International Environmental Modelling and Software Society (iEMSs)
2010 International Congress on Environmental Modelling and Software
Modelling for Environment's Sake, Fifth Biennial Meeting, Ottawa, Canada
David A. Swayne, Wanhong Yang, A. A. Voinov, A. Rizzoli, T. Filatova (Eds.)
http://www.iemss.org/iemss2010/index.php?n=Main.Proceedings

Choosing the Right Data Mining Technique: Classification of Methods and Intelligent Recommendation

Karina Gibert[a,b], Miquel Sànchez-Marrè[a,c], Víctor Codina[a,c]

[a]*Knowledge Engineering and Machine Learning Group (KEMLG)*
[b]*Statistics and Operations Research Dept.*
[c]*Computer Software Dept.*
Universitat Politècnica de Catalunya-BarcelonaTech, Barcelona, Catalonia
(karina.gibert@upc.edu, miquel@lsi.upc.edu,vcodina@lsi.upc.edu)

Abstract: One of the most difficult tasks in the whole KDD process is to choose the right data mining technique, as the commercial software tools provide more and more possibilities together and the decision requires more and more expertise on the methodological point of view. Indeed, there are a lot of data mining techniques available for an environmental scientist wishing to discover some model from her/his data. This diversity can cause some troubles to the scientist who often have not a clear idea of what are the available methods, and moreover, use to have doubts about the most suitable method to be applied to solve a concrete domain problem. Within the data mining literature there is not a common terminology. A classification of the data mining methods would greatly simplify the understanding of the whole space of available methods. Furthermore, most data mining products either do not provide intelligent assistance for addressing the data mining process or tend do so in the form of rudimentary "wizard-like" interfaces that make hard assumptions about the user's background knowledge. In this work, a classification of most common data mining methods is presented in a conceptual map which makes easier the selection process. Also an intelligent data mining assistant is presented. It is oriented to provide model/algorithm selection support, suggesting the user the most suitable data mining techniques for a given problem.

Keywords: Knowledge Discovery from Databases, Data Mining, Intelligent Decision Support Systems, Case-Base Reasoning.

1. INTRODUCTION

The classical scheme of Knowledge Discovery from Data provided by Fayyad in 1996 refers the following steps to complete the high level process of KDD, very often also called simply Data Mining (Spate et al. 2006):

- Developing and ***understanding the domain***, capturing relevant prior knowledge and the goals of the end-user
- Creating the ***target data*** set by selecting a proper set of variables or data samples (including generation of proper queries to a central data warehouse if needed)
- ***Data cleaning and preprocessing***. Quality of result strongly depends on the quality of input data, and therefore the preprocessing step is crucial (Gibert et al 2008b).
- ***Data reduction and projection***: Depending on the problem, it may be convenient to simplify the considered set of variables. The aim here is to keep a relevant set of variables describing the system adequately and efficiently (Núñez et al 2004, Gibert et al 2008b).
- ***Choosing the data mining task***, with reference to the goal of the KDD process. From clustering to time series forecasting, many different techniques exist for different purposes, or with different requirements. See (Kdnuggets 2006) for a survey of the most common data mining techniques.

- ***Selecting the data mining algorithm/s***: once the task is decided and goals are codified, a concrete method (or set of methods) needs to be chosen for searching patterns in the data. Depending on the choice of techniques, parameter optimization may or may not be required
- ***Data mining:*** Searching for patterns in data. This will be significantly improved if previous steps were performed carefully
- ***Interpreting mined patterns***. This is crucial if the discovered patterns have to support effective improvement of expert's knowledge about the analyzed phenomenon or further decision-making (Pérez-Bonilla et al 2007, Gibert et al 2010, Gibert et al 2008). If results look inconsistent possible further iteration of previous steps may be required to refine the analysis.
- ***Consolidating discovered knowledge***: Documenting and reporting results, or using them inside the target system.

