Paper

Briefly on the GUHA method of data mining

Petr Hájek

Abstract — The paper gives brief, user-oriented, information on the GUHA method.

Keywords* — *GUHA method, data mining, exploratory data analysis.

GUHA is a method of computerized generation of hypotheses based on given data. GUHA stands for General Unary Hypotheses Automaton. Its origin goes to mid-sixties of XX century, original authors are P. Hájek, I. Havel and M. Chytil [7]. The basic monograph on GUHA and its theory is P. Hájek and T. Havránek [9]. With time the method has undergone continuous development, both in its theory and implementations.

The main principle of GUHA, formulated in [7] is to describe all the possible assertions which may be hypotheses, to generate them in some optimal manner, to verify each such assertion and to output those found interesting ("interesting" meaning: supported by data and sufficiently "strong"). "*The function of GUHA is to offer hypotheses, not to verify previously formulated hypotheses.*"

This makes GUHA a method of *exploratory data analysis* (as opposed to confirmatory data analysis); since the 90-ties of the 20th century the term "*data mining*" has been in use for such methods (mainly if they deal with very large data sets, see [1, 2]). From the present point of view, GUHA is a very early (and still developing) method of data mining, which, unfortunately, has been rather unknown. Here we describe GUHA very briefly, giving reference to the literature and implementations.

The *data* to be processed can be represented as a rectangular matrix **M** whose rows correspond to some objects and columns to some attributes. (For example, objects may be patients and attributes are symptoms and diseases; or objects are bank transactions and attributes are various items describing them as kind of loan etc. – note that this terminology – transactions and items – is standard in the methods of Agrawal *et al.*). The value in the ith row and jth column is the value of jth attribute for the ith object. The value may be binary (yes-no, coded by 1 and 0), integers or reals, the "classical" case being the first. Missing values can be handled.

Logic is used to code hypotheses. Give each attributes (matrix column) a name (e.g. SEX, $AGE, \dots$). For any subset X of the domain of an attribute P, $P : X$ is the property saying "the value of P is in X, e.g.

$$TEMPERATURE : (\geq 38^\circ C).$$

If P is binary (e.g. HAS−A−CAR) then P stands for $P =$ YES and $\neg P$ (negation) for $P =$ NO. The formula $P : X$ is called a *literal* (atomic formula). You may form conjunctions of literals, e.g. $P_1 : X_1 \,\&\, P_3 : X_3 \,\&\, P_7 : X_7$ is a conjunction of three literals. The ith object satisfies this conjunction if its value (in the ith row of the data matrix) of P_1 is in X_1, value of P_3 is in X_3 and value P_7 is in X_7. Such conjunctions describe *composed properties* of our objects.

Hypotheses in GUHA (more precisely in its particular version called GUHA-assoc) have, roughly, the form "properties φ, ψ are associated" (think, for example, of smoking and cancer). We write $\varphi \sim \psi$ and here $\sim$ denotes some notion of association (logically speaking, a generalized quantifier). φ is called the *antecedent* and ψ the *succedent* of the sentence (hypothesis) $\varphi \sim \psi$.

The pair φ, ψ of properties and the given data matrix determine four frequencies:

$a =$ the number of objects satisfying both φ and ψ,
$b =$ the number of objects satisfying φ but not ψ,
$c =$ the number of objects satisfying ψ but not φ,
$d =$ the number of objects satisfying neither φ nor ψ.

This can be presented as a *four-fold table*

a	b	r
c	d	s
k	l	m

where $r = a+b$, $s = c+d$, $k = a+c$, $l = b+d$ (marginal sums) and $m = a+b+c+d = r+s = k+l$ is the number of objects in our data.

A *quantifier* (notion of association) $\sim$ is given by a function $Tr_\sim(a,b,c,d)$ associated with each four-fold table a,b,c,d either 1 (yes) or 0 (no); the formula $\varphi \sim \psi$ is *true* in the data if and only if for the four-fold table (a,b,c,d) of φ, ψ we get $Tr_{sim})(a,b,c,d) = 1$.

We shall give *four examples of quantifiers* used in GUHA (those are not all quantifiers GUHA uses).

(1) Implicational ($A \Rightarrow B$ says "Many A's are B's)"
FIMPL (founded almost-implication): It has value 1 when

$$a \geq BASE \text{ and } \frac{a}{a+b} \geq p$$

($BASE, p$ are parameters)

LIMPL (lower critical almost-implication): Parameters $BASE, p, \alpha$. It has value 1 when $a \geq BASE$ and

$$\sum_{i=a}^{r} \binom{r}{i} \cdot p^i \cdot (1-p)^{r-i} \leq \alpha.$$

(2) Comparative associational ($A \sim B$ says "B is rather more frequent among A's than among $(\neg A)$'s" (or, if you want, "A makes B more plausible".

