The GUHA method, data preprocessing and mining. (Position paper.)

Hájek, Petr
2002

Institute of Computer Science
Academy of Sciences of the Czech Republic

The GUHA method, data preprocessing and mining. (Position paper.)

Petr Hájek, Tomáš Feglar, Jan Rauch, David Coufal

Technical report No. 867

Pod Vodárenskou věží 2, 182 07 Prague 8, phone: (+4202) 6605 1111, fax: (+4202) 86 585 789,
e-mail:ics@cs.cas.cz

Institute of Computer Science
Academy of Sciences of the Czech Republic

The GUHA method, data preprocessing and mining. (Position paper.)

Petr Hájek, Tomáš Feglar, Jan Rauch[1], David Coufal

Technical report No. 867

Abstract:

The paper surveys basic principles and foundations of the GUHA method, relation to some well-known data mining systems, main publications, existing implementations and future plans.

Keywords:
GUHA method, data mining, knowledge discovery in databases

[1]University of Economics, Prague

1 Introduction: Basic principles

GUHA (General Unary Hypotheses Automaton) is a method originated in Prague (in Czechoslovak Academy of Sciences) in mid-sixties. Its main *principle* is *to let the computer generate and evaluate all hypotheses that may be interesting from the point of view of the given data and the studied problem.* This principle has lead both to a specific theory and to several software implementations. Whereas the latter become quickly obsolete, the theory elaborated in the mean time has its standing value. Typically hypotheses have the form "Many A's are B's" (B is highly frequented in A) of "A,S are mutually positively dependent". (Note that what is now called "association rules" in data mining occurs already in the first 1966 paper [11] on GUHA, see below.) A second feature, very important for GUHA, is its *explicit logical and statistical foundations.*

2 Foundations

Logical foundations include *observational calculi* (a kind of predicate calculi with only finite models and with generalized quantifiers, serving to express relations among the attributes valid in data) and *theoretical calculi* (a kind of modal predicate calculi serving to express probabilistic or other dependencies among the attributes, meaningful in the universe of discourse). Statistical foundations include principles of statistical hypotheses testing and other topics of *exploratory data analysis.* Statistical hypothesis testing is described as a sort of inference in logical sense. But note that GUHA is not bound to generation of *statistical* hypotheses; the logical theory of observational calculi is just logic of *patterns* (associations, dependencies etc.) contained (true) in the data.

The monograph [13] contains detailed exposition of fundamentals of this theory. The underlying logical calculi are analyzed and several basic facts for corresponding algorithms are proved. Special attention is paid to deduction rules serving for optimization of knowledge representation and of intelligent search.

Remark. This book [13] has been not more obtainable since several years ago. We are happy to announce that its publisher, Springer-Verlag, reverted the copyright to the authors which has made possible to *put the text of the book on web for free copying* as a report of the Institute of Computer Science [14]. It is hoped that this contributes to dissemination of the theoretical foundations of GUHA-style data mining, useful for data mining in general.

3 Hypotheses alias rules.

For simplicity imagine the data processed by GUHA as a rectangular matrix of zeros and ones, the rows corresponding to objects and columns to some attributes. (Needless to say, much more general data can be processed.) In the terminology of [1], columns correspond to *items* and rows describe itemsets corresponding to *transactions.* In logical terminology one works with predicates $P_1, \ldots, P_n$ (names of attributes), *negated predicates* $\neg P_1, \ldots \neg P_n$, elementary conjunctions (e.g. $P_1 \& P_3 \& P_7$) and possibly elementary disjunctions ($P_1 \vee \neg P_3 \vee P_7$).

A *hypothesis* (rule, observational statement) has the form $\varphi \Rightarrow^* \psi$ where φ, ψ are elementary conjunctions (subjected to some syntactic restrictions) and $\Rightarrow^*$ is a *generalized quantifier.* The formulas φ, ψ determine four frequencies a, b, c, d (of $\varphi\&\psi, \varphi\&\neg\psi, \neg\varphi\&\psi, \neg\varphi\&\neg\psi$) often presented as a fourfold *table.* The frequencies decide if the formula $\varphi \Rightarrow^* \psi$ is true in data via the semantics of $\Rightarrow^*$. Two quick examples: the quantifier $\Rightarrow_{p,s}$ of *founded implication:* hypothesis true if $a/(a+b) \geq p$ and $a \geq s$. (see [11]). This is almost the semantics of Agrawal [1] (only instead giving a lower bound for the absolute frequency a he gives a lower bound *minsup* for the relative frequency a/m where m is the number of transactions; this is indeed a very unessential difference.) The quantifier $\sim$ of simple deviation: hypothesis is true if $ad > bc$ (equivalently, if $\frac{a}{a+b} > \frac{c}{c+d}$; in words: ψ is more frequent among objects satisfying φ than those satisfying $\neg\varphi$). For statistical variants of both see [11] or [18]. (Note that this is by far not the only kind of patterns analyzed in [11].)

