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

Knowledge Discovery for Advanced Clinical Data Management and Analysis

Ankica Babic[1,2]
1. Medical Informatics, Dept of Biomedical Engineering, Linkoping University, Sweden
2. Faculty of Electrical Engineering, University of Ljubljana, Slovenia

Abstract: Knowledge discovery is a broad research field in which methods are developed to support discovery of novel and potentially useful knowledge from clinical databases and registers in systems for patient care. However, the techniques available are not readily applicable in medical domains, due to, among other reasons, low user friendliness and lack of proper methodological background.
Data mining approaches to be explored and improved are predictive modelling, segmentation, dependency modelling, summarisation, and change and deviation detection/modelling (in data or knowledge).
Another and original contribution of the research is to build up efficient feedback loops. Human experts and available domain expert systems could provide suggestions as how to improve all major steps in the knowledge discovery process such as evaluation of knowledge, choice of data mining methods and data input.
A long tradition of collecting and maintaining clinical and administrative data could be found in fields of oncology, cardiology, coronary surgery, social and primary health care medicine All these areas, that gather data over long periods of time, could benefit from knowledge discovery.

1. Introduction to the field

Knowledge discovery is a process of identifying novel, valid and understandable patterns in data. This is done in order to explain existing data, make predictions or classifications about new data, and summarise the contents of large database to support decision making. Logical data visualisation, which is a part of the process, helps human experts in better understanding of data and in discovering deeper patterns. The approach requires less from the end user allowing him easier hypothesis building compared to conventional statistical methods.
Data mining, which is essential to the whole process, can handle large amounts of data and offer efficient techniques to discover new data patterns which can be visualised in a manner easily understandable by the users. In a multidimensional database, data mining typically begins when even the most advanced query approaches can no longer provide sought answers. It is used to extract general descriptive knowledge (generative models of data, symbolic descriptions of subsets, summarisation), as well as discriminative knowledge (to distinguish between K classes, for accurate classification, to separate spaces).
Practical realisation of the knowledge discovery process is dependent both on the area of the application and the choice of clinical and other tasks to be supported. Therefore, good results can be expected only for well-defined research problems. In all other cases, quality of data management and knowledge extraction will strongly depend on end user skills and support which might be given by knowledge engineers. The most important issues to be taken care of in implementing any knowledge discovery process are given in Table 1 [1].

Table 1. Practical tasks in knowledge discovery.

• automation,
• data transformation and dimensionality reduction,
• guarding against overfitting,
• modes of growth of data,
• interface to derive simplified forms of the extracted models for comprehensibility and visualisation,
• efficient and sufficient sampling schemes,
• in-memory vs. disk-based data processing,
• choice of optimal subset of techniques to span most tasks,
• interfaces to large data warehouses, and use of metadata to optimise access,
• client-server issues, where to perform the processing,
• exploiting parallelism, distributed computing over a network of computers.

The important direction concerns a selection of suitable methods from the data mining library. Depending on the nature of the knowledge discovery task, users preferences and other factors, the following considerations are to be taken into account:

- univariate vs. multivariate analysis,
- numerical data vs. categorical or mixed data,
- explanation requirements or comprehensibility,
- fuzzy vs. precise patterns,
- sample independence assumptions,
- availability of prior knowledge.

Practical motivation to build up an integrated environment for knowledge discovery has a foundation in the following facts:

- data volume is too large for classical analysis regimes,
- networking, increased opportunity for access by many end users,
- end users are also physicians and other non-statisticians.

2. Advanced methodological solutions

Machine learning techniques that have an ability to provide new knowledge have undergone intensive development [2,6]. Many of them provide solutions to complex questions usually offering flexible software implementation and valuable graphical presentation of the results. End users are not, however, generally aware of all the possible disadvantages of choosing a particular method. In addition, an attractive graphical presentation could mislead interpretation of the results. Both these aspects could become an obstacle for knowledge discovery. Therefore, we intend to study effects of learning from the clinical data. Table 2. presents quite detailed performance features to be considered in upgrading of the methods and minimisation of risks of inaccurate outcomes.

Table 2. Classification data methods as represented by their performance features.

