

Biomedical Data Mining

N. Peek¹; C. Combi²; A. Tucker³

¹Department of Medical Informatics, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands;

²Department of Computer Science, University of Verona, Verona, Italy;

³School of Information Systems, Computing and Mathematics, Brunel, London, UK

Keywords

Data mining, machine learning

Summary

Objective: To introduce the special topic of *Methods of Information in Medicine* on data mining in biomedicine, with selected papers from two workshops on Intelligent Data Analysis in bioMedicine (IDAMAP) held in Verona (2006) and Amsterdam (2007).

Methods: Defining the field of biomedical data mining. Characterizing current developments and challenges for researchers in the field. Reporting on current and future activities of IMIA's working group on Intelligent Data Analysis and Data Mining. Describing the content of the selected papers in this special topic.

Results and Conclusions: In the biomedical field, data mining methods are used to develop clinical diagnostic and prognostic systems, to interpret biomedical signal and image data, to discover knowledge from biological and clinical databases, and in bio-surveillance and anomaly detection applications. The main challenges for the field are i) dealing with very large search spaces in a both computationally efficient and statistically valid manner, ii) incorporating and utilizing medical and biological background knowledge in the data analysis process, iii) reasoning with time-oriented data and temporal abstraction, and iv) developing end-user tools for interactive presentation, interpretation, and analysis of large datasets.

Correspondence to:

Niels Peek
Department of Medical Informatics
Academic Medical Center
University of Amsterdam
P.O. Box 22700
1100 DE Amsterdam
The Netherlands
E-mail: n.b.peek@amc.uva.nl
Methods Inf Med 2009; 48: 225–228

What Is Data Mining?

The goal of this special topic is to survey the current state of affairs in biomedical data mining. Data mining is generally described as the (semi-)automatic process of discovering interesting patterns in large amounts of data [1–4]. It is an essential activity to translate the increasing abundance of data in the biomedical field into information that is meaningful and valuable for practitioners. Traditional data analysis methods, such as those originating from statistics, often fail to work when datasets are sizeable, relational in nature, multimedial, or object-oriented. This has led to a stormy development of novel data analysis methods that are increasingly receiving attention in the biomedical literature.

Data mining is a young and interdisciplinary field, drawing from fields such as database systems, data warehousing, machine learning, statistics, signal analysis, data visualization, information retrieval, and high-performance computing. It has been successfully applied in diverse areas such as marketing, finance, engineering, security, games, and science. And rather than comprising a clear-cut set of methods, the term “data mining” refers to an eclectic approach to data analysis where choices are led by pragmatic considerations concerning the problem at hand.

Broadly speaking, the goals of data mining can be classified into two categories: *description* and *prediction* [2–4]. Descriptive data mining attempts to discover implicit and previously unknown knowledge, which can be used by humans in making decisions. In this case, data mining is part of a larger knowledge discovery process that includes data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, and presentation of discovered knowledge to end-users. To arrive at usable results, it is essential that the discovered patterns are comprehensible by humans. Typical descriptive data mining tasks are unsupervised machine learning problems such as mining frequent

patterns, finding interesting associations and correlations in data, cluster analysis, outlier analysis, and evolution analysis.

Predictive data mining seeks to find a model or function that predicts some crucial but (yet) unknown property of a given object, given a set of currently known properties. In prognostic data mining, for instance, one seeks to predict the occurrence of future medical events before they actually occur, based on patients' conditions, medical histories, and treatments [5]. Predictive data mining tasks are typically supervised machine learning problems such as regression and classification. Well-known supervised learning algorithms are decision tree learners, rule-based classifiers, Bayesian classifiers, linear and logistic regression analysis, artificial neural networks, and support vector machines. The models that result from predictive data mining may be embedded in information systems and need not, in that case, to be always comprehensible by humans, even though a sound motivation of the provided prediction is often required in the medical field.

