



Institute of Computer Science
Academy of Sciences of the Czech Republic

A Relational Framework for Data Mining¹

Martin Holeňa

Technical report No. 909

May 2004

Abstract:

The last decade has witnessed a fast increase of the repertoire of available sophisticated data mining methods, based on a broad spectrum of quite diverse paradigms. Besides a number of traditional and recent statistical methods, increasingly frequent are various kinds of graphical dependence models, classification and regression trees, as well as methods based on nonstatistical paradigms, such as artificial neural networks, inductive logic programming, fuzzy sets theory or rough sets theory. Such a diversity leads to problems when interpreting, comparing and consolidating results obtained with different methods. Therefore, a unifying framework of view for different data mining methods would be very useful. Several frameworks of that kind have indeed been proposed in recent years, and also the present paper is a contribution in that direction. It proposes a framework based on the theory of relations, conceived not only in the classical sense of set-theoretical relations, but in the broader sense of fuzzy sets. Underlying assumptions and basic principles of the framework are explained, and a survey of the main classes of data mining methods to which it is applicable is given. The survey points a specificity of applying the framework to the extraction of rules from data by means of artificial neural networks. The applicability of the framework is documented through elaborating it for two particular data mining methods. One of them is the method GUHA, a classical method of exploratory data analysis, the other is a method for rule extraction by means of piecewise-linear multilayer perceptrons.

Keywords:

Data mining frameworks, rules extraction from data, method Guha, observational calculus, neural-networks based rules extraction

¹Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod vodárenskou věží 2, CZ-182 07 Praha 8, e-mail: martin@cs.cas.cz

¹The research reported in this paper has been supported by the COST Action 274, "Theory and Application of Relational Structures as Knowledge Instruments (TARSKI)"