	Advantages	Disadvantages
Decision Trees	• Can deal with high-dimensional data • Breaks data into symbolic partition (leaves, or rules) • Fast execution time with greedy search • Easy to implement	• Quick partitioning of data results in fast deterioration in attribute selection quality • Greedy search is too blind • Limited representation language • Search space is huge, even for limited node tests
Neural Networks	• If examples have attached class probabilities, and densities are smooth, then problem becomes a regression problem and neural nets have been shown to be effective at learning to predict actual probabilities accurately • Representational power • Implicit massive parallelism, though hard to implement in many architectures	• Training neural net to be consistent with training set is known to be NP-hard problem (for classification, NOT regression problems) • Network architecture choice (number of nodes, hidden layers, so forth) are by trial and error • Solution depends on initial weight settings • Learned function not easy to understand, virtually a black box classifier that offers no explanations to humans
Nearest Neighbour Method	• Deal with inexact matches of features • Explanation in terms of records used to reach a decision	• Computationally expensive: all records in the database being mined have to be "touched". Feature weight calculations are dynamic. The method, especially the first variant, is good for small databases, or sparse feature sets • Identifying k: Determining how many records to return can be tricky • Defining "most appropriate": The definition is problem dependent A computable similarity metric may be hard to define • No good method for deciding weights on attributes exists • Generalisation is hard
Bayesian Clustering	• Does not require distance measures • Gives probability of membership in each class rather than a single class assignment • Has natural measure of fit of models to data: probability if the provide data given the model assumptions • Forces user to make all assumptions explicit	• Requires selecting class models a priori. While this is difficult, it is actually better than making this assumption implicitly as any other program would do • Search model parameter space is very large • Have to worry about co-variances between variables

3. Evaluation insights and further development with respect to clinical demands

Common examples of knowledge discovery in oncology are navigation and searching in the databases, hypothesis testing, summarising experiment results, creating predictions and

discrimination of the patient groups. Scientific data analyses puts even stronger demands on the whole process: basic processing, high-level science analysis and, finally, scientific discovery over large data sets.

In our research work we have developed and tested methods of several approaches to extract sought clinical knowledge [7]. Experiences of using multivariate statistics were enriched by implementing procedures of artificial intelligence [8,9,10]. In the field of asymptotic liver disease we have researched and strengthened a complete knowledge discovery cycle from the data to a standardised knowledge representation form [11]. In the field of oncology we have developed procedures for effective graphical representation that gave excellent insights into data clusters [12]. We have designed and implemented customised classification procedures when standard statistical techniques have not proven to be efficient enough [13].

The growing potential of knowledge discovery [1,15-18] enables development of a newer, advanced methods and efficient integration of already existing multivariate methods. *Cluster, discriminant and regression analyses* constitute a powerful analytical toolbox. Case based reasoning is another methodology that could provide significant support trough the integration into a knowledge base system.

4. Empowering knowledge discovering process

Expert validation and evaluation is a normal, required step in declaring any new facts and patterns as novel and relevant knowledge. Figure 1. suggests how this step of quality control could be implemented in an automated manner.

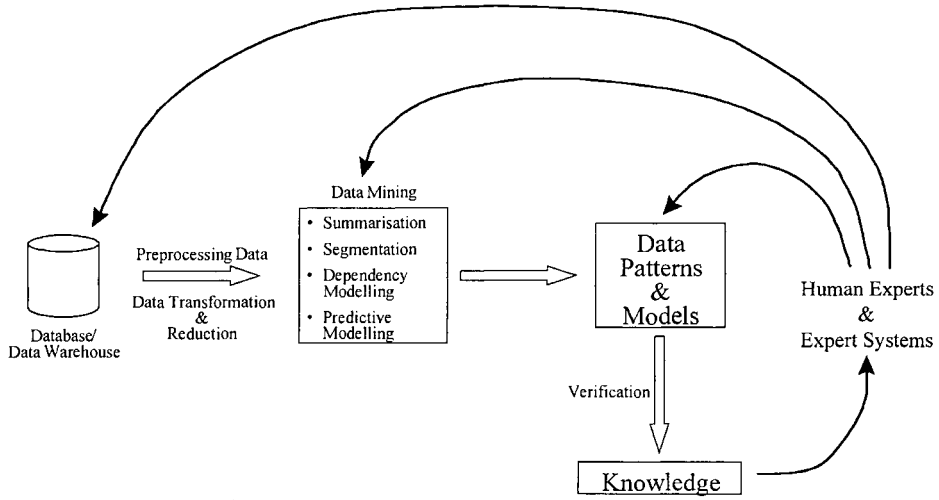

Figure 1. Main directions of updating knowledge discovery process.

Including both human experts and available domain expert systems into the process could provide a feedback information to practically all steps of the process. By analysing obtained data models, or even patterns, a user could come up with new explanations of the knowledge, or go back to either chose another method or select more suitable data. In this manner, and as described in [19], an evaluation of all suggested data analytical methods could be done.

Oncology is just one of the important research and application domains where awareness of

potentials and usefulness of proposed research could be raised. Equally important is to propagate results of the research into other domains where data is collected and maintained, e.g. cardiology, coronary surgery [20,21], and many branches of social medicine [22].

References

[1] Fayyad U., Piatetsky-Shapiro G. and Smyth P.(1995), "From Knowledge Discovery to Data Mining: An overview. a chapter in Advances in Knowledge Discovery and Data Mining, U M Fayyad, G Piatetsky-Shapiro, P J Smyth and R Uthurusamy (Eds.) AAAI/MIT Press

[2] Wu X, Knowledge Acquisition from Data Bases, PhD Thesis 176, pp, Dept of Artificial Intelligence, University of Edinburgh, Edinburgh, Scotland, 1993.