Currently, we are still far from having computational systems that follow this scheme in the globality. Most of the commercial Data Mining systems, provide collections of several preprocessing, data mining and support-interpretation tools, which have to be properly combined by the data miner itself to build a correct KDD process for every particular application. One of the most difficult tasks is to choose the right data mining technique, as the commercial software tools provide more and more possibilities together and the decision requires more and more expertise on the methodological point of view.

In (Gibert et al 2008b) a high level description of a number of Data Mining techniques was presented in order to provide elements to environmental scientists to decide what to do in front a real problem. In that case we presented the techniques that we presumed could be more used for making environmental KDD, and we presented them grouped by technical proximity between them. However, in the last years we have been experiencing that either experts or data miners choose the data mining technique by using two main parameters which have nothing to do with technical characteristics of the choice. After these experiences, we strongly belief that the final choice depends basically on:

- The main goal of the problem to be solved
- The structure of the available data

There are many references in the literature describing collections of data mining techniques organized in many different ways, of course all of them valid by different reasons (Fayyad 1996; Hair et al 2002; Vazirigiannis et al 2003; Kantardzic 2003). However, providing a conceptual map of data mining techniques regarding the parameters used by human beings to decide on the right technique for a particular application, is of great help on:

- Modelling the decision process itself
- Helping non-expert data miners to improve their decisions
- Building technical data miner recommenders that in the future can be included at a higher level in Data Mining systems and contributes to approach the complex scheme proposed in Fayyad 1996. There are not many works in the literature addressing those issues. One of the available works was done by (Charest and Delisle 2006). Authors are not aware of other works trying to solve this task.

In this work, we present a classification of most common Data Mining techniques oriented to support the decisional problem of choising the right one in real applications and the advantage of using it as a reference on the construction of *intelligent data mining techniques recommenders (InDaMiTe-R)* is discussed. In the second part of the paper a first proposal on InDaMiTe-R is presented and evaluated. Finally, conclusions and future work are discussed.

2. CLASSIFICATION OF DATA MINING TECHNIQUES ORIENTED TO DECISION-MAKING

As we said before, we observed that main parameters taken into account by humans to choose the proper data mining technique in a real application are:

- The main goal of the problem to be solved
- The structure of the available data

According to that, we ellaborated the classification displayed in Fig. 1. which includes some of the most popular data mining techniques useful for environmental scientists.

The higher level division is taking into account the basic distinction between having or not a reference variable to be explained (response variable). Left part of the organigram refers to *non-supervised* methods, without response variable, in where the main goal is a better *cognition* of the target phenomenon and description is enough as a result. Whereas right part of the organigram refers those *supervised* models oriented to *re-cognition*, where a response variable is to be explained and prediction is pretended.

At a second level, for methods oriented to description, the main division regards the interest of describing relationships between objects (raws of data matrix), which are labeled as *descriptive methods*, or describing relationships between variables (columns of data matrix), labeled as *associative methods*.

For methods oriented to prediction, here the main distinction regards the nature of the response variable: while *discriminant methods* explain or predict qualitative variables, the classical *predictive methods* refer to quantitative response variables.

Because of variety, discriminant models include a further level of subdivision. *Rule-based reasoning methods* group methods providing explicit knowledge model, which can be expressed by formal rules or not, to be applied for further prediction; in *case-based reasoning methods* the predictive model is implicit in historical data; the third option is a mixture between prior explicit knowledge model and iterative refinements based on future data (*bayesian learning*).

Finally, in the presented conceptual map of Data Mining techniques, different colors have been used for methods coming from the field of Artificial Intelligence or Statistics, and additional information about more recent multidisciplinar proposals which can be classified in the intersection AI&Stats is also provided. As discussed in previous works (Gibert et al 2008b, Gibert et al 2010, Cheeseman et al 1994) these hybrid techniques use to be more powerful for modelling very complex domains, as environmental systems are.

