Logics and Statistics for Association Rules and Beyond

Abstract of Tutorial

Petr Hájek¹ and Jan Rauch²

¹ Institute of Computer Science, Academy of Sciences, Prague

`hajek@uivt.cas.cz`

² Laboratory of Intelligent Systems, University of Economics, Prague

`rauch@vse.cz`

The aim of the tutorial is four-fold:

(1) To present a very natural class of *logical systems suitable for formalizing, generating and evaluating statements on dependences found in given data*. The popular association rules form a particular, but by far not the only example. Technically, our logical systems are monadic observational predicate calculi, i.e. calculi with generalized quantifiers, only finite models and effective semantics. Be not shocked by these terms; no non-trivial knowledge of predicate logic will be assumed. Logic includes *deductive properties*, i.e. possibility to deduce truth of a sentence from other sentences already found true. Transparent deduction rules are very useful for systematic data mining. Special attention will be paid to sentences formalizing expressions of the form “many objects having a certain combination *A* of attributes have a also combination *B*” and, more generally, “combinations *A, B* of attributes are associated (dependent, correlated etc.) in a precisely defined manner”.

(2) To show how suitable observational sentences are related to *statistical hypothesis testing* (this aspect appears sometimes unjustly neglected in data mining). A general pattern of statistical inference will be presented in logical terms (theoretical calculi and their relation to observational ones). In particular, statistical meaning of two variants of “associational rules” as well as of some “symmetric associations” will be explained.

(3) To present short history of the *GUHA method* of automatic generation of hypotheses (General Unary Hypothesis Automaton). It is an original Czech method of exploratory data analysis, one of the oldest approaches to KDD or data mining. Its principle was formulated long before the advent of data mining. Its theoretical foundations (as presented in [5] and later publications) lead to the theory described in points (1), (2) above.

(4) Finally, to show how modern *fuzzy logic* (in the narrow sense of the word, i.e. particular many-valued symbolic logic) may enter the domain of KDD and fruitfully generalize the field. This fourth point will concentrate to open problems and research directions.

The tutorial will be complemented by demonstrations of two recent implementations of the GUHA method.

References

1. Agrawal, R., Mannila, H., Srikant, R., Toivonen, H. and Verkamo, A.: Fast discovery of association rules. In *Advances in Knowledge Discovery and Data Mining*, U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds. AAAI Press, Menlo Park, 1996, pp. 307–328.
2. Hájek, P.: *Logics for data mining (GUHA rediviva)*. (Workshop of Japanese Society for Artificial Intelligence, title in Japanese) Tokio, JSAI 1998, 27–34.
3. Hájek, P.: *Metamathematics of Fuzzy Logic*. Kluwer 1998.
4. Hájek, P., Havel, I. and Chytil, M.: The GUHA-method of automatic hypotheses determination. *Computing* 1 (1966), 293–308.
5. Hájek, P., Havránek T.: *Mechanizing Hypothesis Formation (Mathematical Foundations for a General Theory)*. Springer-Verlag, Berlin-Heidelberg-New York, 1978.
6. Hájek, P., Holeňa, M.: *Formal logics of discovery and hypothesis formation by machine*. Proc. Discovery Science (Fukuoka 1998), Springer-Verlag Editors: S. Arikawa H. Motoda LNAI 1532 pp: 291 – 302
7. Hájek, P., Sochorová, A. and Zvárová, J.: *GUHA for personal computers*. Computational Statistics & Data Analysis 19 (1995), 149–153.
8. Honzíková, Z.: *GUHA+- user's Guide*. (Tech. report, Institute of Computer Science AS CR, Prague 1999.)
9. Rauch, J.: *Classes of Four Fold Table Quantifiers*. Principles of Data Mining and Knowledge Discovery. Red. Zytkow, J - Quafafou, M. Berlin, Springer Verlag 1998, p. 203 -211
10. Rauch, J.: *Four-Fold Table Calculi for Discovery Science*. Discovery Science. Red. Arikava, S. and Motoda, eds.), Springer Verlag, Berlin, pp. 405 - 406.
11. Rauch, J.: *GUHA as a Data Mining Tool*. Practical Aspects of Knowledge Management. Schweizer Informatiker Gesellschaft Basel, 1996, 10 s.
12. Rauch, J.: *Logical Calculi for Knowledge Discovery in Databases*. Principles of Data Mining and Knowledge Discovery. Red. Komorowski, J. - Zytkow, J. Berlin, Springer Verlag 1997, p. 47 - 57
13. Rauch, J.: Four-Fold Table Calculi and Missing Information. In JCIS'98 Proceedings, (Paul P. Wang, editor), Association for Intelligent Machinery, 375-378, 1998.

This work is partly supported by grant 47160008 of the Ministry of Education of the Czech Republic.