Medical Informatics Europe '99
P. Kokol et al. (Eds.)
IOS Press, 1999

Increasing the Diversity of Medical Data Mining through Distributed Object Technology

Martin Holeňa[1], Anna Sochorová[1], Jana Zvárová[2]
Institute of Computer Science, Academy of Sciences, Prague (1),
Euromise, Charles University and Academy of Sciences, Prague (2)

Abstract. Data mining has been recently experiencing a boom of interest from researchers and software producers. In medicine, however, its applications are still rather rare. In this paper, we argue that this is primarily due to the requirements of reproducibility of results and diversity of available data mining tools, both of which are crucially important for medical research. We propose to tackle the diversity requirement by means of distributed object technologies. The results presented here rely on our experience with medical data mining using the method GUHA and with further developing that method.

1. Introduction

Some 15 years ago, the term *data mining* was used rather derogatorily, to denote a search for the best fitting model or the most significant hypothesis by trying a large number of models or hypotheses on the same body of data, not worrying about the requirement of reproducibility of the obtained results [6], [22]. In this paper, we argue that it is this requirement, together with the required diversity of data mining tools, that cause the still low acceptance of medical data mining, compared with the boom of interest data mining experiences in some other application domains such as marketing and finance [14], [23]. To deal with data mining in the context of those requirements is the objective of our paper. In the following two sections, the role of data mining in medicine is discussed, and our experience with medical data mining using the method GUHA is sketched. The last two sections then analyse the diversity requirement and its existing solutions, outlining the possible impact of that analysis on our prospective data mining strategy.

2. Role of data mining in medicine

Data mining is usually defined somehow like a *nontrivial search for interesting relationships and patterns hidden in data* with the aim of *knowledge discovery* from those data [4], [9], [10], [29]. In the context of medicine, it is important to note that the concept of *knowledge* is understood differently in data mining and in medical research:

- Medical research is mostly concerned with law-like, objectively valid knowledge. In some respect, the objective of each medical research study is to extend, detail or correct the existing medical knowledge, typically through confirming or refuting either some in advance postulated conjecture, or some already available piece of knowledge. To achieve an objective validity of the knowledge, all steps through which it can be established, including the performed data analysis, are required to be exactly *reproducible.*
- Data mining, in contrast, is concerned with knowledge about the data in a particular database or dataset. Though that knowledge is usually also established through

evaluating some hypotheses, with the purpose of confirming or rejecting them, these hypotheses are typically not postulated by humans, but are automatically generated during a search in a predefined hypotheses space [3], [14], [29], [34]. Due to the limitation to particular databases and datasets, data mining is not concerned about the reproducibility of results.

This comparison clearly shows that data mining is not able to substitute rigorous data analysis with respect to the cognitive task of medical research. On the other hand, it can be very well employed for the often tedious exploratory pre-processing of the data, which is needed especially in situations when the data have not been collected with any particular conjecture in mind, but simply as a result of an improved computer support of the health care process. It is this kind of data the amount of which has been growing most rapidly since hospital information systems came into use [7], [23]. In the context of exploratory data analysis, data mining can play a twofold role. First, some of its results can serve as conjectures for further research. Due to the exhaustive search that data mining methods perform, they can sometimes suggest interesting conjectures that would not be considered otherwise [21], [23]. Second, the high degree of automation inherent to data mining makes the initial exploratory analysis of the available data less time consuming, thus freeing the data analyst to focus on the ultimate confirmatory analysis [2], [14].

In addition, the importance of data mining for the management of hospitals and healthcare centres should be mentioned. It concerns mining data about costs and revenues, personal and material resources, quantity and quality of the provided care, bed occupancy etc. In this respect, data mining plays its more traditional role – to replace the overwhelming amount of data through a manageable number of association rules, classifications, fitted models and the like [4], [9], [18], [32].