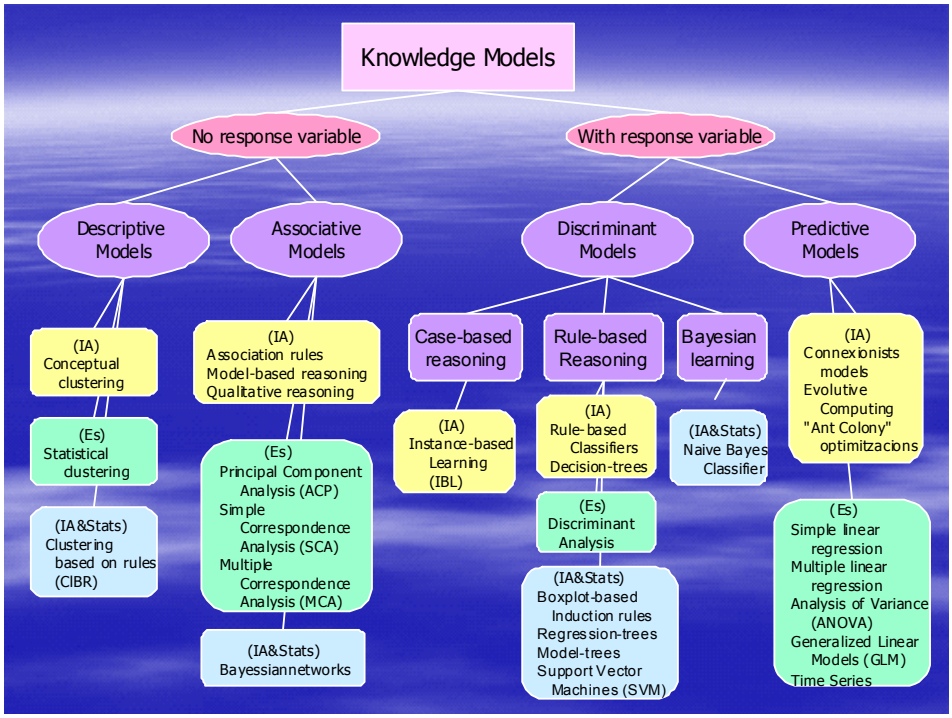


Figure 1: Classification of Data Mining Techniques

It is presented here a very brief description of all the methods included in this classification, just providing the minimum information to make the final choice. See (Gibert et al 2006) for more detailed discussion about the technical assumptions required on data for correct application on every technique.

- Conceptual clustering: Provides grouping of homogeneous objects. Requires hypothesis about the number of classes to be found. Results are directly understandable. Usually do not work with very big data sets
- Statistical clustering: Provides grouping of homogeneous objects. Might not require the number of classes. Can be efficient with big data sets. Sometimes difficult to understand the meaning of grouping provided.
- Clustering based on rules: Provides grouping of homogeneous objects. Do not require number of classes as input. Can introduce prior expert knowledge as semantic bias. Guarantee interpretability of results and coherence with prior expert knowledge.
- Association rules: Provides patterns of associated values of variables and frequencies of appearance. Interpretable results.
- Model-based reasoning: Provides formal model of the causal relationships among the domain variables, by providing models for the dependencies among variables.
- Qualitative reasoning: Provides qualitative model of the causal relationships among the domain variables, by representing which variables increase or decrease values as a consequence of modifications in the values of other variables.
- Principal component analysis: Provides graphical representation to see numerical variables which behave associated or not. Extra work required to interpret results.
- Simple correspondence analysis: Provides graphical representation to see modalities of two qualitative variables which behave associated or not. Extra work required to interpret results.
- Multiple correspondence analysis: Provides graphical representation for associations among modalities of various qualitative variables. Extra work required for interpretation.
- Bayesian networks: Provides graphical interpretation of causal relationships between variables together with conditional probabilities.
- Instance-based learning: Uses historical data to classify a new instance of a problem in a predefined set of classes.

