GUHA-method in Data Mining, Pavelka Style Fuzzy Logic, Many-valued Similarity and Their Applications in Real World Problems

Esko Turunen
Tampere University of Technology
P.O. Box 553
33101 Tampere, Finland
esko.turunen@tut.fi

Abstract: *We present three fundamental tools from soft computing realm to solve real world problems that are vague in nature and not easy to handle with classical mathematical methods. We show by several real world examples how such problems are solved.*
Keywords: Fuzzy logic, many-valued similarity, data mining, control, classification, decision making.

1. Theoretical tools

(a) The GUHA-method in data mining [2]. Knowledge discovery in databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data i.e. in single flat table comprising a number of fields (columns) and records (rows). Data mining is a step in this process concerned with applying computational techniques (i.e., data mining algorithms implemented as computer programs) to actually find patters in the data. GUHA (General Unary Hypotheses Automaton) is a method of automatic generation of hypotheses based on empirical data, thus a method of data mining. The GUHA method is based on first order logic with non-classical quantifiers and finite models. Exploratory analysis means that there is no single specific hypothesis that should be tested by data; rather, the aim is to get orientation in the domain of investigation, analyse the behaviour of chosen variables, interactions among them etc. Such inquiry is not blind but directed by some general (possibly vague) direction of research (some general problem).
(b) Pavelka style fuzzy logic [6] a well-defined sentential logic with truth values in the real unit interval [0,1] (or, more generally, in an injective MV-algebra). Well-defined here means that there is a semantic consideration, i.e. truth functions, as well as syntactic consideration, that is, axioms and many-valued rules of inference, and that these two concepts coincide. This means that Pavelka style logic is complete. Moreover, this fuzzy logic generalizes classical Boolean logic in a reasonable way: everything that can be said in Boolean logic with truth values in {0,1} can be generalized to cover situations with truth values are in the whole interval [0,1]. Such an approach opens new prospects for applications.
(c) Many-valued similarity [4] is a [0,1]-valued fuzzy relation that is reflective, symmetric and weakly transitive, thus, a many-valued extension of classical equivalence relation. In real world applications objects are often not just similar or dissimilar but their

similarity is a matter of degree. Many-valued similarity can be seen as a part of Pavelka style fuzzy logic, where it has many admirable properties, for example combining several (partial) many-valued similarities into a one global (or total) similarity is again a many-valued similarity.

2. Real world applications

We present several applications that exploit the above mentioned mathematical tools.

(1) Constructing a model to predict travel time between two cites in Finland [7]. A data matrix of size 19000x8 was analyzed by the GUHA method and the results were used to construct an IF-THEN inference system of only 5 rules to predict an actual travel time.

(2) Constructing an inference system to control water level in Southern Finland lake area [1]. Given a time series data covering the period 1976-1996, the aim was to construct a computer aided system that would make the required decisions of drainage of water automatically. The problem was solved by constructing a fuzzy inference system based on Pavelka style fuzzy logic.

(3) Several traffic signal control systems based on many-valued similarity and Pavelka style fuzzy logic will be presented [5].

(4) Constructing a computer aided system to define athlete's aerobig and anaerobig thresholds [3]. The problem was solved by using many-valued similarity measures.

(5) Constructing a web based tool for voters use to select a candidate in national elections that best corresponds to a voter’s opinions. An on going project based on global fuzzy similarity.

(6) Constructing mathematical and algorithmic tools in decision making problems. An on going project based on Lukasiewicz style fuzzy logic.

Bibliography

[1] Dubrovin, T., Jolma, A. and Turunen, E.: (2002) Fuzzy model for real-time reservoir operation. Journal of Water Resources Planning and Management 128. 66-73.

[2] Hájek P., Havel I., Chytil M.: (1966) The GUHA method of automatic hypotheses determination, Computing 1, 293-308.

[3] Ketola, J., Saastamoinen, K., Turunen, E.: (2004) Defining Athlete's Aerobig and Anaerobig Thresholds by Using Similarity Measures and Differential Evolution. IEEE Int. Conference of Systems, Man and Cybernetics. 1331-1335.

[4] Kukkurainen, P., Turunen, E.: (2002) Many-valued Similarity Reasoning. An Axiomatic Approach. International Journal of Multiple Valued Logic 8, 751-760.

[5] Niittymaki, J., Turunen, E.: (2003) Traffic signal control on similarity logic reasoning. Fuzzy Sets and Systems 133. 109-131.

[6] Turunen, E.: (1999) Mathematics behind Fuzzy Logic. Advances in Soft Computing. Physica-Verlag, Heidelberg. 191 pp.

[7] Turunen, E., Coufal, D.: (2004) Short Term Prediction of Highway Travel Time Using Data Mining and Neuro-Fuzzy Methods. Neural Network World 3-4. 221-231.

