

Mining Data in Multi-perspectives

João José Furtado Vasco
Universidade de Fortaleza – UNIFOR
Av. Washington Soares 1321
Fortaleza – CE Brazil
Vasco@ufc.br

Extended Abstract

Key Words: Knowledge discover in Database, Data Mining,, Concept Acquisition, Multi-perspectives.

Knowledge discover in Data bases (KDD) concerns the development of models which allow the representation and organization of the knowledge hidden in data. Machine learning systems are an alternative to automate the KDD process. In particular, incremental concept formation systems construct hierarchical abstract representations from *observations* (non-classified description of specific entities, events or situations), by recognizing regularities among them. These systems are especially interesting for KDD because they create concepts that allow better understanding of the data.

In order to create a tool for helping an expert to mine data and to identify concepts of his domain, we have defined architecture called CONFORT. This architecture is based on cognitive psychological studies that suppose that concept formation is a goal-driven process. The CONFORT architecture presupposes that goals of categorization exist (supplied by one or more experts) prior to the initiation of the process. A goal-driven concept formation process leads us naturally to a multi-perspective representation, since goals have influence on the perception of the properties as well as on the determination of relevance for context-specific features. Consequently, this situation favors the generation of different hierarchical organizations. For instance, to achieve the goal of buying a pet for a child, one would consider beauty and low price as relevant properties. As a result, the animal hierarchical organization that reflects this particular situation would probably differ from the perspective of a veterinary surgeon for whom other properties (e.g. physiological) would be relevant.

A perspective can represent the expert's opinion about a problem or it can represent a version of a database. Usually, the notion of perspective corresponds to the concept of *view* in the database (DB) context. The most important characteristic of CONFORT is the possibility to identify a relationship between perspectives, called *bridges*. Depending on which of the above mentioned contexts we work, bridges can represent expert's opinion intersection or time sequence relations between DB's versions, respectively. Two types of bridges are possible: unidirectional and bi-directional. Bi-directional bridges represent set equality relation while unidirectional ones represent set inclusion relation. When observations which are covered by a node (a concept representing a category) C are included into the set of observations which are covered by a node C' in another perspective, a bridge from C (source node) to C' (target node) is established. If the extension of C' is also included in C, a bi-directional bridge is created. Both the set inclusion and set equality relations accept the application of the transitivity property (horizontally, among perspectives), like the vertical transitivity authorized by the specialization relation in a hierarchy. In addition, the specialization relation in one hierarchy allows CONFORT to establish *hidden bridges* between children of a bridge's source node and a bridge's target node.

The core of CONFORT is FORMVIEW, a learning algorithm of incremental concept formation that uses observations to generate multi-perspective concept hierarchies and to establish bridges between them. The main FORMVIEW's input is one or several (following different perspectives) observations. FORMVIEW uses as additional data the degree of relevance (in the interval [0,1]) for each property that the expert considers important to a goal. FORMVIEW constructs *probabilistic concepts*. These concepts have the probability that an observation is classified into the category C, $P(C)$, and all possible values for the C's attributes. Each value has the conditional probability that an observation x has value v for an attribute a , given that x is a member of a category C , or $P(a=v|C)$.

We have used CONFORT in two areas: credit operation analysis and public safety. The first area concerns the credit operations done by the customers of the Brazilian Northeast Bank (BNB) in order to categorize them. BNB is a public bank where credit operations can be viewed from two perspectives: profitability and regional development. Actually, within the bank we have found two distinct areas that are responsible to analyze the customers following these two criteria. Briefly, a BNB's *best credit operation* is related to the maximization of the monetary gain but it is also related to the social advantages that it can bring to the customer's site region.

In this context, CONFORT and FORMVIEW have created hierarchical categories in each of these perspectives and it is also established bridges between them. Many bridges have been identified and analyzed by the user and some of them have been very useful. For example, it can be identified that customers that weren't necessarily the best payers and have low profit operations are considered VIP customers from the social perspective because they create wages in very critical regions.

The second area where we are employing CONFORT is safety public. This is a more recent application and we are in the first phases of analysis. We are mining the State of Ceara civil policy database, where information about crime and criminals are stored. At the present time, we are studying just the *investigation* perspective, where we trying to identify correlations between the *modus operandi* of criminals. We intent to work another perspective related to the sociological aspects of the criminals represented in the database in order to establish bridges between these two perspectives.

REFERENCES

- Barsalou, L.W.: *Ad Hoc Categories*. Memory and Cognition,11(3),1983.
- Fisher, D.H.: *Knowledge Acquisition via Incremental Conceptual Learning*. Machine Learning, vol 2, n. 2, 1987
- Gennari, J.H, Langley, P., Fisher, D.: *Models of Incremental Concept Formation*. Artificial Intelligence, 40, 1989.
- Gluck, M. A., Corter, J.E.: *Information, uncertainty, and the utility of categories*. Proc. of the 7th Annual Conference of the Cognitive Science Society. Lawrence Erlbaum, 1985.
- Hampton, J. Dubois, D.: *Psychological Models of Concepts: Introduction*. In Categories and Concepts: Theoretical Views and Inductive Data Analysis. Academic Press, 1993.
- Michalski, R., Carbonnel, J., Mitchell, T.: *Machine Learning, An Intelligence Approach*. Vol II. Morgan Kaufmann, CA. 1986.
- Reich, Y.: *Macro and Micro Perspectives of Multistrategy Learning*. In Michalski and Tecuci(Eds), Machine Learning: A multistrategy approach. Vol.IV. Morgan Kauffmann,1994.
- Seifert, C.: *A Retrieval Model Using Feature Selection*. Proc. of the 6th International Workshop on Machine Learning. Morgan Kauffmann. 1989.
- Thaise: *L'approche logique de l'intelligence artificiel*. Vol 4,1991.
- Vasco, J.J.F., Faucher, C., Chouraqui, E.: *A Knowledge Acquisition Tool for Multi-perspective Concept Formation*. European Knowledge Acquisition Workshop EKAW-96, Springer-Verlag, 1996.
- Vasco, J.J.F.: *Formation de Concepts dans un Langage de Schémas*. PhD thesis, Université d'Aix Marseille III, 1997.
- Vasco, J.J.F.: *Determining Property Relevance in Concept Formation by Determining Correlation between Properties*. European Conference on Machine Learning, ECML98, Chemnitz, Springer Verlag, 1998.