- Rule-based classifiers: Provide a set of classification rules that can be used later to evaluate a new case and classify in a predefined set of classes.
- Decision trees: Provide a graphical representation of a tree with conditions associated to nodes that permit to classify a new instance in a predefined set of classes. Problems with very big data sets. It works with qualitative variables.
- Discriminant analysis: Provides an algebraic discriminant function and a cut-off as the rule to decide between two groups for a new instance. Only for numerical variables, two predefined classes and works only under linear separably classes.
- Support Vector Machines (SVMs): They can provide discriminant functions to distinguish between two predefined classes that can be non-lineary separable.
- Boxplot-based induction rules: Provide a set of probabilistic classification rules that can be used later to classify a new instance in a predefined set of classes.
- Regression-trees: Provide decision trees for prediction of numerical values. Each leaf has a numerical value, which is the average of all the training set values that the leaf, or rule, applies to.
- Model trees: Provide regression trees combined with regression equations. The leaves of these trees contain regression equations rather than single predicted values. A model tree approximates continuous functions by several linear submodels.
- Naïve Bayes classifier: Provides an adaptative classifier that can improve initial knowledge-based predictions for the class of a new instance by refining the model on the basis of the evidences provided by the whole history of processed cases.
- Connexionist models: Include all artificial neural networks models. Permit to predict the value of one or more variables for a new instance on the basis of non-linear combination of the values of several input variables and intermediary layers.
- Evolutionary computation: Provides the optimization of a certain objective function through the evolution of a population of individuals, which are subjected to several genetic operators. Include techniques simulating the theory of evolution, like genetic algorithms and genetic programming.
- Swarm Intelligence (SI): Provides predictions of numerical variables by training the system under the metaphore of the collective behavior of decentralized, self-organized systems, natural or artificial. Local interactions between very simple agents lead to the emergence of intelligent global behavior. Natural examples of SI include ant colonies, bird flocking, animal herding, bacterial growth, and fish schooling.
- Simple linear regression: Predicts the value of a quantitative variable for a new instance as a linear equation of a single numerical variable. Requires normality, lineality and homocedasticity.
- Multiple linear regression: Predicts the value of a quantitative variable for a new instance a s a linear equation of several numerical variables. Requires normality, linearity, homocedasticity and independence
- Analysis of Variance: Predicts the value of a quantitative variable for a new instance as a linear combination of one or two qualitative variables. Requires conditional normality, linearity, homocedasticity and independence.
- Generalized Linear Models: Predicts the value of a quantitative variable for a new instance as a linear combination of several numerical and qualitative variables. Same hipothesis as previous methods, all of them particular cases of that one.
- Time series: Predict the value of a quantitative variable for a future instance as a linear combination of past values of the same variable. See (Box et al 76) for technical hipothesis required.

3. AN INTELLIGENT DATA MINING RECOMMENDER

As evidenced in the previous section, there are many difficult and technical decisions that the data miner has to face in order to obtain the best outcome for a given dataset and user's goals. Selecting the machine learning or statistical method more appropriate, once a family of methods is found, deciding which training parameters are most appropriate or which particular technique is more convenient are some examples. Furthermore, most data mining commercial software tools either do not provide intelligent assistance for addressing the

data mining process or tend to do so in the form of rudimentary "wizard-like" interfaces that make hard assumptions about the level of background knowledge required by a user in order to effectively use the system.

At a first sight, it seems that building some knowledge-based system including as decision-rules some translation of the conceptual map presented in previous section should be the better option for building an intelligent assistant to choose the right data mining technique to be used in a specific application.

However, it is obvious that, being that map a non-exhaustive classification of data mining techniques, but the most common ones in environmental sciences, as the tendencies or needs change in the future, new refinements of the map will be required, with the consequent modifications on the assistant. Also, the number of data mining techniques available grows incredibly fast every day and this means that the knowledge-based approach is constrained to continuous reviews and upgrades. On the other hand, the number of decisions made by an expert data miner to find the right subfamily of techniques (hierarchical clustering, or partitional or fuzzy), the right parameters of execution (once decided hierarchical clustering, choose the algorithm, the metric, the aggregation criterion, sometimes, weight on the metrics) in a particular case is so complex that it becomes difficult to make explicit in a conceptual map.

