on Machine Learning*, Prieditis, A., Russell, S. (Eds.). Morgan Kaufmann, pp. 115-123.

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

Comas, J., Llorens, E., Martí, E., Puig, M.A., Riera, J.L., Sabater, F. and Poch, M. 2003. Knowledge Acquisition in the STREAMES Project: The Key Process in the Environmental Decision Support System Development. *AI Communications*, 16(4): 253-265.

Cortés, U., Rodríguez-Roda, I., Sànchez-Marrè, M., Comas, J., Cortés, C. and Poch, M. 2002. DAI-DEPUR: An environmental decision support system for supervision of municipal waste water treatment plants. In: *Proceedings of the 15th European Conference on Artificial Intelligence (ECAI'2002)*. pp. 603-607.

Dillon, W. and Goldstein, M. 1984. *Multivariate Analysis*. Wiley, USA.

Dixon, M., Gallop, J.R., Lambert, S.C. and Healy, J.V. 2007. Experience with data mining for the anaerobic wastewater treatment process. *Environmental Modelling and Software*, 22: 315-322.

Domingos, P. 1996. Unifying Instance-Based and Rule-Based Induction. *Machine Learning*, 24: 141−168.

Draper, N. and Smith, H. 1998. *Applied Regression Analysis*. Wiley.

Dubes, R. and Jain, A. 1988. *Algorithms for Clustering Data*. Prentice Hall.

Dzeroski, S. and Drumm, D. 2003. Using regression trees to identify the habitat preference of the sea cucumber (holothuria leucospilota) on rarotonga, cook islands. *Ecological Modelling*, 170(2-3): 219-226.

Dzeroski, S., Grbovic, J., Walley, W. and Kompare, B. 1997. Using machine learning techniques in the construction of models. ii. data analysis with rule induction. *Ecological Modelling*, 95(1): 95-111.

Ekasingh, B., Ngamsomsuke, K., Letcher, R. and Spate, J. 2003. A data mining approach to simulating land use decisions: Modelling farmer's crop choice from farm level data for integrated water resource management. In: *Advances in Hydrology: Proceedings of the International Conference on Water and Environment*, Singh, V., Yadava, R. (Eds.), pp. 175-188.

Ekasingh, B., Ngamsomsuke, K., Letcher, R. and Spate, J. 2005. A data mining approach to simulating land use decisions: Modelling farmer's crop choice from farm level data for integrated water resource management. *Journal of Environmental Management*.

Ellis, F. 1996. The application of machine learning techniques to erosion modelling. In: *Proceedings of the Third International Conference on Integrating GIS and Environmental modelling*. National Center for Geographic Information and Analysis.

Espert, V., López, P. A. and Izquierdo, J., 1999. Fundamentals of a water quality model solution for dissolved oxygen in one-dimensional receiving system. Numerical Modelling of Hydrodynamic Systems. Proc. of Intnl. Workshop, 444-445.

Faye, R. M., Sawadogo, S., Lishou, C. and Mora-Camino, F. Long-term fuzzy management of water resource systems. *Applied Mathematics and Computation* 37: 459-475, 2003.

Fayyad, U. and Piatetsky-Shapiro, G., P, S. 1996a. Advances in knowledge discovery and data mining. In: *Data Mining to Knowledge Discovery: an Overview.* American Association for Artificial Intelligence, pp. 1-34.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. 1996b. From data mining to knowledge discovery in databases (a survey). *AI Magazine*, 3(17): 37-54.

Fernández, A., García, S., del Jesus, M.J. and Herrera, F. A Study of the Behaviour of Linguistic Fuzzy Rule Based Classification Systems in the Framework of Imbalanced Data-Sets. *Fuzzy Sets and Systems*, doi:10.1016/j.fss.2007.12.023, 2007

Frank E, Witten I A (1999): Making better use of global discretization. In Proc 16th International Conference on Machine Learning, Bled, Slovenia, pages 115-123. Morgan Kaufmann.

Gatts, C., Ovalle, A. and Silva, C. 2005. Neural pattern recognition and multivariate data: Water typology of the Paraiba do Sul River, Brazil. *Environmental Modelling and Software*, 20(7): 883-889.

Gibbs, M., Morgan, N., Maier, H., Dandy, GC, H. M. and Nixon, J. 2003. Use of artificial neural networks for modelling chlorine residuals in water distribution systems. In: *MODSIM 2003: Proceedings of the 2003 International Congress on Modelling and Simulation*. pp. 789-794.

