Jul 1st, 12:00 AM

GESCONDA: from Environmental Data Mining to Environmental Decision Support

Miquel Sànchez-Marrè

Karina Gibert

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

GESCONDA: from Environmental Data Mining to Environmental Decision Support

M.Sànchez-Marrè[a] **and K. Gibert** [a,b]

[a] *Knowledge Engineering and Machine Learning Group, Technical University of Catalonia Catalonia, Jordi Girona 1-3, 08034 Barcelona, Catalonia.*
(miquel@lsi.upc.edu, karina.gibert@upc.edu)
[b] *Department of Statistics and Operations Research, Technical University of Catalonia, Jordi Girona 1-3, 08034 Barcelona, Catalonia.*

Keywords: Knowledge Discovery from Databases; Data Mining; Environmental applications; integrated approach

GESCONDA is a tool for intelligent data analysis and implicit knowledge management of databases, with special focus on environmental databases. Differing from existing commercial systems, the more relevant aspects of this proposal are the incorporation of the statistical data filtering and pre-processing in the same software tool together with the intelligent data analysis techniques as well as the interaction of different data mining methods. Either statistical techniques or Artificial Intelligence techniques or even mixed techniques are combined and used to extract the knowledge contained within data.

The authors are not aware of the existence of a specific software for knowledge discovery and data mining of environmental databases, taking into account the special features of environmental domains, such as the temporal and dynamic aspects of data, including both statistical data mining and statistical modelling methods, or the problem of noisy data, and data filtering with no clear relevant or irrelevant features. In fact, these are major differences with other commercial or freeware software.

The issue of our work aims at designing and building an Intelligent Knowledge Data Discovery and Data Mining System, especially suitable for environmental data analysis. The software tool, which is called GESCONDA [Sànchez-Marrè et al., 2002], is built-up. Addition of new functionalities will take place in the near future.

On the basis of previous experiences described in Sànchez-Marrè et al. [1997, 1999], it was decided that GESCONDA would have a multi-layer architecture of 4 levels connecting the user with the environmental system or process. These 4 levels are briefly described below:

- Data Filtering: Provides statistical tools for data cleaning, including one-way or two-way analysis, even graphical representations, missing data or outlier analysis, as well as management and variable transformations.

- Recommendation and Meta-Knowledge Management: supporting the formal definition of problem goals, meta-knowledge of variables and examples, methods recommender, parameter setting, and domain knowledge elicitation.

- Knowledge Discovery: including several statistical and machine learning data mining algorithms like clustering, decision tree induction or case based resoning, among others, as well as some mixed techniques, some of them developed by the authors, as detailed in Gibert, [1998], Comas et al. [2001], Sánchez-Marrè et al. [1999], and Gibert [2004].

• Knowledge Management: making possible the integration of different knowledge patterns for a predictive task, or planning, or system supervision, as well as the validation of the knowledge patterns produced in the previous step. User interaction is important in this phase, and the system supports it.

GESCONDA will be useful to acquire relevant knowledge from environmental systems, on the basis of available databases. This knowledge will be used afterwards in the implementation of reliable EDSSs. The portability of the software is provided by a common Java platform. In next sections suggestions on the use of the software are provided.

GESCONDA is a standard Java application with a friendly graphical user interface (GUI).

Input data files can be analysed by GESCONDA. They follow the standard format of instance arranged in rows, and attributes in columns. Prior to the data file loading, the user should introduce the meta-information associated with each variable into the system. The variable type must be specified: quantitative or qualitative, and in this case, also the list of modalities. For ordered qualitative variables, the ordering of modalities must be also provided. Also, the weight of the variable can be modified. In addition, the variables can be declared as active for the analyses or not, depending on the user's requirements.

After data file loading, all changes can be saved into a GESCONDA database file format (GSP/GCDA file), in order to recover the work in future working sessions with the tool. Recovering the work is done through the opening of a previously created database (GSP/GCDA file). Once data are loaded, the first thing to do in order to extract knowledge patterns from data is the descriptive statistical analysis and the data filtering task. This operations let the user check whether there are errors, outliers, badly codified data, missing data, as well as summarise main data features, such as the minimum and maximum values, the mean, standard deviation, variance, and so on. In addition, tools for arranging the variables according to the user needs are provided, when variable transformations, such as linear transformation, variable re-coding or variable standardisation, are needed prior to analysis. Other facilities such as random variable generation following several distributions such as Bernouilli, Binomial, Gaussian, Exponential, Uniform, Discrete or specific probability value computation are also available.

According to the user's goal, which could be to discover some concepts hidden in the data (clustering or grouping), or to discover some discriminant knowledge (decision trees, classification rules) to predict to which kind of concept (class or cluster) each instance belongs, different data mining techniques can be used. Furthermore, they could be combined to make a more accurate data analysis. Different scenarios are possible.

A common scenario for our system could be when an unknown environmental database, with a huge amount of instances and/or features, is faced. In this case, one possibility is to start using a clustering technique to identify typical situations in the target environmental process. Several methods are available in GESCONDA, such as K-means [Dubes and Jain, 1988], Isodata [Ball and Hall, 1965], Nearest-Neighbour classifier [Cover and Hart, 1968; Duda and Hart, 1973], Marata and COBWEB/3 [McKusick and Thompson, 1990].

After experimenting with different techniques, and trying several parameter values of the methods, the tool provides the user with a sensitivity analysis regarding the applied methods, which can be used for finding the stable set of classes, as detailed in Gibert et al. [2004]. The obtained classes and prototypes can be visualized. The resulting class is recorded as a new attribute or variable.