That is the reason why we propose to move to a case-based reasoning approach, much more flexible to changes in the future on the methodological framework and using implicitly the expertise of data miners by means of past experiences. In fact, a key characteristic that any intelligent data mining assistant should possess is the capacity to learn from past user's experiences, so the system can help the user to avoid the repetition of mistakes and motivates the knowledge reuse, and on the other hand, to adapt to new possibilities by including them in the system easily. Thus, a non-expert data miner could take advantage of the experiences of others users facing similar problems [Charest *et al*., 2006]. And an expert can propose new solutions based on more recent trends.

3.1 System overview

Thus, an intelligent DM assistant was developed, based on a pure Case-Based Reasoning (CBR) approach. This DM assistant is integrated within the GESCONDA tool (Sànchez-Marrè et al 2004). CBR is a problem solving methodology that tries to solve new problems by re-using specific past experiences stored in a case base, and its core proposition is that new problems can be solved by reusing the solutions to similar problems that have been solved in the past. A more detailed explanation of this problem solving method can be found in Kolodner [1993]. For this reason, a CBR approach naturally fits with the above mentioned challenge of promoting the knowledge reuse. Furthermore, this approach allows the construction of a dynamic system able to refine and adapt the suggestions over time, something that is of vital importance in a domain like the data mining processes, where is extremely difficult to find the best solution due to the huge number of possible combinations of data mining techniques and training parameters.

As in most CBR approaches, the system relies on a unique case base in which past user's experiences are captured. Each case/experience is composed of:

- Case description, a set of dataset characteristics describing the type of problem the data miner wants to solve (e.g. the number of attributes and records, the ratio of symbolic attributes and the relative probability of missing values);
- Case solution, a set of DM techniques with its own configuration parameters and also associated with its corresponding evaluation measures such as the execution time and the error rate.

CBR is a cyclic and integrated problem solving process that supports learning from experience and can be described by four main phases. Retrieve the most similar case(s) to the new case, Reuse the solution(s) of the retrieved case(s) to solve the new case, Revise the proposed solution, Retain the new case into the existing case base. A more detailed

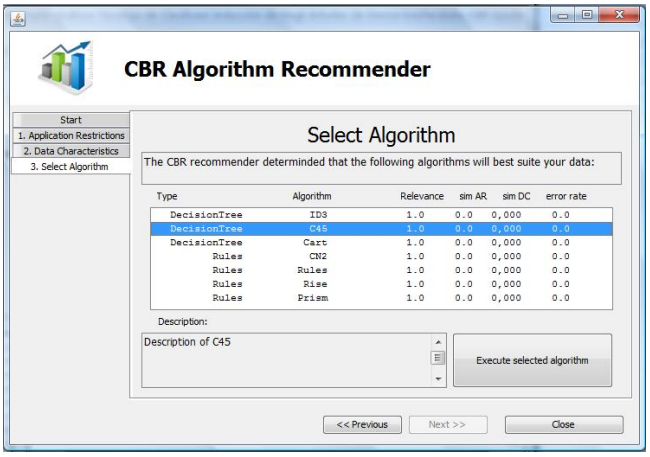
explanation of the CBR phases can be found in Aamod et al. [1994]. For the particular problem domain of DM processes our system conducts each one of these phases in the following manner:

1. Retrieve. The well-known K-nearest neighbour classification algorithm is employed in order to find the most similar cases. To do this, the system automatically extracts the metadata information of the dataset more relevant for the general type of task the data miner wants to solve (i.e. classification or regression task). The extracted dataset characteristics conform the case description as previously said.
2. Reuse. The solutions of the most similar cases are filtered and ranked according to how well fit with the application restrictions provided by the user (if any), and how similar is the new case to the past case to which the solution belongs. The ordered list of DM algorithms forms the preliminary case solution.
3. Revise. The user can execute the suggested algorithms or can modify/propose new configuration parameters and algorithms. During the user session each execution is automatically evaluated using a particular error rate measure according to the type of task the user is conducting. Furthermore, the user can explicitly evaluate the results of an execution validating so if the results are good or bad. Finally, the executed DM algorithms constitute the final case solution.
4. Retain. The system will retain the new case if its solution contains at least the execution of one DM algorithm whose evaluation results either have been validated by the user or its corresponding error rate is greater than a threshold.