Gibert, K., Spate, J., Sànchez-Marrè, M., Comas, J., Athanasiadis, I. 2008: Data Mining for Environmental Systems, In State of the art and Futures in Environmental Modelling and Software. IDEA Series (Jackeman, A. J., Rizzoli, A., Voinov, A. and Chen, S. (eds), (in press). Elsevier, Amsterdam, The Netherlands.

Gibert, K., Sanchez-Marre, M. and Rodriguez-Roda, I. 2006. GESCONDA: an intelligent data analysis system for knowledge discovery and management in environmental databases. *Environmental Modelling and Software*, 21:115-120.

Gibert, K., Annicchiarico, R., Cortés, U. and Caltagirone, C. 2005a. *Knowledge Discovery on Functional Disabilities: Clustering Based on Rules Versus Other Approaches*. IOS Press.

Gibert, K., Flores, X., Rodríguez-Roda, I. and Sànchez-Marrè, M. 2004. Knowledge discovery in environmental data bases using GESCONDA. In: *Proceedings of IEMSS 2004: International Environmental Modelling and Software Society Conference*, Osnabruck, Germany.

Gibert, K., Nonell, R., Velarde, JM, Colillas, MM 2005b. Knowledge discovery with clustering: Impact of metrics and reporting phase by using klass. *Neural Network World*, 319-326.

Gibert, K, Rodríguez-Silva, G (2007) Knowledge Discovery in a Wastewater Treatment Plant with Clustering based on Rules by States. In Artificial Intelligence Research and Development. Frontiers in Artificial Intelligence and Applications, v 163, pp 359—367. IOS Press.

Gibert, K, Sànchez-Marrè, M, Comas, J. (2006b) The impact of feature weighting in environmental class discovery using GESCONDA, In BESAI proceedings, ECAI 2006.

Gibert, K., Sànchez Marrè, M. and Flores, X. 2005c. Cluster discovery in environmental databases using GESCONDA: The added value of comparisons. *AI Communications,* 4(18): 319-331.

Gibert, K. and Sonicki, Z. 1999. Clustering based on rules and medical research. *Journal on Applied Stochastic Models in Business and Industry*, formerly JASMDA 15(4): 319-324.

Gibert K, Saxena, K. Morris, J., et al. 2008: The advantages of using clustering for missing inputation, World Health Organization (in press).

Grzymala-Busse, J.W., Goodwin, L.K. and Zhang, X. Increasing sensitivity of preterm birth by changing rule strengths, *Pattern Recognition Letters* 24 (6), 903-910, 2003.

Guo, H. and Viktor, H. L. Learning from imbalanced data sets with boosting and data generation: The databoosting approach, *SIGKDD Explorations* 6 (1), 30-39, 2004.

Guo, Q., Kelly, M. and Graham, C. 2005. Support vector machines for predicting distribution of sudden oak death in california. *Ecological Modelling*, 182(1): 75-90.

Hall, M. 1999. *Feature selection for discrete and numeric class machine learning*. Tech. rep., Department of Computer Science, University of Waikato, working Paper 99/4. URL http://www.cs.waikato.ac.nz/~ml/publications1999.html

Han, J. and Kamber, M. 2001. *Data Mining: Concepts and Techniques*. Morgan Kaufmann.

Hastie, T., Tibshirani, R. and Friedman, J. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag.

Holmes, G., Cunningham, S., Dela Rue, B. and Bollen, A. 1998. Predicting apple bruising using machine learning. In: Proceedings of the Model-IT Conference, *Acta Horticulturae*, 476: 289-296.

Izquierdo, J., Pérez, R., López, P. A., Iglesias, P. L., 2006. Neural Identification of Fuzzy Anomalies in Pressurized Water Systems, Summit on Environmental Modelling and Software, 3rd Biennial meeting of the International Environmental Modelling and Software Society, Proceedings, Burlington (Vt), USA.

Izquierdo, J., López, P.A., Martínez, F.J. and Pérez, R., 2007. Fault detection in water supply systems using hybrid (theory and data-driven) modelling, *Mathematical and Computer Modelling*, 46, 3-4, 341-350.

Izquierdo, J., Montalvo, I., Pérez, R. and Herrera, M., 2008. Sensitivity analysis to assess the relative importance of pipes in water distribution networks, *Mathematical and Computer Modelling*, 48, 268–278.

Japkowicz, N. and Stephen, S. The class imbalance problem: A systematic study, *Intelligent Data Analysis* 6 (5), 429-450, 2002.

Juang, Ch. F. Temporal problems solved by dynamic fuzzy network based on genetic algorithm with variable-length chromosomes. *Fuzzy Sets and Systems* 142: 199-219, 2003.