The distinction between descriptive and predictive data mining is not always clear-cut. Interesting patterns that were found with descriptive data mining techniques can sometimes be used for predictive purposes. Conversely, a comprehensible predictive model (e.g. a decision tree) may highlight interesting patterns and thus have descriptive qualities. It may also be useful to alternate between descriptive and predictive activities within a data mining process. In all cases, the results of descriptive and predictive data mining should be valid on new, unseen data in order to be valuable to, and trusted by, end-users.

Data Mining in Biomedicine

Data mining can be applied in biomedicine for a large variety of purposes, and is thus connected to diverse biomedical subfields. Traditionally, data mining and machine learning applications focused on clinical applications, such as decision support to medical practitioners and interpretation of signal and image data. More recently, applications in epidemiology, bioinformatics, and biosurveillance have received increasing attention.

Clinical data mining applications are mostly predictive in nature and attempt to derive models that use patient-specific information to predict a patient's diagnosis, prognosis, or any other outcome of interest and to thereby support clinical decision-making [6]. Historically, diagnostic applications have received most attention [7–9], but in the last decade prognostic models are becoming more popular [5, 10, 11]. Other tasks that are addressed with clinical data mining are detection of data artifacts [12] and adverse events [13], discovering homogeneous subgroups of patients [14], and extracting meaningful features from signal and image data [15].

Several characteristic features of clinical data may complicate the data mining process, such as the frequent and often meaningful occurrence of missing values, and the fact that data values (e.g. diagnostic categories) may stem from structured and very large medical vocabularies such as the ICD [16]. Furthermore, when the data were collected in routine care settings, it may be misleading to draw conclusions from the data with respect to causal effects of therapies. Data from randomized controlled studies enable researchers to compute unbiased estimates of causal effects, as these studies ensure exchangeability of patient groups [17]. In observational data, however, the analysis is biased due to the lack of this exchangeability.

In recent years, biomedical data mining has received a strong impulse from research in molecular biology. In this field, datasets fall into three classes: i) sequence data, often represented by a collection of single nucleotide polymorphisms (SNPs) [18]; ii) gene expression data, which can be measured with DNA microarrays to obtain a snapshot of the activity of all genes in one tissue at a given time [19], and iii) protein expression data, which can include a complete set of protein profiles obtained with mass spectra technologies, or a few protein markers [20]. Initially, most genomic and proteomic research focused upon working with individual data sources and achieved considerable success. However, a number of key barriers have been met concerning for example the variability in microarray data [21] and the enormous search spaces involved with identifying protein-protein interactions and folding which require substantial data samples. An alternative approach to dealing directly with ge-

nomomic and proteomic data is to perform literature mining which aims to discover related genes and proteins through analysis of biomedical publications [22]. Recent developments have explored methods to combine data sources such as meta-analysis and consensus algorithms for homogenous data [23] and Bayesian priors for heterogeneous data [24]. Another major recent development that aims to combine data and knowledge is system biology. This is an emerging field that attempts to model an entire organism (or a major system within an organism) as a whole [25]. It is starting to show genuine promise when combined with data mining [26], particularly in certain biological subsystems such as the cardiovascular and immune systems.

Although many data mining concepts are today well-established and toolsets are available to readily apply data mining algorithms [27, 28], various challenges remain for researchers in the biomedical data mining field. First and foremost, biomedical datasets continue to grow in terms of the number of variables (measurements) per patient. This results in exponentially growing search spaces of hypotheses that are explored by data mining algorithms. An important challenge is to deal with these very large search spaces in a manner that is both computationally efficient and statistically valid. Second, knowledge discovery activities are only meaningful when they take advantage of existing background knowledge in the application area at hand [29]. Biomedical knowledge typically resides in structured medical vocabularies and ontologies, clinical practice guidelines and protocols, and results from scientific studies. Few existing data mining methods are capable of utilizing any of these forms of background knowledge. Third, a most characteristic feature of medical data is its temporal dimension. All clinical observations and interventions must occur at some point in time or during a time period, and the medical jargon abounds with references to time and temporality [30]. Although the attention to temporal reasoning and data analysis has increased over the last decade [31–33], there is still a lack of established data mining methods that deal with temporality. The fourth and final challenge is the development of software tools for end users (such as biologists and medical professionals). With the growing amounts of data available, there is an

increasing need for interactive tools that support users in the presentation, interpretation, and analysis of datasets.