3.2 Case study

An example of use is presented for illustration. A certain data miner (Maria) wants to conduct a classification for a given dataset (D1). She wonders the better DM classification algorithm and configuration parameters for D1, and she starts the DM assistant.

First of all, Maria is asked to provide the type of task, that in this case it consists of a classification problem. After that, the system automatically extracts the most relevant metadata from D1 taking into account that the user's goal is a classification task. Some examples of metadata information that may be obtained in this case would be the number of classes, the entropy of the classes and the percent of the mode category of the class. Then, the user is asked to provide some application restriction, and in this case, the restrictions of Maria are that the model has to be as accurate as possible and interpretable. With all this information the system generates a recommendation consisting of two DM algorithms: the ID3 (decision tree type) and the CN2 (rule induction type). Maria executes the two proposed algorithms with the predefined configuration parameters and validates the results. As the evaluation of the ID3 is not satisfactory enough, Maria decides to execute again the ID3 but with different parameters. Now the results are much better and Maria is satisfied with the obtained results. Finally, she saves her work and logs out the system. At this point, the system learns the new experience with its corresponding solution: the ID3 algorithm with the parameters defined by Maria and the CN2 with the default parameters. In figure 2 can be observed some screenshots of the DM assistant interface.



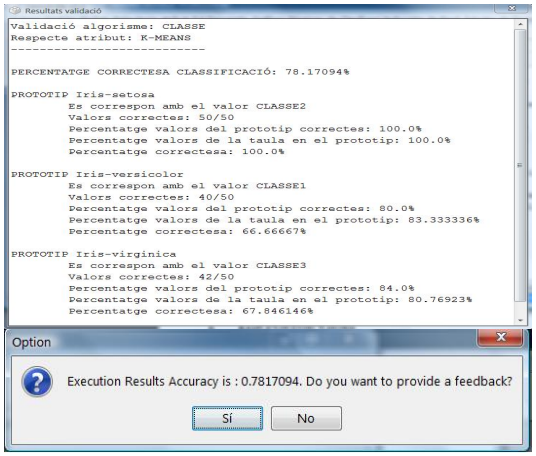


Figure 2. *(left)* screenshot of the DM assistant with an algorithm suggestion; *(right)* results evaluation results after some algorithm run and, asking to the user for explicit feedback.

3.3 Evaluation

A small-scale evaluation was carried out with 5 data miners that knew the original system without the DM assistant integrated. The experiment consisted of each user trying to solve 3 classification problems (P1, P2, P3), quite similar in terms of dataset characteristics, in different user's sessions and in a sequential manner (i.e. first all users had to solve P1, then P2 and finally P3). Doing it so, all users should be able to reuse the past experiences of others users or/and their own experiences during the small experiment.
After the execution of the experiment the users provided us some feedback about the usefulness of the assistant. In general, the user's opinions were satisfactory, especially in users that were the last ones solving some of the problems, since they could take more advantage of the past experiences of the rest of users. The results shown that the assistant is able to refine the case solutions over time, and therefore is gradually better giving support to the users that try to solve problems similar to the ones that have been solved previously.
However, the strong limitation of the system, usual in CBR, is the need of a number of varied past experiences before the assistant provides appropriate support to most users.

4. CONCLUSIONS AND FUTURE WORK

Choosing the proper data mining method is one of the most critical and difficult tasks in the KDD process. In this paper, a conceptual map of the most common data mining techniques has been proposed. There is not a unique and consensual classification of data mining methods in the literature. First main decisional criteria used by human experts in real decisions have been identified and the conceptual map is organized based on them. The proposal helps environmental data miners in the conceptual organization and rational understanding of the broad scope of data mining methods; also helps non-expert data miners to improve decisions in real applications. Finally, this provides formal expert knowledge representation to be transmitted to automatic intelligent recommenders, contributing to approach the integral conception of KDD system.