Kolodner, J. 1993. *Case-Based Reasoning*. Morgan Kaufmann.

Kohavi and John, 1998 R. Kohavi, and G-H. John, The Wrapper Approach, in Feature Selection for Knowledge Discovery and Data Mining, H. Liu & H. Motoda (eds.), Kluwer Academic Publishers, pp33-50. 1998

Kralisch, S., Fink, M., Flügel, W.-A. and Beckstein, C. 2001. Using neural network techniques to optimize agricultural land management for minimisation of nitrogen loading. In: *MODSIM 2001: Proceedings of the 2001 International Congress on Modelling and Simulation*. pp. 203-208.

Krasnopolsky, V.; Fox-Rabinovitz, M. S., 2006. Complex hybrid models combining deterministic and machine learning components for numerical climate modeling and weather prediction. Neural Networks, 19, 122-134

Kubat, M., Holte, R. and Matwin, S. Machine learning for the detection of oil spills in satellite radar images, *Machine Learning* 30 (2-3), 195-215, 1998.

Kuncheva, L. I., Wrench, J., Jain, L. C. and Al-Zaidan, A. S. A fuzzy model of heavy metal loadings in Liverpool bay. Env. *Modeling and Software* 15: 161-167, 2000.

Langley, P., and Iba, W. (1993). Average-case analysis of a nearest neighbor algorithm. Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence (pp. 889-894). Chambery, France.

Larose, D. 2004. *Discovering Knowledge in Data: An Introduction to Data Mining.* John Wiley.

Lebart, L., Morineau, A. and Warwick, K. 1984. *Multivariate Descriptive Statistical Analysis*. Wiley, New York, USA.

Little, R. and Rubin, D. 1987. *Statistical Analysis with Missing Data*. Wiley.

Martín, F.J. and Plaza, E. 2004. Ceaseless Case-based Reasoning. In: *Procc. of 7th European Conference on Case-Based Reasoning (ECCBR'2004)*, pp. 287-301, LNAI-3155, Madrid, Spain.

X. L. Martorell, R. Massanet, K. Gibert, M. Sánchez-Marrè, J. C. Martín-Sánchez, A. Martorell (2007) A Knowledge Discovery methodology for identifying vulnerability factors of mental disorder in an intellectually disabled population. In Artificial Intelligence Research and Development. Frontiers in Artificial Intelligence and Applications, v 163, pp 426—435. IOS Press

Mas, J., Puig, H., Palacio, J. and Sosa-Lopez, A. 2004. Modelling deforestation using GIS and artificial neural networks. *Environmental Modelling and Software*, 19(5): 461-471.

McKenzie, N. and Ryan, P. 1999. Spatial prediction of soil properties using environmental correlation. *Geoderma* (89): 67-94.

Michalski, R. and Chilausky, R. 1980. Learning by being told and learning by examples: An experimental comparison of the two methods of knowledge acquisition in the context of developing an expert system for soybean disease diagnosis. *International Journal of Policy Analysis and Information Systems*, 4(2): 125-161.

Miller, H. J. (in press, to appear in 2007). Geographic data mining and knowledge discovery. In: *Handbook of Geographic Information Science*, Wilson, J. P., Fotheringham, A. S. (Eds.). Blackwell Publishing.

Molina, L., Belanche, L. and Nebot, A. 2002. Feature selection algorithms: A survey and experimental evaluation. In: *ICDM 2002: Proceedings of the IEEE International Conference on Data Mining*. pp. 306-313.

Moore, D. and McCabe, G. 1993. *Introduction to the practice of statistics*. WH Freeman, New York, second edition.

Mora-López, L. and Conejo, R. 1998. Qualitative reasoning model for the prediction of climatic data. In: *ECAI 1998: Proceedings of the 13th European Conference on Artificial Intelligence.* pp. 61-75.

Núñez, H. 2004. Feature Weighting in Plain Case-Based Reasoning. Ph.D. Thesis. Universitat Politècnica de Catalunya.

Núñez, H., Sànchez-Marrè, M., Cortés, U., Comas, J., Martinez, M., Rodríguez-Roda, I. and Poch, M. 2004. A comparative study on the use of similarity measures in case-based reasoning to improve the classification of environmental system situations. *Environmental Modelling and Software*, 19(9): 809-819.

Núñez, H., Sànchez-Marrè, M. and Cortés, U. 2003. Improving similarity assessment with entropy-based local weighting. In: *Lecture Notes in Artificial Intelligence, (LNAI-2689): Proceedings of the 5th International Conference on Case-Based Reasoning (ICCBR2003)*. Springer-Verlag, pp. 377-391.