4 Relation to data mining and discovery science

Both data mining and discovery science are terms that emerged recently and have aims similar to each other as well as to the main ideas of GUHA declared from its beginning: to develop methods of discovering (mining) knowledge form data (usually large data). Relations of GUHA and its ancestors to data mining and discovery science were analyzed in [32], [33], [15], [24], [9].

In particular, our hypotheses described above are more general than Agrawal's association rules particularly by (1) explicit use of negations and (2) choice from a variety of quantifiers, not just FIMPL. On the other hand, Agrawal's data mining is particularly developed for processing extremely huge data, which influence the choice of techniques for pruning the system (tree) of hypotheses = rules. These aspects and possibilities of mutual influence of GUHA and Agrawal's approach are analyzed especially in [9].

The fact that Agrawal's notion of an association rule [1] occurs in fact in [11] has remained unnoticed for long; but note that it is explicitly stated e.g. in [28]. Let us stress that priority questions are by far not the most important thing; what is valuable is possible mutual influence. (This is discussed e.g. in [9].

5 The relation of the GUHA method to modern database systems

The GUHA implementations were used in various research domains (e.g. in medicine [19, 26, 37, 38], pharmacology [20, 21, 22], banking [2, 3, 29, 30] or in meteorology [4]), but admittedly, they never got a broad use.

It is obvious that if GUHA method aims to be efficiently applied it has to be implemented in a software form on a computer. During the rather long history of the method there were created several implementations of the method reflecting the rapid development in information technologies (IT). The first implementations were realized on MINSK22 computer in 60's and on mainframes in 70's. In the 80's implementations were transferred on IBM PC platform which gave PC-GUHA implementation for MS DOS operation system and two present implementations named GUHA +– [27] and 4FT-Miner [35] for MS Windows operation system. This evolution can be considered as a "hardware driven" because it was mainly caused by the hardware progress in IT, but there is also other type of evolution in GUHA implementations which can be seen as "software driven".

Actually, the software driven evolution has at least three main directions. The first one is in a sense of new algorithms developed for hypotheses generation and testing on base of dichotomized data. In fact, these algorithms were established very early in the history of the method and they are not so affected by progress in IT. The second direction can be characterized as the one solving the problem of accessing of raw data (not dichotomized) by the method. In contrast to the first direction this second one is highly affected by progress in IT. The third direction concerns post-processing of the set of found hypotheses. Inspite of the fact that one part of input information are syntactical restrictions to the generated hypotheses, the resulting set of hypotheses supported by the data may be huge. Several techniques of sorting, ordering and other post-processing have been proposed and implemented.

In the first implementations of the method data were stored in stand alone plain files. In these files data were already dichotomized (the oldest implementations) or in a raw form, consequently dichotomized in an automatic way on base of scripts determined by user (PC-GUHA). However, such a way of access to raw data was untenable in a light of developments in a database software industry. Therefore present implementations (GUHA+–, 4FT-Miner) employ universal ODBC interface to access raw data typically stored in form of tables in MS Access or MS Excel software. But progress continues.

Note that a GUHA-DBS database system was proposed in early 1980's by Pokorný and Rauch [31], see also [34]. This is now rather obsolete and this development was neglected in GUHA for long; but since recently there has been a new interest in this, see [5].

Contemporary trends in data mining area are driven by requirements for processing huge data sets which brings new research and implementation problems for GUHA and its logical and statistical theory. Standard sources of huge databases - created typically in a dynamic way - are hypermarkets, banks, internet applications, etc. These data sets are enormous and professional databases as Oracle or MS SQL Server have to be used to manage them. To cope with these modern trends there was established a research group formed around COST Action 274 aiming on a new implementation of the GUHA method enabling to work efficiently with large data sets.

