Part II: Tools for Supporting Basic Creative Processes

4 Knowledge Acquisition by Machine Learning and Data Mining

Tu Bao Ho, Saori Kawasaki[1] and Janusz Granat[2]

4.1 Introductory Remarks

A critical problem in the development of knowledge-based systems is capturing knowledge from the experts. There are many knowledge elicitation techniques that might aid this process, but the fundamental problem remains: tacit knowledge that is normally implicit, inside the expert's head, must be externalised and made explicit. Knowledge acquisition (KA) thus has been well recognised as a bottleneck in the development of knowledge-based systems and is a key issue in knowledge engineering. Traditionally, KA techniques can be grouped into three categories: manual, semi-automated (interactive), and automated (machine learning and data mining). Since the early days of artificial intelligence (AI), the problem of knowledge acquisition, the elicitation of expert knowledge in building knowledge bases, has been recognised as a fundamental issue in knowledge engineering.

Fifteen years ago the Encyclopaedia of Artificial Intelligence wrote, "Acquiring and modelling knowledge effectively can be the most time-consuming portion of the knowledge engineering process. Little methodology is practiced beyond unstructured interviewing. Automated methods are, for the most part, still in the research stage." (Shapiro 1992, Vol. 1, p. 719).Much has changed

[1] School of Knowledge Science Japan Advanced Institute of Science and Technology 1-1 Asahidai, Nomi-city, Ishikawa, 923-1292 Japan
[2] National Institute of Telecommunications, Szachowa 1, 04-894 Warsaw, Poland

since that day. On the one hand, various modelling methodologies and tools for KA have been constructed; for example, CommonKADS, a comprehensive methodology that covers the complete route of KA,[3] is widely used by companies and educational institutions (Schreiber et al., 2000). One the other hand, the last decade has witnessed much progress in machine learning (ML) research and applications (Bishop 2006), (Langley and Simon 1995), (Mitchell 1997), and in the emerging interdisciplinary field of knowledge discovery and data mining (KDD), or, in short, data mining (Han and Kamber 2006, Hand et al. 2001). There have been numerous achievements in machine learning and data mining, such as automated methods, in the context of knowledge acquisition.

The use of machine learning and data mining can ease the knowledge acquisition problem. Experts may find it hard to say what rules they use to assess a situation, but they can usually tell you what factors they take into account. A machine learning and data mining program can take descriptions of situations couched in terms of these factors and then infer rules that match the behaviour of the expert. The expert can then critique these rules and verify that they seem reasonable (it is usually easier to recognise correct rules than to generate them). If the rules are wrong, the expert may be able to suggest counter-examples that can guide further learning.

This chapter aims to show that the progress in machine learning and data mining research has made them valuable knowledge acquisition tools. We start by introducing the basic concepts of machine learning and data mining and follow by describing some progress in these fields. We then address and illustrate some problems and results in scientific data mining as a tool for scientific knowledge acquisition. We also address some of the major opportunities and challenges of machine learning and data mining in the task of knowledge acquisition.

[3] "Apparently complete" would be more precise, since this methodology does not yet fully respond to the challenge of *human centred knowledge acquisition* discussed later.

There is, however, one caveat: too strong a concentration on the further development of automated machine learning and data mining tools might result in missing a basic point: knowledge acquisition can be done with the goal of making computers learn more and become more intelligent *only if it also helps people gain more knowledge and make correct decisions.* Thus, the focus should not only be on machine learning, but also on human learning and decision making in the interaction with a data mining computer. This is a very clear conclusion from data mining and knowledge acquisition applications in the very demanding field of modern telecommunications, as reported in detail in one of the later sections. From these applications, new challenges will result in more *human centred* knowledge acquisition.