IMIA's Working Group on Intelligent Data Analysis and Data Mining

In 2000, a Working Group on Intelligent Data Analysis and Data Mining was established as part of the International Medical Informatics Association (IMIA) [34]. The objectives of the working group are i) to increase the awareness and acceptance of intelligent data analysis and data mining methods in the biomedical community, ii) to foster scientific discussion and disseminate new knowledge on AI-based methods for data analysis and data mining techniques applied to medicine, iii) to promote the development of standardized platforms and solutions for biomedical data analysis, iv) to provide a forum for presentation of successful intelligent data analysis and data mining implementations in medicine.

The working group's main activity is organization of a yearly workshop called Intelligent Data Analysis in bioMedicine and Pharmacology (IDAMAP) [35]. IDAMAP workshops are devoted to computational methods for data analysis in medicine, biology and pharmacology that present results of analysis in the form communicable to domain experts and that somehow exploit knowledge of the problem domain. Typical methods include data visualization, data exploration, machine learning, and data mining. Gathering in an informal setting, workshop participants have the opportunity to meet and discuss selected technical topics in an atmosphere that fosters the active exchange of ideas among researchers and practitioners. IDAMAP workshops have been organized since 1996. The most recent workshops were held in Aberdeen (2005), Verona (2006), Amsterdam (2007), and Washington (2008).

Other activities of the working group include the organization of tutorials and panel discussions at international conferences on the topics of intelligent data analysis and data mining in biomedicine. In all its activities, there is a close collaboration with the Working Group on Knowledge Discovery and Data Mining of AMIA [36].

Selected Papers

From a total of 35 papers presented at the IDAMAP-2006 and IDAMAP-2007 workshops, the ten best papers were selected based on the workshop review reports, and the authors were invited to submit an extended version of their paper for the special topic. Eight authors responded positively, from which five papers were finally accepted after blinded peer review. To our opinion, these papers form a representative sample of the current developments in biomedical data mining.

The paper by Curk et al. [37] considers the problem of *knowledge discovery from gene expression data*, by searching for patterns of gene regulation in microarray datasets. Knowledge of gene regulation mechanisms is essential for understanding gene function and interactions between genes. Curk et al. present a novel descriptive data mining algorithm, called *rule-based clustering*, that finds groups of genes sharing combinations of promoter elements (regions of DNA that facilitate gene transcription). The main methodological challenge is the vast number of candidate combinations of genes and promoter regions, which is handled by the algorithm by employing a heuristic search method. Interesting features of this algorithm are that it yields a symbolic cluster representation, and, in contrast to traditional clustering techniques, allows for overlapping clusters.

Also Bielza et al. [38] discuss on the analysis of microarray gene expression data, but focus on predictive data mining, using *logistic regression analysis*. As discussed in the previous section, microarray datasets have created new methodological challenges for existing data analysis algorithms. In particular, the number of data attributes (genes) is typically much larger than the number of observations (patients), potentially resulting in unreliable statistical inferences due to a severe 'multiple testing' problem. One popular solution in biostatistics is *regularization* of the model parameters by setting a penalty on total size of the estimated regression coefficients. However, estimation of regularized model parameters is a complex numeric optimization problem. The paper by Bielza et al. presents an evolutionary algorithm to solve the problem.

The third paper, by Andreassen et al. [39], is clinically oriented and uses a Bayesian

learning method to solve a well-known problem in pharmacoepidemiology: discovering which bacterial pathogenic organisms can be treated with particular antibiotic drugs. Again, the large number of possible combinations that need to be considered poses problems for traditional data analysis methods. More specifically, many pathogen-drug combinations will not even occur in the data, or in such small numbers that reliable direct inferences are not possible. Andreassen et al. propose to solve this problem by borrowing statistical strength from observations on similar pathogens using *hierarchical Dirichlet learning*.