Afterwards, an inductive decision tree technique can be used to discover a predictive knowledge model, such a decision tree, to find the best set of attributes predicting the class label for new instances of the environmental database. In GESCONDA, the user can select and test ID3 [Quinlan, 1986], CART [Breiman et al., 1984], C4.5 [Quinlan, 1993], with optional pruning techniques. Another complementary action, to predict the class label for a new instance, is to directly induce classification rules. Several methods exist in the machine learning field; RULES [Pham & Aksoy, 1995], PRISM [Cendrowska, 1987], CN2 [Clark & Niblett, 1989], and RISE [Domingos, 1996] are implemented in GESCONDA. The user can test several parameters and validate the obtained classification rules. Some validation

techniques (simple validation, cross-validation) are also available to test the quality of the induced models.

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

For data without qualitative variables, a statistical modelling component is also available. In that case, quantitative models with several charts, graphs and model parameter estimation are found after a validation process, like multiple linear regression, ANOVA analysis [Lebart, 1990] or correlation models.

As an example of the use of GESCONDA, an IEDSS was built-up to supervise a wastewater treatment plant operation as described in Rodríguez-Roda et al. [2002].

Currently, GESCONDA is composed by several statistical data filtering analysis methods, such as one-way and two-way descriptive statistics, missing data analysis, clustering, relationship between variables, hybrid Artificial Intelligence and Statistical methods, as well as several machine learning techniques, coming from Artificial Intelligence, such as conceptual clustering methods, decision tree induction and classification rule induction.

The project purpose is to extend these intelligent system with some new agents and computational modules, such as case-based reasoning techniques, soft computing methods, support vector machines approaches, statistical models, dynamical analysis techniques, and hybrid methods integrating Artificial Intelligence approaches and Statistical ones.

The prototype is evolving from a simple data mining and knowledge discovery tool to a more complex intelligent environmental decision support tool, with a high emphasis on environmental features like:

- Huge amount of data
- Incomplete information: many missing values
- Many descriptive features: feature relevance problem
- Temporal / Spatial feature: Dynamic and Spatial data analysis
- Different Data format: Spatial data formats

ACKNOWLEDGEMENTS

The authors wish to thank the Spanish Ministry of Science and Technology for the partial financial support of the project TIN2004-01368.

REFERENCES

Ball G.H. and Hall D.J. ISODATA, a novel method of data analysis and pattern classification. Technical Report, Stanford Research Institute, 1965.

Bratko I., Dzeroski S., Kompare B. And Urbancic T. *Analysis of environmental data with machine learning methods*. Jozef Stefan Institute, 2000.

Breiman, L., Friedman, J.H., Olshen R.A. and Stone, C.J. *Classification and Regression Trees*, Wadsworth, Belmont, 1984.

Cendrowska,. J 1987. PRISM: an algorithm for inducing module rules. *Int'l Journal of Man-Machine Studies* 27(4):349-370.

Clark & Niblett, 1989. P. Clark & T. Niblett. The CN2 induction algorithm. *Machine Learning* 3:261-283.

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

Cover T.M. and Hart P.E. Nearest neighbour pattern classificat-ion. *IEEE Transactions on Information Theory*,13,21-27,1968

Domingos, P 1996 Unifying Instance-Based and Rule-Based Induction, *Machine Learning* 24(2):141-168.

Dubes R. and Kain A. Algorithms for Clustering Data. Prentice-Hall, Englewood Cliffs, USA, 1988.

Gibert K. AI and Statistics techniques for Knowledge Discovery and Data Mining. In *Tendencias de la Minería de Datos en España*. (Eds.) R. Giráldez, J. C. Riquelme, Jesús S. Aguilar-Ruiz. In press, 2004.

Gibert K. and Cortés U. Clustering based on rules and Knowl-edge Discovery in ill-structured domains. *Computación y Sis-temas* 1(4):213-227. CIC, Instituto Politécnico Nacional, 1998.

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

Gibert K., Flores X. and Sànchez-Marrè M. Comparison of classificat-ions in environmental databases using GESCONDA. *In 4th ECAI Workshop on Binding Environmental Sciences and Artificial Inte-lligence (BESAI'2004)*, 13-1:13-10, at ECAI'2004. València, 2004.

Kanevski M., R. Parkin, A. Pozdnukhov, V. Timonin, M. Maignan, V. Demyanov and S. Canu. Environmental Data Mining & Modelling based on Machine Learning algorithms and Geostatistics, *Environmental Modelling & Software* 19(9): 845-856, 2004.

Lébart, L. Traitement statistique des données. DUNOD, Paris, 90

McKusick K. and Thompson K. COBWEB/3: A Portable Implementation, Technical Report FIA-90-6-18-2, NASA Ames Research Center, June 20, 1990.

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

Pham & Aksoy, 1995. RULES: a simple ruler extraction system. *Expert Systems with Applications* 8(1).

Quinlan J.R. *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA, 1993.

Quinlan J.R. Induction of decision trees, *Machine Learning* 1(1) (1986), 81- 106.

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

Sànchez-Marrè M., Gibert K., Rodríguez-Roda R., Bueno E., Mozo L., Clavell A., Martín M. and Rougé P. (2002). Development of an Intelligent Data Analysis System for Knowledge Management in Environmental Data Bases. In Rizzoli, A. E. and Jakeman, A. E. (eds.) Integrated Assessment and Decission Supoort. *Proceedings of the First biennial Meeting of the International Environmental Modelling and Software Society*, Vol. 3, pp:420-425. iEMSs 2002.

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

Sànchez-Marrè M., Béjar J., Cortés U., Gràcia J., Lafuente J. and Poch M. Concept formation in WWTP by means of classification techniques: a compared study. *Applied Intelligence*, 7(2), 147-166. 1997.