Oh, S. K. and Pedrycz, W. Self-organizing polynomial neural networks based on polynomial and fuzzy polynomial neurons: analysis and design. *Fuzzy Sets and Systems* 142, 163-198, 2004.

Parr Rud, O. 2001. *Data Mining Cookbook- Modelling data for marketing, risk, and CRM*. Wiley.

Pérez-Bonilla A, Gibert, K (2007) Automatic generation of conceptual interpretation of clustering. In Progress in Pattern Recognition, Image analysis and Applications.Lecture Notes in Computer Science v 4756, pp 653-663. Springer.

Poch, M., Comas, J., Rodríguez-Roda, I., Sànchez-Marrè, M. and Cortés, U. 2004. Designing and building real environmental decision support systems. *Environmental Modelling and Software*, (19): 857-873.

Provost, F. and Fawcett, T. Robust classification for imprecise environments, *Machine Learning* 42 (3), 203-231, 2001.

Recknagel, F. 2001. Applications of machine learning to ecological modelling. *Ecological Modelling*, 146(1-3): 303-310.

Riaño, D. 1998. Learning rules within the framework of environmental sciences. In: *ECAI 1998: Proceedings of the 13th European Conference on Artificial Intelligence*. pp. 151-165.

Robertson, A., Kirshner, S. and Smyth, P. 2003. Hidden markov models for modelling daily rainfall occurrence over brazil. Tech. Rep. UCI-ICS 03-27. URL http://www.datalab.uci.edu/papers-by-date.html

Rodríguez-Roda, I., Comas, J., Colprim, J., Poch, M., Sànchez-Marrè, M., Cortés, U., Baeza, J. and Lafuente, J. 2002. A hybrid supervisory system to support wastewater treatment plant operation: Implementation and validation. *Water Science and Technology*, 45(4-5): 289-297.

Rodríguez-Roda, I., Comas, J., Poch, M., Sànchez-Marrè, M. and Cortés, U. 2001. Automatic knowledge acquisition from complex processes for the development of knowledge based systems. *Industrial and Engineering Chemistry Research*, 15(40), 3353-3360.

Rodríguez-Roda, I., Poch, Sànchez-Marrè, M., Cortés, U. and Lafuente, J. 1999. Consider a case-based system for control of complex processes. *Chemical Engineering Progress,* 6(95): 39-48.

Rubin, D. 1987. *Multiple Imputation for Nonresponse in Surveys.* Wiley.

Sanborn, S. and Bledsoe, B. 2005. Predicting streamflow regime metrics for ungauaged streams in colorado, washington, and oregon. *Journal of Hydrology* (*under review*).

Siebes, A. 1996. *Data mining: What it is and how it is done*. In: SEBD. p. 329.

Smith, C. and Spate, J. 2005. Gully erosion in the ben chifley catchment (*in preparation*).

Sànchez-Marrè, M., Cortés, U., Martínez, M., Comas, J. and Rodríguez-Roda, I. 2005. An Approach for Temporal Case-Based Reasoning: Episode-Based Reasoning. In: *Procc. of 6th International Conference on Case-Based Reasoning (ICCBR'2005)*. LNAI-3620, pp. 465 – 476.

Sànchez-Marrè, M., Gibert, K. and Rodríguez-Roda, I. 2004. GESCONDA: A Tool for Knowledge Discovery and Data Mining in Environmental Databases. Vol. 11 of Research on Computing Science. Centro de Investigación en Computación, Instituto Politécnico Nacional, México DF, México, pp. 348-364.

Sànchez-Marrè, M., Cortés, U., Béjar, J., de Gracia, J., Lafuente, J. and Poch, M. 1997. Concept formation in wastewater treatment plants by means of classification techniques: a compared study. *Applied Intelligence*, 7(2):147-165 .

Stadler, M., Ahlers, D., Bekker, R.M., Finke, J., Kunzmann, D. and Sonnenschein, M. 2006. Web-based tools for data analysis and quality assurance on a life-history trait database of plants of Northwest Europe. *Environmental Modelling and Software*, 21:1536-1543.

Spate, J. 2002. *Data in hydrology: Existing uses and new approaches*. Australian National University, Master thesis.

Spate, J. 2005. Modelling the relationship between streamflow and electrical conductivity in Hollin Creek, southeastern Australia. In: *Proceedings of the 6th International Symposium on Intelligent Data Analysis*, Fazel Famili, A., Kok, J., Peña, J. (Eds.). pp. 419-440.

Spate, J. 2006. *Machine learning as a tool for investigating environmental systems*. Australian National University, PhD Thesis.