4.2 Machine Learning, Knowledge Discovery and Data Mining

Machine learning (ML), a broad subfield of AI, is concerned with the development of algorithms and techniques for building computer systems that can "learn" (Bishop 2006, Mitchell 1997). Essentially, we can say that in machine learning methods, computers learn new knowledge from supervised and unsupervised data. Machine learning research began in the 1960s, and soon had various applications in diverse domains (Langley and Simon 1995). The annual International Conference on Machine Learning (ICML), first held in 1982, is one of the highest quality and most competitive conferences in computer science. The European conference on machine learning (ECML), started in 1989, is another high-quality event in the field.

Knowledge discovery and data mining (KDD), also called knowledge discovery in databases or shortly data mining (Han and Kamber 2006, Hand et al. 2001) emerged in the early 1990s as an interdisciplinary field, and became widely recognised at the first ACM SIGKDD international conference in Montreal in 1995. Data mining is concerned with the development of algorithms and techniques to extract knowledge from large and complex databases.

Machine learning and data mining share the same broad goal of finding novel and useful knowledge in data, and thus they have most techniques and processes in common. The fundamental difference between machine learning and data mining exists in volume of data being processed. However, both require theoretical soundness and experimental effectiveness.

Throughout this chapter we will illustrate diverse notions with real-world databases. Below are examples from the meningitis database collected at the Medical Research Institute, Tokyo Medical and Dental University from 1979 to 1993. This database contains data of patients who suffered from meningitis and who were admitted to the departments of emergency and neurology in several hospitals. Table 4.1 presents attributes used in this database.

Categories	Type of attributes	# attributes
Present History	Numerical and Categorical	7
Physical Examination	Numerical and Categorical	8
Laboratory Examination	Numerical	11
Diagnosis	Categorical	2
Therapy	Categorical	2
Clinical Course	Categorical	4
Final Status	Categorical	2
Risk Factor	Categorical	2
Total		38

Table 1 Attributes in the meningitis database

Below are two patient data records in this database that have mixed numerical and categorical data, as well as missing values (denoted by "?"):

10, M, ABSCESS, BACTERIA, 0, 10, 10, 0, 0, 0, SUBACUTE, 37,2, 1, 0,
15, -, -6000, 2, 0, abnormal, abnormal, -, 2852, 2148, 712, 97, 49, F,
-, multiple, ?, 2137, negative, n, n, n

12, M, BACTERIA, VIRUS, 0, 5, 5, 0, 0, 0, ACUTE, 38.5, 2,1, 0,
15, -, -, 10700, 4, 0, normal, abnormal, +, 1080, 680, 400, 71,
59, F, -, ABPC+CZX, ?, 70, negative, n, n, n

A pattern discovered from this database in the language of IF-THEN rules is given below; the quality of the pattern is measured by its accuracy (in this case, 87.5%):

```
IF          Poly-nuclear cell count in CFS <= 220
      and   Risk factor = n
      and   Loss of consciousness = positive
      and   When nausea starts > 15
THEN        Prediction = Virus [accuracy = 87.5%]
```

The process of knowledge discovery can be viewed as inherently consisting of five steps, as shown in Fig. 4.1 (essentially, the same applies to a machine learning process). The main tasks in each step of the KDD process are shown in Fig. 4.2.