SIMPLE (simple deviation): Parameters: $BASE$ (integer), $K \geq 1$. It has value 1 when $a \geq BASE$ and $a \cdot d > K \cdot b \cdot c$. (Note that the second formula is equivalent to

$$\frac{a}{a+Kb} > \frac{c}{c+d}).$$

Fisher quantifier: Parameters $BASE, \alpha$. It has value 1 when $a \geq BASE$, $a \cdot d > b \cdot c$ and

$$\sum_{i=0}^{\min(b,c)} \frac{(a+b)!(a+c)!(b+d)!(c+d)!}{m!(a+i)!(b-i)!(c-i)!(d+i)!} \leq \alpha .$$

In both groups, the first quantifier (FIMPL, SIMPLE) just expresses an observation on frequencies in the data; the second, more complicated, is a test of a statistical hypothesis concerning probabilities in the unknown universe from which our data are a sample. (High conditional probability in the case of LIMPL, positive dependence in the case of FISHER).

Note that the GUHA theory has a notion of associational and implicational quantifiers; FIMPL and LIMPL are examples of implicational quantifiers whereas SIMPLE and FISHER are examples of associational quantifiers that are not implicational (conversely, each implicational quantifier is associational). We do not go into any details; see [9] for this theory.

Let us also mention that FIMPL is almost the same notion of association as what Agrawal and his school calls "association rule", reinventing in fact our FIMPL (the only difference is that our BASE gives a lower bound for a whereas his "support" gives a lower bound to a/m, where m is the number of objects). Also note that examples of other quantifiers are found in [9] and in the papers by Rauch [21–25].

Now we shortly describe a GUHA procedure ASSOC working with associational quantifiers. The application of the procedure takes place in three steps:

- preprocessing – input of the data matrix and parameters determining syntactic form of antecedents and succedents to be generated (e.g. maximal length of conjunctions, attributes allowed to occur only in antecedents, in the succedents; choice of the quantifier used and its parameters etc.); preparing the internal representation of the data matrix in a form suitable for quick generation and evaluation of hypotheses;

- kernel – generating and evaluating hypotheses; a system of "interesting" ones is produced;

- postprocessing – browsing in the hypotheses found and their interpretation; sorting, reordering etc.

There have been several implementations in the history of GUHA; but the reader will agree that implementations get obsolete very quickly. There is a working implementation for PC [15] and two present implementations under Windows, freely available: GUHA+- and 4ft-Miner [28, 29]. The references below contain a selection of works devoted to GUHA theory and selected applications. *Don't overlook* the fact that the basic monograph on GUHA theory, [9] is now freely available on the web [10].

The antiquity of GUHA is not the most important thing; more important is the fact that the theory developed for GUHA is also presently useful and inspirative for data mining. One can hope that it will become more broadly known in the future.

Acknowledgement

Partial support by the COST action 274 TARSKI is acknowledged.

References

[1] R. Agrawal, H. Manilla, R. Sukent, A. Toivonen, and A. Verkamo, "Fast discovery of association rules" in *Advance in Knowledge Discovery and Data Mining*. AAA Press, 1996, pp. 307–328.

[2] J. M. Adamo, *Data Mining for Associational Rules and Sequential Patterns, Sequential and Parallel Algorithms*. Springer, 2001.

[3] D. Coufal, "GUHA analysis of air pollution data" in *Proc. Int. Conf. Artificial Neural Nets and Genetic Algorithms ICANNGA'2001*, V. Kůrková, N. C. Steele, R. Neruda, and M. Kárný, Eds., Prague, Czech Rep., 2001. Springer, 2001, pp. 465–468.

[4] T. Feglar, "The GUHA architecture" in *Proc. Relmics 6*, Tilburg, The Netherlands, 2001, pp. 358–364.

[5] P. Hájek, "The GUHA method and mining association rules" in *Proc. CIMA'2001*, Bangor, Wales, 2001, pp. 533–539.

[6] P. Hájek, "The new version of the GUHA procedure ASSOC" in *Proc. COMPSTAT*, Physica-Verlag, 1984, pp. 360–365.

[7] P. Hájek, I. Havel, and M. Chytil, "The GUHA method of automatic hypotheses determination", *Computing*, no. 1, pp. 293–308, 1966.

[8] P. Hájek, K. Bendová, and Z. Renc, "The GUHA method and three-valued logic", *Kybernetika*, no. 7, pp. 421–431, 1971.

[9] P. Hájek and T. Havránek, *Mechanizing Hypothesis Formation (Mathematical Foundations for a General Theory)*. Springer, 1978.

[10] P. Hájek and T. Havránek, *Mechanizing Hypothesis Formation (Mathematical Foundations for a General Theory)*. Internet edition, http://www.cs.cas.cz/∼hajek/guhabook/

[11] P. Hájek and M. Holeňa, "Formal logics of discovery and hypothesis formation by machine", *Theor. Comput. Sci.*, vol. 292, pp. 345–357, 2003.