3. Our own experience – the method GUHA

So far, we have tackled our medical data mining tasks using a sophisticated method called *General Unary Hypotheses Automaton (GUHA)*. It is one of the earliest methods for automated discovery of knowledge from data, theoretically elaborated and first implemented in the seventies [13]. However, it is only in the last years that, as the result of an increased performance of the ubiquitous personal computers, GUHA has found a broad acceptance, witnessed by more than a dozen new applications published since 1995.

Basically, GUHA relies on interconnecting logic and statistics. Such a connection, encountered also in the more recent data mining systems Explora and 49er [21], [34], allows to combine the strengths of both approaches:

- ♦ logic helps to reduce the complexity of the search for valid hypotheses;
- ♦ statistics helps to filter patterns inherent to the data from random influences or noise.

The interested readers are referred to the book [13] for the original GUHA method, and to [12], [15], [16], [27], [28] for its most important later extensions, developed under the participation of the present authors.

The hypotheses generated by GUHA are sentences of a *observational calculus*, for which *dichotomous data matrices* are used as *models*, which in turn are viewed as realizations of *random samples*. By means of common dichotomization transformations of nominal or ordinal features, GUHA can handle those kinds of features as well. *A monadic observational predicate calculus* (MOPC) is a collection of a finite number of *unary predicates*, and of an at most countable number of *generalized quantifiers*. From predicates, *open formulae* can be built by means of the usual connectives ¬, & and ∨, whereas generalized quantifiers yield *closed formulae* when applied to open formulae.

As an example, the following two sentences were among those output by GUHA in a research study of epilepsy, reported in [33]:

SEX = MALE & DISEASE DURATION > 10 YEARS & ONLY GRAND MAL SEIZURES $\sim^{F}$ MEMORY QUOTIENT > 90,

DISEASE DURATION > 10 YEARS & ONLY GRAND MAL SEIZURES & DISEASE COURSE = GOOD $\sim^{F}$ MEMORY QUOTIENT > 90.

Both sentences are closed formulae of a MOPC containing the unary predicates SEX = MALE, DISEASE DURATION > 10 YEARS, ONLY GRAND MAL SEIZURES, DISEASE COURSE = GOOD, and MEMORY QUOTIENT > 90. The first four of them are used to build the open formulae on the left side of the symbol $\sim^{F}$, denoting the *Fisher quantifier*. This generalized quantifier captures performing, on the considered data matrix, the one-sided Fisher exact test of independence in contingency tables. Its truth function is defined

$$\mathrm{Tr}_{\sim_{\mathrm{F}}}(M) = \begin{cases} 1 & \text{if} \quad a_M \geq n_M s \;\&\; a_M d_M \geq b_M c_M \;\&\; \displaystyle\sum_{i=a_M}^{\min(r_M, k_M)} \frac{\binom{k_M}{i}\binom{l_M}{r_M - i}}{\binom{n_M}{r_M}} \leq \alpha \\ 0 & \text{else,} \end{cases}$$

where M is a two-column model, viewed as a realization of a random sample, a_M, b_M, c_M, d_M are cardinalities of occurrences of the vectors (1,1), (1,0), (0,1), (0,0), respectively, among the rows of M (i.e., $a_M + b_M + c_M + d_M = n_M$, where n_M is the sample size), $s \in (0,1]$ is a prescribed minimal support for a sentence to be generated, and $\alpha \in (0,½]$ is a given constant capturing the significance level of the test.

As a typical data mining method, GUHA tests each hypothesis separately. However, it can be combined with multiple hypotheses testing methods (e.g., Bonferroni, Hochberg, Holm, Shaffer, Simes) to arrive to sets of sentences valid at global or multiple significance levels.

Among the recent applications of GUHA, there were four belonging to the area of medical data mining:

- In the *biochemistry of cancer cells*, hypotheses generated by GUHA helped to elucidate some of the mechanisms driving the rise of metastases of a primary tumour [20].
- In a *clinical research of epilepsy*, GUHA was applied to search possible relationships of 13 clinical variables commonly used to characterize the disease to the occurrence of memory disturbances in epileptic patients [33].
- In a *preventive care screening study*, GUHA helped to find dependencies between the life style, the family history, and the state of health of first-year medical students [26].
- In the *psychotherapeutic medicine*, the GUHA method has been used to investigate the patterns of conflict episodes of clinically inconspicuous persons [25], which should serve as a control-group information for clinical research studies of various psychic diseases.