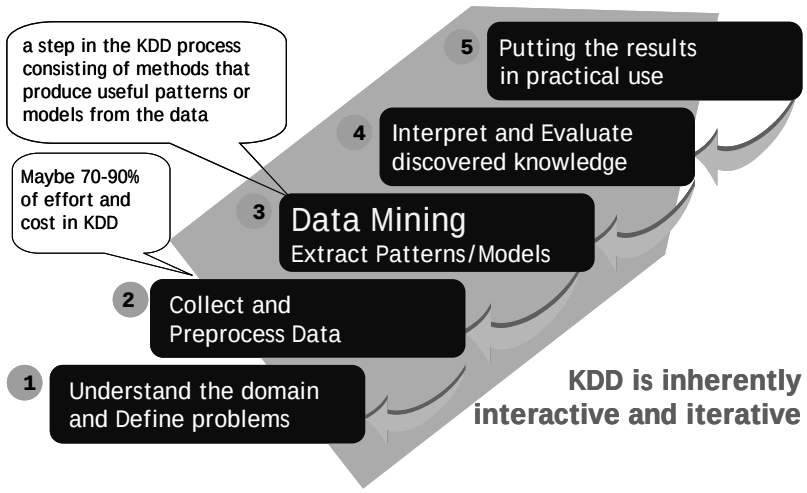

Fig. 4.1: The KDD process

The first step is to understand the application domain and formulate the problem. This step is clearly a prerequisite for extracting useful knowledge and for choosing appropriate machine learning and data mining methods in the third step according to the application target and the nature of the data.

The second step is to collect and preprocess the data, including the selection of the data sources, the removal of noise or outliers, the treatment of missing data, the transformation (discretisation if necessary) and reduction of data, etc. This step usually takes most of the time needed for the whole KDD process.

The third step is learning and data mining to extract patterns and/or models hidden in the data. A model can be viewed as a

global representation of a structure that summarises the systematic components underlying the data or that describes how the data may have arisen. In contrast, a pattern is a local structure, perhaps relating to just a handful of variables and a few cases. The major classes of data mining methods are: predictive modelling such as classification and regression; segmentation (clustering); dependency modelling, such as graphical models or density estimation; summarisation, such as finding the relations between fields; association; visualisation; and change and deviation detection and modelling in data and knowledge.

The fourth step is to interpret (post-process) the discovered knowledge, especially in terms of description and prediction—the two primary goals of discovery systems in practice. Experiments show that discovered patterns or models from data are not always of interest or direct use, and the KDD process is necessarily iterative with the judgment of discovered knowledge. One standard way to evaluate induced rules is to divide the data into two sets, training on the first set and testing on the second. One can repeat this process a number of times with different splits, and then average the results to estimate the rules performance.

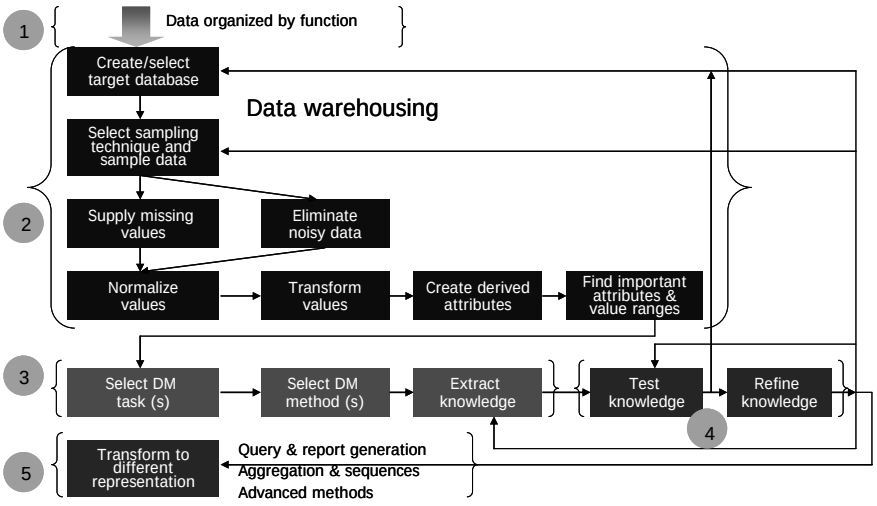

Fig. 4.2: Main tasks in each step of the KDD process

The final step is to put discovered knowledge into practical use. Sometimes, one can use discovered knowledge without embedding it in a computer system. In other cases, the user may expect that the discovered knowledge can be put on computers and exploited by various programs. Putting the results into practical use is certainly the ultimate goal of knowledge discovery.