Actually, there are generally two ways possible of raw data access in a new GUHA method's implementation. The first, the simpler way is to transform respective data objects of modern database systems (e.g., data cubes of MS SQL Server 2000) into standard tables and then access data by old algorithms already programmed in GUHA +– or 4FT-Miner implementations. The second approach which is more tending and we aim on it preferably within the work in COST Action 274, is to homogeneously interconnect GUHA core algorithms with data access algorithms offered by modern data base systems issuing into a qualitatively new (faster) processing of huge data sets. Detail description of work on this task can be found in the report [6].

6 Conclusion

Research tasks include: Further comparison of GUHA theory with the approaches of data mining for mutual benefits. Systematic development of the theory in relation to fuzzy logic (in the style of Hájek's monograph [7]). Development of observational calculi for temporal hypotheses (reflecting time). Systematic development of the database aspects, in particular improving existing methods and elaborating new methods of data pre-processing and post-processing of GUHA results. Design of a new GUHA-style system based on data received from distributed network resources, and construction of a model of the customer decision processes.

Acknowledgement:

Theoretical and practical development of the GUHA method is the subject of Czech participation in the EU COST project 274 (TARSKI).

References

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

[2] Coufal D., Holeňa Martin, Sochorová A.: Coping with Discovery Challenge by GUHA. In: Discovery Challenge. A Collaborative Effort in Knowledge. Discovery from Databases. - Prague, University of Economics 1999, pp. 7-16, PKDD'99 European Conference on Principles and Practice of Knowledge Discovery in Databases /3./, Prague, Czech Rep., 99.09.15-99.09.18

[3] Coufal D.: Financial Data Set Analysis - Hierarchical Testing with GUHA Method. In: Discovery Challenge. (Ed.: Siebens A., Berka P.) - Prague, [VSE] 2000, PKDD 2000 European Conference on Principles and Practice of Knowledge Discovery in Databases /4./, Lyon, France, 00.09.12-00.09.16

[4] Coufal D.: GUHA Analysis of Air Pollution Data. In: Artificial Neural Nets and Genetic Algorithms. Proceedings of the International conference. (Ed.: Kůrková V., Steele N.C., Neruda R., Kárný M.) - Wien, Springer 2001, pp. 465-468,), ICANNGA'2001 /5./, Prague, Czech Rep., 01.04.22-01.04.25

[5] Feglar T.: The GUHA architecture. Proc. Relmics 6, Tilburg (The Netherlands), pp. 358–364.

[6] Feglar T.: The GUHA Virtual Machine - Frameworks and Key Concept - Frameworks and Key Concept, Research Report COST 274, Year 2001.

[7] Hájek P.: Metamathematics of Fuzzy Logic, Kluwer 1998.

[8] Hájek P.: Relations in GUHA style data mining. Proc. Relmics 6, Tilburg (The Netherlands) 91-96.

[9] Hájek P.: The GUHA method and mining association rules. Proc. CIMA'2001 (Bangor, Wales) 533-539

[10] Hájek P.: The new version of the GUHA procedure ASSOC, COMPSTAT 1984, pp. 360-365.

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

[12] Hájek P., Bendová K., Renc Z.: The GUHA method and three-valued logic, Kybernetika 7(1971) 421-431.

[13] Hájek P., Havránek T.: Mechanizing Hypothesis Formation (Mathematical Foundations for a General Theory), Springer-Verlag 1978, 396 pp.

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

[15] Hájek P., Holeňa M.: Formal logics of discovery and hypothesis formation by machine. To appear in Theoretical Computer Science.

[16] Hájek P. (guest editor): International Journal of man-Machine Studies, vol. 10, No 1 (special issue on Guha). Introductory paper of the volume is Hájek, Havránek: The GUHA method - its aims and techniques. Int. J. Man-Machine Studies 10(1977) 3-22.

[17] Hájek P. (guest editor): International Journal for Man-Machine Studies, vol. 15, No 3 (second special issue on GUHA)

[18] Hájek P., Sochorová A., Zvárová J.: GUHA for personal computers, Comp. Stat., Data Arch. 19, pp. 149-153.