[12] P. Hájek, J. Rauch, T. Feglar, and D. Coufal, "The GUHA method, data preprocessing and mining", in *Proc. Database Technologies for Data Mining DTDM'02*, Praha, Czech Rep., 2002, pp. 29–36.

[13] P. Hájek, Guest Ed., *Int. J. Man-Mach. Stud.*, vol. 10, no. 1, 1977 (special issue on Guha). Introductory paper, P. Hájek and T. Havránek, "The GUHA method – its aims and techniques", *Int. J. Man-Mach. Stud.*, no. 10, pp. 3–22, 1977.

[14] P. Hájek, Guest Ed., *Int. J. Man-Mach. Stud.*, vol. 15, no. 3, 1981 (second special issue on GUHA).

[15] P. Hájek, A. Sochorová, and J. Zvárová, "GUHA for personal computers", *Comp. Stat. Data Anal.*, no. 19, pp. 149–153, 1995.

[16] J. Hálová and P. Žák, "Coping discovery challenge of mutagenes discovery with GUHA+/- for windows" in *Int. Workshop on KDD Challenge on Real-world Data*, Kyoto, Japan, 2000, pp. 55–60; *Pacific-Asia Conf. Knowledge Discovery and Data Mining*, Kyoto, Japan, 2000.

[17] T. Havránek, "The statistical modification and interpretation of GUHA method", *Kybernetika*, no. 7, pp. 13–21, 1971.

[18] M. Holeňa, "Fuzzy hypotheses for GUHA implications", *Fuzzy Sets Syst.*, vol. 98, pp. 101–125, 1998.

[19] M. Holeňa, "Exploratory data processing using a fuzzy generalization of the GUHA approach" in *Fuzzy Logic*, Baldwin *et al.*, Eds., New York: Willey, 1996, pp. 213–229.

[20] L. Pecen, E. Pelikán, H. Beran, and D. Pivka, "Short-term fx market analysis and prediction" in *Neural Networks in Financial Engeneering*, A. P. Refenes, J. Moody, and Y. Abu-Mustafa, Eds., *World Scientific*, 1995, pp. 189–196.

[21] J. Rauch, "GUHA as a data mining tool, practical aspects of knowledge management". Schweizer Informatiker Gesellshaft Basel, 1996.

[22] J. Rauch, "Logical calculi for knowledge discovery" in *Principles of Data Mining and Knowledge Discovery*, J. Komorowski and J. Zytkow, Eds. Berlin: Springer, 1997, pp. 47–57.

[23] J. Rauch, "Logical problems of statistical data analysis in databases" in *Proc. Eleventh Int. Sem. Database Manag. Syst.*, 1988, pp. 53–63.

[24] J. Rauch and M. Šimůnek, "Mining for 4ft association rules" in *Proc. Discovery Science*, 2000, Kyoto. Springer, 2000, pp. 268–272.

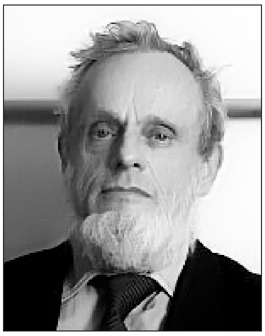
[25] J. Rauch and M. Šimůnek, "Mining for statistical association rules" in *Proc. PAKDD*, Hong Kong, 2001, pp. 149–158.

[26] V. Šebesta V and L. Straka, "Determination of suitable markers by the GUHA method for the prediction of bleeding at patients with chronic lymphoblastic leukemia" in *Mediter. Conf. Med. Biolog. Eng. Comput. Medicon 98*, Lemesos, Cyprus, 1998.

[27] J. Zvárová, J. Preiss, and A. Sochorová, "Analysis of data about epileptic patients using GUHA method" in *EuroMISE 95: Information, Health and Education*, J. Zvárová and I. Malá, Eds., Prague, Czech Rep., 1995; *TEMPUS Int. Conf.*, Prague, Czech Rep., 1995.

[28] "GUHA+– project web site", http://www.cs.cas.cz/ics/software.html

[29] LISp-Miner overview, http://lispminer.vse.cz/overview/4ftminer.html



Petr Hájek graduated in mathematics in Charles University in Prague. He obtained Ph.D. and D.Sc. degrees in mathematical logic in the Czechoslovak Academy of Sciences. In 1998 he obtained professorship in mathematics (awarded by the president of the Czech Republic). He worked in Mathematical Institute of the Czechoslovak Academy of Sciences (now Academy of Sciences of the Czech Republic) and then became a director of the Institute of Computer Science of AS CR, where he is now a senior scientist. He is an author and co-author of 6 monographs and more than 200 scientific papers in the field of mathematical logic. Member of the Association for Symbolic Logic and of editorial boards of *Archive for Mathematical Logic*, *Fundamenta Informaticae*, *Soft Computing* and *Czechoslovak Mathematical Journal*.
e-mail: hajek@cs.cas.cz
Institute of Computer Science
Academy of Sciences
Pod vodarenskou vezi 2
182 07 Prague, Czech Republic