We suggest a good view of data mining by considering two dimensions; one is the type of data to be mined and the other is the mining tasks and methods. Typically, various kinds of data are stored in different data schemes (Han and Kamber 2006):

- Flat data tables
- Relational databases
- Object-oriented databases
- Object-relational databases
- Transactional databases
- Spatial databases
- Temporal databases and time-series databases
- Text databases and multimedia databases
- Heterogeneous databases and legacy databases
- World Wide Web data

Data mining tasks and methods basically can be divided into two groups: classification with prediction and description. Classification with prediction is the process of finding a set of models or patterns or functions that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. Description is the process of characterizing the general properties of the data in a database.

Typical methods of classification with prediction include (Bishop 2006, Han and Kamber 2006, Hand et al. 2001, Mitchell 1997):

- *Decision tree induction* aims to find classification/prediction models in the tree structure. Typical decision tree methods are C4.5 and CART. Current research on decision trees concentrates on converting large trees into sets of rules, tree visualisation, and data access (to very large databases).

♦ *Neural networks* are information processing devices that consist of a large number of simple nonlinear processing modules, connected by elements that have information storage and programming functions. Extracting or making sense of numeric weights associated with the interconnections of neurons to come up with a higher level of knowledge has been and will continue to be a challenging problem in data mining.

♦ *Bayesian inference* is a statistical inference in which evidence or observations are used to update or to newly infer the probability that a hypothesis may be true. The name Bayesian comes from the frequent use of the Bayes theorem in the inference process. The most widely used methods are Naïve Bayesian classification, assuming that attributes are all independent, and Bayesian belief networks, assuming that dependencies exist among subsets of attributes. Representing dependencies among random variables by a graph in which each random variable is a node and the edges between the nodes represent conditional dependencies, is the essence of the graphical models that are playing an increasingly important role in machine learning and data mining (Jordan 1998).

♦ *Rule induction* produces a set of IF-THEN rules from a database. Unlike decision tree methods that employ the "divide-and-conquer" strategy, rule induction methods usually employ the "separate-and-conquer" strategy. Some popular methods include CN2, IREP, RIPPER, and LUPC (Ho and Nguyen 2003, Pham and Ho 2007).

♦ *Hidden Markov models* (HMM) – a widely used finite-state-machine method – are statistical models in which the system being modeled is assumed to be a Markov process with unknown parameters, and the challenge is to determine the hidden parameters from the observations. Recent finite-state-machine methods, including maximum entropy Markov models (MEMM) and conditional random fields (CRFs), have shown high performance in various structured prediction problems (Phan et al. 2005).

♦ *Support vector machines* (SVMs) are typical kernel methods that also apply linear classification techniques to non-linear classification problems by increasing their dimensionality (Nguyen and Ho 2006a, 2007, Schölkopf and Smola 2001, Tran

and Ho 2006a, 2007, Schölkopf and Smola 2001, Tran et al. 2006).

- etc.

Following are some typical description methods (Bishop 2006, Han and Kamber 2006, Hand et al. 2001, Mitchell 1997):

- *Association rule mining*, which aims to discover elements that co-occur frequently within a dataset consisting of multiple independent selections of elements (such as purchasing transactions), and to discover rules, such as implication or correlation, which relate co-occurring elements. Questions such as "if a customer purchases product A, how likely is he to purchase product B?" and "what products will a customer buy if she buys products C and D?" are answered by association mining algorithms. Typical association mining algorithms are Apriori (Agrawal et al., 1994) and FP-tree (Han et al., 2001).

- *Clustering* seeks to identify a finite set of categories or clusters to describe the data. The categories may be mutually exclusive and exhaustive, or consist of richer representations such as hierarchical or overlapping categories. Examples of clustering in a knowledge discovery context include discovering homogeneous sub-populations for consumers in marketing databases and the identification of sub-categories of spectra from infrared measurements. Data mining research focuses on efficient and effective clustering methods for large and complex databases (scalability, complex shapes and types of data, high dimensional clustering, mixed numerical and categorical data, etc.).

- *Summarisation* involves methods for finding a compact description for a subset of data. A simple example would be tabulating the mean and standard deviations for all fields. More sophisticated methods involve the derivation of summary rules, multivariate visualisation techniques, and the discovery of functional relationships between variables. Summarisation techniques are often applied to interactive exploratory data analysis and automated report generation.

