

# **DATAMINING: Supervised and non-supervised intelligent knowledge discovery, Database and Taxonomy.**

## **ABSTRACT**

This study investigates an approach of knowledge discovery and data mining in insufficient databases. An application of Computational Taxonomy analysis demonstrates that the approach is effective in such a data mining process. The approach is characterized by the use of both the second type of domain knowledge and visualization. This type of knowledge is newly defined in this study and deduced from supposition about background situations of the domain. The supposition is triggered by strong intuition about the extracted features in a recurrent process of data mining. This type of domain knowledge is useful not only for discovering interesting knowledge but also for guiding the subsequent search for more explicit and interesting knowledge. The visualization is very useful for triggering the supposition.

In this paper we describe both, a new supervised technique based on clustering detection and a non-supervised one based on genetic algorithms. The final classification of the attributes is done applying different Bayesian criteria. The techniques are used in different sequences leading to different procedures of classification.

We applied the derived methods to a data base with the final purpose of determining client profiles in order to give, through a realistic example, the advantages and disadvantages of the different procedures. Even when there is not a unique conclusion it is possible to establish the required refinements needed on each combination of techniques.

## **DATABASE AND TAXONOMY**

We have a chance of making a clustering of a set of objects using their states or characters values. Some data mining technologies are applied to discovering the knowledge useful for the clustering analysis. This leads to investigation of effective technologies in discovering the target knowledge. The Data Matrix is, however, primarily used for obtaining the distances between objects. Therefore, the database seems insufficient for the cluster structure analysis. This leads to another investigation of appropriate technologies for data mining in insufficient databases.

In performing the data mining in insufficient databases, domain knowledge is especially effective not only in extracting interesting knowledge but also in guiding and containing the search for the interesting knowledge. Data visualization also significantly assists the data mining process as an interface between human and computer for iterative mining based on the domain knowledge.

In the course of data mining for the clustering analysis, it is noted that two types of domain knowledge are required:

The first is the domain knowledge which is typically defined and usually provided by some domain experts, in this study by Numerical Taxonomy researchers, and applications. The data mining problem involves many contextual constraints to be taken into account, which are only in experts' mind but not explicitly represented anywhere. The first type of domain knowledge brings to mind such important constraints.

The second is the domain knowledge which is newly defined in this study and deduced from supposition about background situations of a domain. The data mining process yields many incomplete features which can never be discarded to discover the target knowledge. The supposition is triggered by strong intuition about such features. The second type of domain knowledge is useful for guiding and containing the subsequent search for more explicit and interesting knowledge in the data mining process in insufficient databases.

To direct the search for the target knowledge, interaction is required between human relating to the domain knowledge and computer to do the search.

This leads to an iterative process of data mining with a preferred hierarchy for the interested set of data. This processing provides perhaps the best opportunity for the knowledge discovery in insufficient databases.

## **BAYESIAN KNOWLEDGE DISCOVERY**

Bayesian statistics is a methodological tool for data analysis, knowledge discovery and machine learning. The Bayesian paradigm is based on the Laplace's criteria which states that probability is the state of

knowledge about some event. As Laplace said, ‘Probability theory is nothing but common sense reduced to calculation’. The Bayes’ theorem in its simplest form relates the probability of two events or hypothesis  $C_i$  and events  $E$ . The theorem states that the joint probability distribution function of  $C_i$  and  $E$  can be expressed in term of the marginal and conditional distributions, given an state of knowledge we will named as  $I$  is given by:

$$P(C_i|E, I) = \frac{P(C_i|I) P(E|C_i)}{\sum P(E|C_j) P(C_j|I)}$$

Probably, the more important fact of the Bayes’ theorem is its bi-directionality, and the possibility to evaluate the probable causes of some event.

In data mining we appeal also to the concept of Bayesian network, defined as the graphical model for probabilistic relationships among a set of variables. This BN gives us the opportunity to encode the domain knowledge given in terms of a joint probability distribution, and, in combination with Bayesian techniques, to extract knowledge from data.