Spate, J., Croke, B. and Jakeman, A. 2003. Data mining in hydrology. In: *MODSIM 2003: Proceedings of the 2003 International Congress on Modelling and Simulation*. pp. 422-427.

Spate, J. and Jakeman, A., 2006 Review of data mining techniques and their application to environmental problems. *Environmental Modelling and Software* (*under review*).

Su, F., Zhou, C., Lyne, V., Du, Y. and Shi, W. 2004. A data-mining approach to determine the spatio-temporal relationship between environmental factors and fish distribution. *Ecological Modelling*, 174(4): 421-431.

Swayne, D., Cook, D. and Buja, A. 1998. Xgobi: Interactive dynamic data visualization in the x window system. *Journal of Computational and Graphical Statistics*, 7(1).

Sweeney, A., Beebe, N. and Cooper, R. 2007. Analysis of environmental factors influencing the range of anopheline mosquitoes in northern Australia using a genetic algorithm and data mining methods. *Ecological Modelling*, 203(3-4): 375-386.

Ter Braak, C., Hoijtink, H., Akkermans, W. and Verdonschot, P. 2003. Bayesian model-based cluster analysis of predicting macrofaunal communities. *Ecological Modelling*, 160(3): 235-248.

Vellido, A., Martí, J., Comas, I., Rodríguez-Roda, I. and Sabater, F. 2007. Exploring the ecological status of human altered streams through generative topographic mapping. Environmental Modelling and Software, 22(7): 1053-1065.

Vojinovic, Z.; Kecman, V; Babovic, V., 2003. Hybrid approach for modelling wet weather response in wastewater systems. J. of Water Resources, Planning and Mgment., 129(6), 511-521

Voss, H., Wachowicz, M., Dzeroski, S. and Lanza, A. (Eds.). 2004. *Knowledge Discovery for Environmental Management*. Knowledge-Based Services for the Public Sector Conference. KDnet.

Ward, J. 1963. *Hierarchical Grouping to Optimize an Objective Function*.

Weiss, G. and Provost, F. 2001. *The effect of class distribution on classier learning: An empirical study*. Tech. rep., Department of Computer Science, Rutgers University, technical Report ML-TR-44.

Weiss, G. and Provost, F. Learning when training data are costly: The effect of class distribution on tree induction, *Journal of Artificial Intelligence Research* 19, 315-354, 2003.

URL http://www.research.rutgers.edu/\~{}gweiss/papers/ml-tr-44.pdf

Witten I. H., Cunningham S., Holmes G., McQueen R. J., and Smith L.A. (1993): Practical Machine Learning and its Potential Application to Problems in Agriculture. In

Proc New Zealand Computer Conference, volume 1, pages 308-325, Auckland, New Zealand, 1993.

Whitten, I. and Frank, E. 1991. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann Publishers.

Wettschereck, D., D. W. Aha, and T. Mohri. 1997. A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. *Artificial Intelligence Review*, Special Issue on lazy learning Algorithms.

Wong, I.W., Bloom, R., McNicol, D.K., Fong, P., Russell, R. and Chen, X. 2007. Species at risk: data and knowledge management within the WILDSPACETM Decision Support System. *Environmental Modelling and Software*, 22: 423-430.

Wnek, J. and Michalski, R. 1991. Hypothesis-driven constructive induction in aQ17: A method and experiments. In: *Proceedings of the IJCAI-91 Workshop on Evaluating and Changing Representation in Machine Learning*. pp. 13-22.

Xu, L., Chow, M. and Taylor, L. Power distribution fault cause identification with imbalanced data using the data mining-based fuzzy classification e-algorithm, *IEEE Transactions on Power Systems* 22 (1), 164-171, 2007.

Yeates, S. and Thomson, K. 1996. Applications of machine learning on two agricultural datasets. In: *Proceedings of the New Zealand Conference of Postgraduate Students in Engineering and Technology*. pp. 495-496.

Zadeh, L. Fuzzy sets. *Information and Control*, 8, 338-353, 1965.

Zadeh, L. Probability theory and fuzzy logic are complementary rather than competitive. *Technometrics*, 37(3), 271-276, 1995.

Zhu, X. and Simpson, A. 1996. Expert system for water treatment plant operation. *Journal of Environmental Engineering*, 822–-829.

Zoppou, C., Neilsen, O. and Zhang, L. 2002. Regionalization of daily stream flow in Australia using wavelets and k-means. Tech. rep., Australian National University, (http://wwwmaths.anu.edu.au/research.reports/mrr/mrr02.003/abs.html), accessed 15/10/2002.