[3] Wu X., K Eshell2: An Intelligent Learning Data Base System, Research and Development in Expert Systems IX, M A Bramer and R W Milne (Eds) Cambridge University Press, UK 1992.

[4] Wu X., Inductive Learning: Algorithms and Frontiers, Artificial Intelligence review, 7 (1993), 2:93-108.

[5] Wu X, Research Issues in Intelligent Learning Database Systems, Proceedings of the Seventh Annual Florida AI Research Symposium, Pensacola Beach, Florida, USA May 5-7, 1994.

[6] Wu X., Book: Knowledge Acquisition from databases, Ablex, USA, 1995.

[7] Babic Ankica. *Medical knowledge extraction, Application of data analysis methods to support clinical decisions*. PhD. dissertation. No. 322, Linköping University, 1993.

[8] Babic A, Aahlfeld H, Wigertz O, Bodemar G, Mathiesen U. Artificial neural networks in clustering and classification data on unspecified liver diseases. XIth Nordic Meeting on Medical and Biological Engineering, Lund, Sweden, p 136, 1993.

[9] Babic A, Krusinska E, Strömberg J.-E. Extraction of diagnostic rules using recursive partitioning systems. A Comparison of Approaches. Artificial Intelligence in Medicine, 4:373-387, Elsevier, 1992.

[10] Babic A. Case Studies in Machine Learning for Medical Knowledge Extraction. The SAIS94, Swedish artificial Intelligence workshop, Ronneby, Sweden, 1994

[11] Babic A, Bodemar G, Mathiesen U, Aahlfeldt H, Franzen L, Wigertz O. Machine Learning to Support Diagnostics in the Domain of Asymptomatic Liver Disease. In R.A. Greeneset al. (Eds) MEDINFO'95 Proceedings, IMIA, 8th Wrold Congress on Medical Informatics, Vancouver, Canada, pp 809-813, 1995

[12] Zganec M, Babic A., M Us-Krasovec, B Palcic. 3D Presentation of the nuclear cell features in quantitative cytometry. In Proc AMIA Annual Fall Symposium, Washington DC, James J Cimino, ed. Hanley & Belfus, 1996, pp 679-683.

[13] Siska A, Babic A., Pavesic N., Studies of responsiveness in the R-EGFR and Rcerb-B2 oncogenes spaces: a customized application for thyroid lesions. In Proc MIE'97, Technology and Informatics 43, C Pappas, N Maglaveras, J-R Scherrer eds, Greece. IOS Press, Netherlands, 1997, pp 634-637.

[14] Matheus C, Chan P, G Piatetsky-Shapiro, Systems for Knowledge Discovery in Databases, Special Issue on Learning and Discovery in Databases, IEEE Transactions on Knowledge and Data Engineering, Dec. 1993

[15] Parsaye K. and Chignell M., 1993. Intelligent Database Tools & Applications, John Wiley.

[16] N Cerone and M Tsuchiya, guest editors, Special Issue on Learning and Discovery in Databases, IEEE Transactions on Knowledge and Data Engineering, 5(6) Dec 1993.

[17] G Piatetsky-Shapiro, guest editor, Special issue on Knowledge Discovery in Databases and Knowledge Bases, International Journal of Intelligent Systems, Vol 7, no 7, Sep 1992,

[18] G Piatetsky-Sharpio, guest editor , Special issue on Knowledge Discovery in Databases, Journal of Intelligent Information Systems, 3(4), Dec 1994,.

[19] Harrell Jr FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;**15**:361-87.

[20] Higgins TL, Estafanous FG, Loop FD, Beck GJ, Blum JM, Paranandi L. Stratification of morbidity and mortality outcome by preoperative risk factors in coronary artery bypass patients. A clinical severity score [published erratum appears in jama 1992 oct 14;268(14):1860] [see comments]. *JAMA* 1992;**267**:2344-8.

[21] O'Connor G.T., Plume S.K., Olmstead E.M., Morton J.R., Maloney C.T., Nugent W.C., Hernandez F., Clough R., Leavitt, B.J., Coffin L.H., Marrin C.A.S., Wennberg J.E., Birkmeyer J.D., Charlesworth D.C., Malenka D.J., Quinton H.B., Kasper J.F., A Regional Intervention to Improve the Hospital Mortality Associated With Coronary Artery Bypass Graft Surgery, *JAMA* 1996, 275: 841-6.

[22] Iezzoni LI, Ash AS, Shwartz M, Daley J, Hughes JS, Mackiernan YD. Predicting who dies depends on how severity is measured: implications for evaluating patient outcomes [see comments]. *Ann Intern Med* 1995;123:763-70.