## THE MACHINE LEARNING APPROACH

The machine learning methods are primarily oriented towards developing symbolic logic-style descriptions of data, which may characterize one or more sets of data quantitatively, differentiate between different classes (defined by different values of designated output variables), create a ‘‘conceptual’’ classification of data, select the most representative cases, qualitatively predict sequences, and others. These techniques are particularly well suited for developing descriptions that nominal (categorical) and rank variables in data.

Another important distinction between the approaches to data analysis is that statistical methods are typically used for globally characterizing a class of objects (a table of data), but not for determining a description for predicting class memberships of future objects.

The discovered knowledge is primarily classified into two. One class is deeply concerned with the characteristics of the objects, subject matter of the classification. The other relates to a set of objects which belong to each cluster, and the characteristics of the clusters.

## REFERENCES

- Brachman, R. J., Khabaza, T., Kloesgen, W., Piatetsky-Shapiro, G., Simoudis, E., 1996. *Mining Business Databases*, Communications of the ACM.
- Codd E. F. ‘‘The Relational Model for Database Management: Version 2’’. Addison Wesley. 1990.
- Cramer, Harald. ‘‘Mathematics Methods in Statistics’’. Aguilar Edition. Madrid. Spanish. 1958.
- Crisci, J.V. , Lopez Armengol, M.F. ‘‘Introduction to Theory and Practice of the Numerical Taxonomy’’, A.S.O. Regional Program of Science and Technology for Development. Washington D.C. Spanish. 1983.
- Date, C.J. ‘‘An Introduction to Database Systems Vol. I’’. 6<sup>a</sup> Ed. Addison Wesley. 1995.
- Date, C.J. ‘‘Date on Databases’’ On proceeding of the Codd & Date Relational Database Symposium’’. Madrid. 1992.
- de Miguel, A., Piattini, M. ‘‘Concepts and Design of Databases.’’ Addison Wesley. 1994. Spanish.
- Dzeroski, Todorovski, 1994. *Discovering Dynamics, Intelligent Information Systems*.
- Evangelos, S., Han, J, (eds). 1996. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Portland, EE.UU.
- Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., Uhturudamy, R. (eds), 1996. *Advances in Knowledge Discovery and Data Mining*, San Mateo, AAAI Press.
- Fenton, N.E., Pfleeger, Sh.L. ‘‘Software Metrics’’. PWS Publishing Company. 1997.
- García Martínez, R. 1993. *Aprendizaje Automático basado en Método Heurístico de Formación y Ponderación de Teorías*. Revista Tecnología. Brasil. Volumen 15. Número 1-2. Páginas 159-182.