[19] Hálová J., Žák P.: Coping Discovery challenge of mutagenes discovery with GUHA+/- for windows. In: The Fourth Pacific-Asia Conference on Knowledge Discovery and Data Mining. Workshop KDD Challenge 2000. International Workshop on KDD Challenge on Real-world Data. - Kyoto, - 2000, pp. 55-60, Pacific-Asia Conference on Knowledge Discovery and Data Mining /4./, Kyoto, Japan, 00.04.18-00.04.20

[20] Hálová J., Žák P.: Drug Tailoring by GUHA +/- for Windows. In: Challenges for MCDM in the New Millenium. Abstracts. - Ankara, Middle East Technical University 2000, pp. 50, International Conference MCDM /15./, Ankara, Turkey, 00.07.10-00.07.14

[21] Hálová J., Žák P.: Fingerprint Descriptors in Tailoring New Drugs Using GUHA Method. In: 51th Meeting of the European Working Group Multicriteria Aid for Decisions. Program and Abstracts. - Madrid, - 2000, pp. 25, Meeting of European Working Group Multicriteria Aid for Decisions /51./, Madrid, Spain, 00.03.30-00.03.3

[22] Hálová J., Žák P.: Quantitative Structure-Activity Relaitionship by GUHA Method. In: 52nd Meeting of the European Working Group Multicriteria Aids for Decision. Program and Abstracts. - Vilnius, - 2000, pp. 40 ,Meeting of European Working Group Multicriteria Aid for Decisions /52./, Vilnius, Lithuania, 00.10.05-00.10.06

[23] Havránek T.: The statistical modification ond interpretation of GUHA method, Kybernetika 7(1971) 13-21.

[24] Holeňa M., Fuzzy hypotheses for GUHA implications, Fuzzy Sets and Systems 98 (1998), 101-125.

[25] Holeňa M.: Exploratory data processing using a fuzzy generalization of the GUHA approach, Fuzzy Logic, Baldwin et al., ed. Willey et Sons, New York, 1996, pp. 213-229.

[26] Holubec L., jr., Topolcan O., Pikner R., Pecen L., Holubec L., sen., Fínek J., Ludvíkov M.: Discriminative Level of Tumor Markers after Primary Therapy in Colorectal Carcinoma Patients. In: ISOBM Meeting. Abstract Book. - Barcelona, - 2001, pp. 173, ISOBM Meeting /29./ International Society for Oncodevelopmental Biology and Medicine, International Symposium /8./ Biology and Clinical Usefulness of Tumor Markers, Barcelona, Spain, 01.09.29-01.10.03

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

[28] Lin W., Alvarez S. A., Ruiz C.: Collaborative recommendation via adaptive association rule mining. Web-KDD 2000.

[29] Pecen L., Ramešová N., Pelikán E., Beran H.: Application of the GUHA method on financial data. Neural Network World 5 (1995), 565–571

[30] Pecen L., Pelikán E., Beran H., and Pivka D.: Short-term fx market analysis and prediction. In Neural Networks in Financial Engeneering (1996), pp.189-196

[31] Pokorný D., Rauch J.: The GUHA-DBS Data Base System. Int. Journ. Math. Machine Studies 15 (1981), pp. 289–298.

[32] Rauch J.: GUHA as a Data Mining Tool, Practical Aspects of Knowledge management. Schweizer Informatiker Gesellshaft Basel, 1996, 10 s.

[33] Rauch J.: Logical Calculi for Knowledge Discovery. Red. Komorowski, J. - Zytkow, J. Berlin, Springer Verlag 1997, pp. 47-57.

[34] Rauch J.: Logical problems of statistical data analysis in databases. Proc. Eleventh Int. Seminar on Database Management Systems (1988), pp. 53–63.

[35] Rauch J., Šimůnek M.: Mining for 4ft association rules. Proc. Discovery Science 2000 Kyoto, Springer Verlag 2000, 268-272

[36] Rauch J., Šimůnek M.: Mining for statistical association rules. Proc. PAKDD 2001 Hong Kong, 149-158.

[37] Šebesta V., Straka L.: Determination of Suitable Markers by the GUHA Method for the Prediction of Bleeding at Patients with Chronic Lymphoblastic Leukemia. In: Medicon 98, Mediterranean Conference on Medical and Biological Engineering and Computing /8./, Lemesos, Cyprus

[38] Zvárová J., Preiss J., Sochorová A.: Analysis of Data about Epileptic Patients Using GUHA Method. In: EuroMISE 95: Information, Health and Education. (Ed.: Zvárová J., Malá I.) - Prague, EuroMISE Center 1995, pp. 87, TEMPUS International Conference, Prague, Czech Republic, 95.10.20-95.10.23