4. The diversity requirement and distributed object technologies

In the application domains in which data mining has been most successful so far, it typically has been used for some specific repeatedly performed task, such as analysing consumer behaviour, or proposing bank clients likely to take high credits [8], [24]. In medicine, the situation is completely different. Different research studies may lead to different data mining tasks (e.g., classification, clustering, association rules mining, sequence mining). Moreover, even in the case of a particular task, different methods may be appropriate (e.g., Bayesian classification, decision trees, or artificial neural networks in the case of classification).

Even large data mining suites (such as Decision Support Suite, Enterprise Miner, Intelligent Miner, or MineSet), besides being usually too expensive for the healthcare sector, do not implement every existing or newly emerging method. Therefore, successful application of data mining in medical research requires diverse data mining tools to be simultaneously accessible, or to be easily added if necessary.

A general solution to the diversity requirement provide distributed object technologies (DOTs), i.e. middleware technologies allowing clients to pass requests to objects implementing diverse data mining functionality in a distributed environment. Such a situation is schematically depicted in Figure 1.

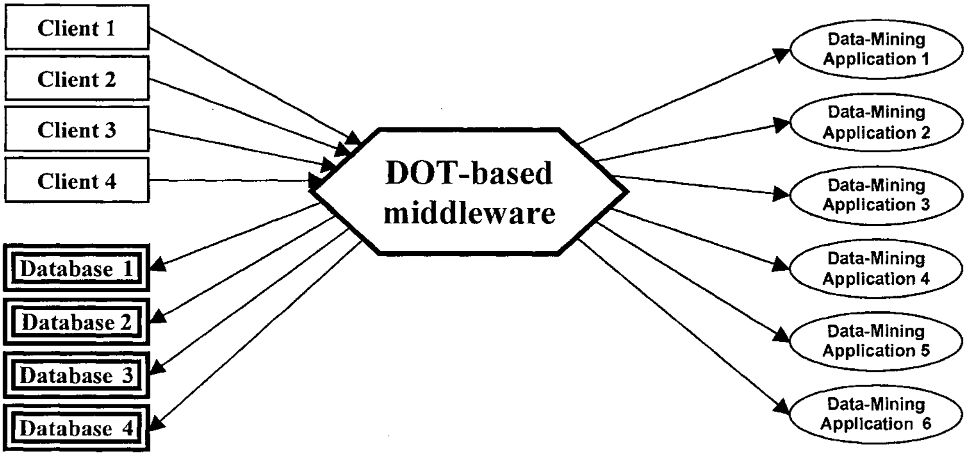
Client 1
Client 2
Client 3
Client 4
Database 1
Database 2
Database 3
Database 4
DOT-based middleware
Data-Mining Application 1
Data-Mining Application 2
Data-Mining Application 3
Data-Mining Application 4
Data-Mining Application 5
Data-Mining Application 6

Figure 1: Data mining using a DOT-based middleware

Developers of data mining tools increasingly often realize the need for an underlying DOT, and they start providing their products with corresponding interfaces. Recently, we have performed a market analysis of available data mining tools supporting some DOT. The main results of that analysis are summarized in Table 1.

Table 1: Important data mining tools supporting distributed object technologies.
Based on [1], [5], [11], [19], [30], [31] and the web pages of the producers.