- García Martínez, r. 1994. *Un Sistema con Aprendizaje No-supervisado basado en Método Heurístico de Formación y Ponderación de Teorías*. Revista Latino Americana de Ingeniería. Volumen 2. Número 2. Páginas 105-127.
- García Martínez, R. & Borrajo Millán, D. 1996. *Unsupervised Machine Learning Embedded in Autonomous Intelligent Systems*. Proceedings of the XIV International Conference on Applied Informatics. Páginas 71-73. Innsbruck. Austria.
- García Martínez, R. Fernández, V & Merlo, G. 1998. *Learning in Autonomous Intelligent Systems: A Theoretical Framework*. Proceedings IV International Congresso of Informatcs Engineering. Pps 106-113. Edited by Department of Publishing of the Faculty of Engineering. Buenos Aires. Argentine.
- García Martínez, R. & Borrajo Millán, D. 1998. *Learning in Unknown Environments by Knowledge Sharing*. Proceedings of the Seventh European Workshop on Learning Robots. Pp 22-32. Edited University of Edinburg Press.
- Gennari,J.H. "A Survey of Clustering Methods" (b). Technical Report 89-38. Department of Computer Science and Informatics. University of California., Irvine, CA 92717. 1989.
- Grossman, Robert, Simon Kasif, Reagan Moore, David Rocke, and Jeff Ullman, *Data Mining Research: Opportunities and Challenges*, A Report of three NSF Workshops on Mining Large, Massive, and Distributed Data, January 1999, Chicago
- Hamming, R.W. "Coding and information theory". Englewood Clifs, NJ: Prentice Hall. 1980.
- Heckerman David. 1997. "Bayesian networks for data mining" Data mining and Knowledge discovery 1(1997) 79
- Horvitz Eric, Breese John and Henrion Max. 1988. Decision Theory in expert systems and AI. International Journal of Approximate reasoning 2 (1988) 247.
- Kitchenham,B., Pickard,L., Pfleeger, S.L. "Case studies for method and tool evaluation". IEEE Software, 12(4) pp 52-62. 1995.
- Michalski, R. S., 1991. *Searching for Knowledge in a World Flooded with Facts, Applied Stochastic Models and Data Analysis*.
- Michalski, R. S., I. Bratko, M. Kubat (eds.). *Machine Learning and Data Mining, Methods and Applications*, John Wiley & Sons Ltd, 1998, West Sussex, England
- Perichinsky, G. et Al. "Data Base Model Manager Structured on Independent Domains". Faculty of Science. National University of La Plata. Spanish. 1992.
- Perichinsky,G., Feldgen, M., Clúa,O. "Dynamic Data Bases and Taxonomy" in Proceedings International Association of Science and Technology for Development. 15<sup>th</sup>Applied Informatics-Conference.Innsbruck.Austria.1997.
- Perichinsky, G., Jimenez Rey, E.; Grossi, M. "Application of Dynamic Data Bases in Astronomic Taxonomy" in Proceedings International Association of Science and Technology for Development. 17<sup>th</sup> Applied Informatics-Conference.Innsbruck.Austria.1999.
- Sokal, R.R., Sneath, P.H.A. "Numerical Taxonomy".W.H.Freeman and Company. 1973.
- Perichinsky, G, Jiménez Rey, E., Grossi, M. D., Orellana, R. B. y Plastino, A. 1999. *Spectra of Taxonomic Evidence on the Dynamic Databases. Applied to Celestial Bodies. Asteroids Families*. Proceedings Proceedings IV International Congresso of Informatcs Engineering. Pps 301-313. Edited by Editorial Nueva Librería. Buenos Aires. Argentine.
- Perichinsky, G. *Dynamic Data Base and Taxonomy*. 1997. Proceeding of the Fiftennth International Conference Applied Informatics. Pág. 37-40 Innsbruck (Austria).
- Perichinsky, G., Jiménez Rey, E., Grossi, M. 1998a. *Spectra of Objects of Taxonomic Evidence on the Dynamic Data Bases*. Proceedings XVI International Congress on Applied Informatics. Pág. 165 - 168. Garmish-Portenkirchen. Alemania.
- Perichinsky, G., Jiménez Rey, E., Grossi, M. 1998b. *Domain Standardization of Operational Taxonomic Units (OTU'S) on Dynamic Data Bases*. Proceedings XVI International Congress on Applied Informatics. Pág. 191 - 195. Garmish-Portenkirchen. Alemania.
- Perichinsky, G., Jiménez Rey, E., Grossi, M. D. 1998c. *Spectra of Taxonomic Evidence on Databases. Proceeding IV Argentine Congress on Computer Science*. Pps. 1060 Department of Informatics and Statistics. Faculty of Ecomics and Administration. National University of the Comahue. Neuquén. Argentine.

- Quinlan, J. R., 1990. *Induction of Decision Trees*, *Machine Learning*, 1986 Quinlan, J. R., *Learning Logic Definitions from Relations*, *Machine Learning*.
- Ramoni Marco and Sebastiani Paola. An Introduction to robust bayesian classifier. March 1999. KMI technical Report KMI-TR79 March 1999.
- Thrun, Sebastian, Christos Faloutsos, Tom Mitchell, Larry Wasserman, *Automated Learning and Discovery: State-of-The-Art and Research Topics in a Rapidly Growing Field*, September 1998, Center for Automated Learning and Discovery, Carnegie Mellon University, Pittsburgh
- Van Mechelen, I., Hampton, J., Michalski, R.S., Thelus, P. (eds). *Categories and Concepts: Theoretical Views and Inductive Data Analysis*, Londres, Academic Press, 1993