Castellani et al. [40] consider the identification of tumor areas in *dynamic contrast enhanced magnetic resonance imaging* (DCE-MRI), a technique that has recently expanded the range and application of imaging assessment in clinical research. DCE-MRI data consists of serial sets of images obtained before and after the injection of a paramagnetic contrast agent. Rapid acquisition of images allows an analysis of the variation of the MR signal intensity over time for each image voxel, which is indicative for the type of tissue represented by the voxel. Castellani et al. use *support vector machines* to classify the signal intensity time curves associated with image voxels.

The fifth and final paper of the special topic, written by Klimov et al. [41], deals with the *visual exploration of temporal clinical data*. They present a new workbench, called VISITORS (VISUALization of Time-Oriented Records), which integrates knowledge-based temporal reasoning mechanisms with information visualization methods. The underlying concept is the *temporal association chart*, a list of raw or abstracted observations. The VISITORS system allows users to interactively visualize temporal data from a set of patient records.

References

1. Fayyad U, Piatetsky-Shapiro G, Smyth P. The KDD process for extracting useful knowledge from volumes of data. *Commun ACM* 1996; 39 (11): 27–34.
2. Hand DJ, Mannila H, Smyth P. *Principles of Data Mining*. Cambridge, Massachusetts: MIT Press; 2001.
3. Giudici P. *Applied Data Mining Statistical Methods for Business and Industry*. London: John Wiley & Sons; 2003.