Additionally, an intelligent data mining techniques recommender is being developed based on same decisional criteria, in order to automatically provide recommendations on the best data mining technique. In order to gain flexibility and adaptability to new methods, a pure case-base reasoning aproach has been used. The prototype has been deployed and integrated in GESCONDA software and preliminary tested with successful results on expert's opinion.

Pure CBR approach requires enormous case bases to be realiable. Moving to a mixt knowledge/case-based approach can mitigate this limitation. Currently, the presented conceptual map is being included into the system to improve searches and refinements, gaining efficiency and accellerating the adaptative behaviour of the recommender. In the future, the deployment of the recommender would be more reliable in a shared environment (e.g. distributed Web system), where multiple data miners could contribute to the enrichment of the knowledge base reducing so the learning/training time of the DM assistant.

REFERENCES

Aamodt A. and E. Plaza, "Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches," *AI Communications*, vol. 7, 1994, pp. 39-59.

Box, George; Jenkins, Gwilym (1976), Time series analysis: forecasting and control CA:Holden-Day

Charest, M and S. Delisle: Ontology-guided intelligent data mining assistance: Combining declarative and procedural knowledge. Artificial Intelligence and Soft Computing 2006: 9-14

Cheeseman P and R. W. Oldford (eds) 1994: Selecting models from data. LNStats 89, Springer.

Fayyad U, et al 1996: From Data Mining to Knowledge Discovery: An overview. In Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press.

Gibert K, A. García-Rudolph, G. Rodríguez-Silva 2008: The role of KDD Support-Interpretation tools in the conceptualization of medical profiles: An application to neurorehabilitation. Acta Informatica Medica 16(4) 178-182

Gibert K, Rodríguez-Silva G, Rodríguez-Roda I, 2010: Knowledge Discovery with Clustering based on rules by States: A water treatment application. *Environmental Modelling&Software* 25:712-723

Gibert K, J. Spate, M. Sànchez-Marrè, I. Athanasiadis, J. Comas (2008b): Data Mining for Environmental Systems. In Environmental Modeling, Software and Decision Support. State of the art and New Perspectives. IDEA Series v3 (Jackeman, A. J., Voinov, A., Rizzoli, A., and Chen, S. eds), pp 205-228. Elsevier NL.

Hair JF, Jr. Robert P. Bush David J. Ortinau. 2002: Marketing Research.Within a Changing Information Environment. McGraw-Hill 2002

Mehmed Kantardzic (2003): Review of Data Mining: Concepts, Models, Methods, and Algorithms. Technometrics vol. 45, no. 3, p. 277-277

Kdnuggets (2006): http://www.kdnuggets.com/polls/2006/data_mining_methods.htm Data Mining Methods (Apr 2006)

Kolodner J. Case-Based Reasoning. Morgan Kaufmann Publishers, San Mateo,CA, 1993. Berry, N.S.M., The effect of metering on water consumption in Honiara-British Solomon

Núñez, H. and Sànchez-Marrè, M. (2004). Instance-based Learning Techniques of Unsupervised Feature Weighting do not perform so badly! In Procs 16th ECAI'2004, pp. 102-106. IOS Press.

Pérez-Bonilla A, K. Gibert 2007: Automatic generation of conceptual interpretation of clustering. In Progress in Pattern Recognition, Image analysis and Applications. LNCS-4756:653-663. Springer

Sànchez-Marrè M., Gibert K. and Rodríguez-Roda I. (2004). GESCONDA: A tool for Knowledge Discovery and Data Mining in Environmental Databases. In e-Environment: Progress and Challenge Series on Research on Computing Science, Vol. 11 pp. 348-364. CIC, IPN, México.

Spate J, K. Gibert, M. Sànchez-Marrè, E. Frank, J. Comas, I. Athanasiadis, R. Letcher 2006. Data Mining as a tool for environmental scientist. In procs 1srts iEMSs Workshop DM-TES 2006 Third Biennal Meeting: "Summit on Environmental Modelling and Software". Burlington, VE USA.

Vazirigiannis M, M. Halkidi, D. Gunopulos 2003: Uncertainty handling and quality assessment in data mining. Springer-Verlag.