It is worth noting that data mining methods are rather specialised; for example, the decision tree algorithm C4.5, originally de-

signed for flat data tables, should be changed appropriately when it is applied to different types of data, such as text or sequential data.

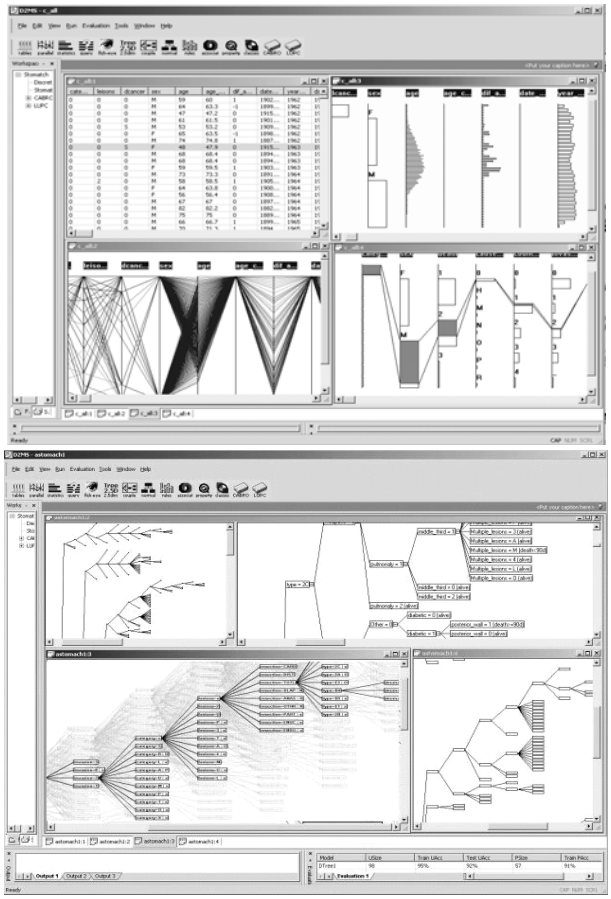

Fig. 4.3: Visual data mining system D2MS (Ho et al. 2003a)

Visualisation has proven its effectiveness in exploratory data analysis and has high potential in data mining. Various methods and systems have been developed for visualizing large datasets and discovered knowledge (large decision trees, huge numbers of associations, etc.) as well as visualizing the knowledge discovery process. They facilitate and support the active role of the user in

all knowledge discovery steps, from preprocessing to interpretation and evaluation (Fayyad et al. 2001, Ho et al. 2003b).

Finding scalable algorithms is an essential task in data mining in order to deal with huge datasets. An algorithm is said to be scalable if, given an amount of main memory, its runtime increases linearly with the number of input instances. Massively parallel processing is another strategy for dealing with huge datasets when the algorithm in nature cannot be so nearly linear to be scalable (Freitas et al. 1998). The increased computing power allows solving many problems in practice with advanced methods, such as computing second order conditional random fields on large databases (Phan et al. 2005).

4.3 Examples of Progress in Machine Learning and Data Mining

In a study presented at the IEEE International Conference on Data Mining in 2005 (ICML'05) many active and influential researchers were interviewed to identify the ten most challenging problems in this field. Here are these ten problems:

1. Developing a unifying theory of data mining
2. Scaling up for high dimensional data/high speed streams
3. Mining sequence data and time series data
4. Mining complex knowledge from complex data
5. Data mining in a network setting
6. Distributed data mining and mining multi-agent data
7. Data mining for biological and environmental problems
8. Data-mining-process related problems
9. Security, privacy and data integrity
10. Dealing with non-static, unbalanced, and cost-sensitive data

Various new techniques have been developed to attack these problems. In the last few years, kernel methods, graphical models, and semi-supervised learning have emerged among the most exciting research directions in machine learning and data mining.