4. Han J, Kamber M. *Data Mining. Concepts and Techniques*. San Francisco, California: Morgan Kaufmann Publishers; 2006.
5. Abu-Hanna A, Lucas PJ. Prognostic models in medicine: AI and statistical approaches. *Methods Inf Med* 2001; 40 (1): 1–5.
6. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. *Int J Med Inform* 2008; 77 (2): 81–97.
7. Lavrac N, Kononenko I, Keravnou E, Kukar M, Zupan B. Intelligent data analysis for medical diagnosis: using machine learning and temporal abstraction. *AI Commun* 1998; 11: 191–218.
8. Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. *Artif Intell Med* 2001; 23: 89–109.
9. Sakai S, Kobayashi K, Nakamura J, Toyabe S, Akazawa K. Accuracy in the diagnostic prediction of acute appendicitis based on the Bayesian network model. *Methods Inf Med* 2007; 46 (6): 723–726.
10. Pfaff M, Weller K, Woetzel D, Guthke R, Schroeder K, Stein G, Pohlmeier R, Vienken J. Prediction of cardiovascular risk in hemodialysis patients by data mining. *Methods Inf Med* 2004; 43 (1): 106–113.
11. Tjortjis C, Saraee M, Theodoulidis B, Keane JA. Using T3, an improved decision tree classifier, for mining stroke-related medical data. *Methods Inf Med* 2007; 46 (5): 523–529.
12. Verduijn M, Peek N, de Keizer NF, van Lieshout EJ, de Pont AC, Schultz MJ, de Jonge E, de Mol BA. Individual and joint expert judgments as reference standards in artifact detection. *J Am Med Inform Assoc* 2008; 15 (2): 227–234.
13. Jakkula V, Cook DJ. Anomaly detection using temporal data mining in a smart home environment. *Methods Inf Med* 2008; 47 (1): 70–75.
14. Nannings B, Bosman RJ, Abu-Hanna A. A subgroup discovery approach for scrutinizing blood glucose management guidelines by the identification of hyperglycemia determinants in ICU patients. *Methods Inf Med* 2008; 47 (6): 480–488.
15. Lessmann B, Nattkemper TW, Hans VH, Degenhard A. A method for linking computed image features to histological semantics in neuropathology. *J Biomed Inform* 2007; 40 (6): 631–641.
16. www.who.int/whosis/icd10. Last accessed Mar 3, 2009.
17. Hernán MA. A definition of causal effect for epidemiological research. *J Epidemiol Community Health* 2004; 58: 265–271.
18. Barker G, Batley J, O'Sullivan H, Edwards KJ, Edwards D. Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. *Bioinformatics* 2003; 19 (3): 421–422.
19. Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. *J Comput Biol* 2000; 7 (3–4): 601–620.
20. Lobley A, Swindells MB, Orengo CA, Jones DT. Inferring function using patterns of native disorder in proteins. *PLoS Comput Biol* 2007; 3 (8): e162.
21. Choi JK, Yu U, Kim S, Yoo OJ. Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics* 2003; 19 (Suppl 1): i84–i90.
22. Jelier R, Schuemie MJ, Roes PJ, van Mulligen EM, Kors JA. Literature-based concept profiles for gene annotation: the issue of weighting. *Int J Med Inform* 2008; 77 (5): 354–362.
23. Steele E, Tucker A. Consensus and meta-analysis regulatory networks for combining multiple microarray gene expression datasets. *J Biomed Inform* 2008; 41 (6): 914–926.
24. Husmeier D, Werhli AV. Bayesian integration of biological prior knowledge into the reconstruction of gene regulatory networks with Bayesian networks. In: Markstein P, Xu Y, editors. *Computational Systems Bioinformatics, Volume 6: Proceedings of the CSB 2007 Conference*. World Scientific; 2007. pp 85–95.
25. Kitano H. *Computational systems biology*. Nature 2002; 420: 206–210.
26. Zhang M. Interactive analysis of systems biology molecular expression data. *BMC Systems Biol* 2008; 2: 2–23.
27. Witten IH, Frank E. *Data Mining. Practical Machine Learning Tools and Techniques*. San Francisco, California: Morgan Kaufmann Publishers; 2005.
28. <http://www.ailab.si/orange>. Last accessed Mar 3, 2009.
29. Zupan B, Holmes JH, Bellazzi R. Knowledge-based data analysis and interpretation. *Artif Intell Med* 2006; 37 (3): 163–165.
30. Shahar Y. Dimensions of time in illness: an objective view. *Ann Intern Med* 2000; 132 (1): 45–53.
31. Combi C, Shahar Y. Temporal reasoning and temporal data maintenance in medicine: issues and challenges. *Comput Biol Med* 1997; 27 (5): 353–368.
32. Adlassnig KP, Combi C, Das AK, Keravnou ET, Pozzi G. Temporal representation and reasoning in medicine: Research directions and challenges. *Artif Intell Med* 2006; 38 (2): 101–113.
33. Stacey M, McGregor C. Temporal abstraction in intelligent clinical data analysis: a survey. *Artif Intell Med* 2007; 39 (1): 1–24.
34. <http://magix.fri.uni-lj.si/idadm>. Last accessed Mar 3, 2009.
35. <http://www.idamap.org>. Last accessed Mar 3, 2009.
36. <http://www.amia.org/mbrcenter/wg/kddm>. Last accessed Mar 3, 2009.
37. Curk T, Petrovic U, Shaulsky G, Zupan B. Rule-based clustering for gene promoter structure discovery. *Methods Inf Med* 2009; 48: 229–235.
38. Bielza C, Robles V, Larrañaga P. Estimation of distribution algorithms as logistic regression regularizers of microarray classifiers. *Methods Inf Med* 2009; 48: 236–241.
39. Andreassen S, Zalounina A, Leibovici L, Paul M. Learning susceptibility of a pathogen to antibiotics using data from similar pathogens. *Methods Inf Med* 2009; 48: 242–247.
40. Castellani U, Cristani M, Daducci A, Farace P, Marzola P, Murino V, Sbarbati V. DCE-MRI data analysis for cancer area classification. *Methods Inf Med* 2009; 48: 248–253.
41. Klimov D, Shahar Y, Taieb-Maimon M. Intelligent interactive visual exploration of temporal associations among multiple time-oriented patient records. *Methods Inf Med* 2009; 48: 254–262.