data-mining tool	producer	supported DOT	basic feature(s)
ActiveSuite	Sheridan	ActiveX	concentrates on hierarchical data; mainly OLAP
Daisy (Data Analysis Interactively)	Daisy Analysis	ActiveX	conceived as an ActiveX component for interactive, mainly graphical, data analysis
Darwin / Spectrum	Thinking Machines / Cabletron	CORBA	the Darwin tool above the Spectrum data warehouse
Decision Centre	Imperial College / Fujitsu	CORBA, JavaBeans	both DOTs used to invoke data mining applications; databases accessed via JDBC
Decision Support Suite	Pilot Software	ActiveX	local / remote (in web) mining; mainly OLAP capabilities
DynamiCube	Hallogram	ActiveX	an ActiveX component; mainly OLAP capabilities
Expression / Sequence Explorer	Millenium Pharmaceuticals	CORBA	specialized for RNA expression + DNA sequence mining
Infosleuth	MCC	CORBA, JavaBeans	data mining agent – one of many agents, which are invoked either via RMI, or via CORBA
JWAVE	Visual Numerics	JavaBeans	PV-WAVE applications for data mining and visualisation
KnowledgeStudio	Angoss	ActiveX	8 own methods + bi-directional interface to SAS
OASIS	UCLA Data Mining Laboratory	CORBA	for mining scientific data; Java GUI above CORBA
ONISA	Frauenhofer Computer Graphics Institute	CORBA	CORBA-based metaserver routes data mining requests to database servers
SuperNova	Meditech	ActiveX	only simple data mining, but specialized for medicine
VisualAge	IBM	JavaBeans	for data mining against DB2 Universal Databases
XpertRule	Attar Software	ActiveX	association rules and decision trees; SQL client-server

5. Conclusion and consequences for our data mining strategy

This paper has dealt with medical data mining in the context of specific requirements of medical research. We have sketched our experience with medical data mining using the method GUHA and have summarized the main results of our recent comprehensive market analysis of data mining tools supporting DOTs.

Each of the DOTs supported by the analysed data mining tools has its own advantages and disadvantages. The most important of them are juxtaposed below in Table 2. Observe that each DOT allows access to quite a large spectrum of legacy applications. Therefore, they can be actually used with a much broader class of data mining tools than only those listed in Table 1. In particular, we want to use CORBA to enable simultaneous access to GUHA and two tools based on artificial neural networks and fuzzy logic. Our decision to use CORBA as the underlying DOT is due to our previous engagement in promoting CORBA in the medical domain [17].

Table 2: Main pros and contras of the distributed object technologies supported by data mining tools.

DOT	advantages	disadvantages
ActiveX	easy to use, access to all applications supporting OLE (often encountered in the medical domain)	not fully object-oriented, proprietary
CORBA	fully object-oriented, medical standardization (CORBAmed) access to applications in IDL-mapped languages (C, C++, Java...),	complex, little support by important software producers
JavaBeans	fully object-oriented, access to all Java applications + to CORBA applications via the Java-IDL mapping	dependent on the Java language, complex

References

[1] A. Al-Attar. Data mining – beyond algorithms. Technical report, Attar Software, 1998.

[2] C. Chatfield. Model uncertainty, data mining and statistical inference. Journal of the Royal Statistical Society. Series A, 158:419–466, 1995.

[3] P. Cheeseman and J. Stutz. Bayesian classification (autoclass): Theory and results. In [8], pages 153–180.

[4] M.S. Chen, J. Han, and P.S. Yu. Data mining: An overview from a database perspective. IEEE Transactions on Knowledge and Data Engineering, 8:866–883, 1996.

[5] T. Clark. Response to CORBAmed life sciences research request for information. Technical report, Millenium Pharmaceuticals, Inc., 1997.

[6] F.T. Denton. Data mining as an industry. The Review of Economics and Statistics, 67:124–127, 1985.

[7] J. Dudeck, B. Blobel, W. Lordieck, and T. Bürkle, editors. New Technologies in Hospital Information Systems. IOS Press, Amsterdam, 1997.

[8] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors. Advances in Knowledge Discovery and Data Mining. AAAI Press, Menlo Park, 1996.

[9] U.M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery: An overview. In [8], pages 1–36.

[10] W.J. Frawley, G. Piatetsky-Shapiro, and C.J. Matheus. Knowledge discovery in databases: An overview. In G. Piatetsky-Shapiro and W. Frawley, editors, Knowledge Discovery in Databases, pages 1–27. AAAI Press, Menlo Park, 1991.

[11] Y. Guo, D. Yang, and S. Hedvall. Software architecture for DecisionCentre: A web-based distributed data mining system. Technical report, Imperial College London, 1998.

[12] P. Hájek. The new version of the GUHA procedure ASSOC (generating hypotheses on associations) – mathematical foundations. In COMPSTAT 1984 – Proceedings in Computational Statistics, pages 360–365, 1984.

[13] P. Hájek and T. Havránek. Mechanizing Hypothesis Formation. Springer-Verlag, Berlin, 1978.

[14] D. Harmancová, M. Holeňa, and A. Sochorová. Overview of the GUHA method for automating knowledge discovery in statistical data sets. In [24], pages 39–52.

[15] M. Holeňa. Exploratory data processing using a fuzzy generalization of the GUHA approach. In J.F. Baldwin, editor, Fuzzy Logic, pages 213–229. John Wiley and Sons, New York, 1996.

[16] M. Holeňa. Fuzzy hypotheses for GUHA implications. Fuzzy Sets and Systems, 98:101–125, 1998.

[17] M. Holeňa and B. Blobel. Healthcare information system approaches based on middleware concepts. In [7], pages 178–185.

[18] M. Holsheimer and A. Siebes. Data mining. The search for knowledge in databases. Technical report, CWI, Amsterdam, 1994.

[19] M. Jern. Information drill-down using web tools. Technical report, Advanced Visual Systems, 1998.

[20] J. Kaušitz, P. Kulliffay, B. Puterová, and L. Pecen. Prognostic meaning of cystolic concentrations of ER, PS2, Cath-D, TPS, TK and cAMP in primary breast carcinomas for patient risk estimation and therapy selection. To appear in International Journal of Human Tumor Markers.

[21] W. Klösgen. Explora: A multipattern and multistrategy discovery assistant. In [8], pages 249–272.

[22] M.C. Lovell. Data mining. The Review of Economics and Statistics, 65:1–12, 1983.

[23] N. Lavrac, E.T. Keravnou, and B. Zupan. Intelligent data analysis in medicine and pharmacology. Proceedings of the First International Workshop. Kluwer Academic Publishers, Dordrecht, 1996.

[24] M. Noirhomme-Fraiture, editor. Proceedings of KESDA'98 – International Conference on Knowledge Extraction from Statistical Data, 1998.

[25] D. Pokorný and A. Sochorová. Do we use stereotypes in our relationship to the others? Technical report, Dept. of Psychotherapy, University of Ulm, 1996.

[26] H. Provazníková, N. Štullerová, and J. Štuller. Health and options in preventive care for first year students at the 3rd medical faculty of Charles University in Prague. To appear in Journal of American College Health.

[27] J. Rauch. Logical problems of statistical data analysis in data bases. In Proceedings of the Eleventh International Seminar on Data Base Management Systems, pages 53–63, 1988.

[28] J. Rauch. Logical calculi for knowledge discovery in databases. In J. Komorowski and J. Żytkov, editors, Principles of Data Mining and Knowledge Discovery, pages 47–57. Springer-Verlag, Berlin, 1997.

[29] A. Siebes. KESO: data mining and statistics. In [24], pages 1–13.

[30] M.A. Sokolewicz. ONISA: The open networked information system architecture. Technical report, Frauenhofer Computer Graphics Institute, 1998.

[31] Visual Numerics, Inc. JWAVE System Introduction, 1998.

[32] M.J. Zaki, S. Parathasarathy, M. Ogihara, and W. Li. New parallel algorithms for fast discovery of association rules. Data Mining and Knowledge Discovery, 1:343–373, 1997.

[33] J. Zvárová, J. Preiss, and A. Sochorová. Analysis of data about epileptic patients using GUHA method. International Journal of Medical Informatics, 45:59–64, 1997.

[34] J.M. Żytkov and R. Zembowicz. Contingency tables as the foundation for concepts, concept hierarchies and rules: The 49er system approach. Fundamenta Informaticae, 30:383–399, 1997.