Kernel methods in general, and support vector machines (SVMs) in particular, are increasingly used to solve diverse prob-

lems, especially in scientific data. They offer versatile tools to process, analyse, and compare many types of data, and offer state-of-the-art performance in many cases (Nguyen and Ho 2007, Schölkopf and Smola 2001, Schölkopf et al. 2004). The big problem with simple models of linear learning machines (say, perceptrons, developed in 1956) is that of insufficient capacity, as highlighted by (Minsky and Pappert 1969). The first wave of neural networks (since the mid-1980s) overcame the problem by glueing together many thresholded linear units (multi-layer neural networks). That solved the problem of capacity, but there were training problems in terms of speed and multiple local minima. The kernel methods approach (since 2000), which retains linear functions but works in another, higher dimensional feature space, can be viewed as the second wave of linear learning machines. Kernel methods can operate on very general types of data and can detect very general relations. Overall, kernel methods have two kernel function components: a kernal matrix (to map the data from the original space into a higher dimensional feature space), and a kernel machine (to look for solutions from the kernel matrix by finding a linear or other easy pattern in the feature space, using a well-known algorithm that works on the kernel matrix). By applying an inverse map, a linear pattern in the feature space can be found to correspond to a complex pattern in the original space.

Graphical models are a marriage between graph theory and probability theory (Jordan 1998). They clarify the relationship between neural networks and related network-based models such as hidden Markov model (HMMs), Markov random fields (MRFs), Kalman filters, conditional random fields (CRFs), etc. Typical advantages of graphical models are: inference and learning are treated together; supervised and unsupervised learning are merged seamlessly; missing data is handled nicely; there is a focus on conditional independence; and there is a high interpretability of the results.

Semi-supervised learning (SSL) is halfway between supervised and unsupervised learning (Chapelle et al. 2006): “Traditional classifiers use only labelled data (feature/label pairs) to train. Labelled instances however are often difficult, expensive, or time consuming to obtain, as they require the efforts of experienced

human annotators. Meanwhile unlabeled data may be relatively easy to collect, but there have been few ways to use them. Semi-supervised learning addresses this problem by using a large amount of unlabeled data, together with the labeled data, to build better classifiers. Because semi-supervised learning requires less human effort and gives higher accuracy, it is of great interest both in theory and in practice".

4.4. Scientific Data Mining

Due to the rapid progress of network and data acquisition technologies in the last decades, a huge amount of data has been accumulated and archived in many scientific areas, such as astronomy, medicine, biology, chemistry, and physics. To find useful information in these data sets, scientists and engineers are turning to data analysis techniques. There has been a fundamental shift from more conventional techniques to computer-aided scientific discovery in various sciences, especially by the use of machine learning and data mining methods to exploit huge and precious scientific databases (Augen 2005, Gilbert 1991, Lacroix and Critchlow 2003, Langley and Simon 1995, Larson and Totowa 2006, Ramakrishnan and Grama 2001). It is worth noting that scientific data are essentially complexly structured data (relational data, sequences, molecules, graphs, trees, etc.) that create a number of difficult problems when being analysed (e.g., structured output interdependency, imbalanced, heterogeneous, large-scale, etc.) (Fayyad et al. 1996).

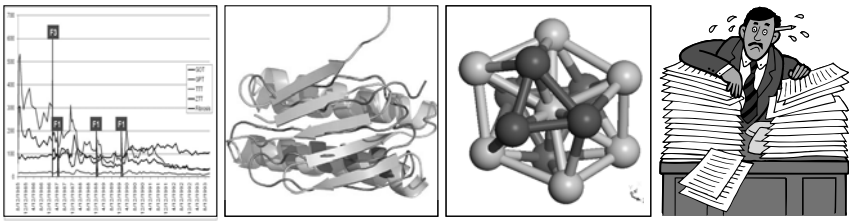

Fig. 4.4 Scientific data typically have a complex structure

This section introduces a new branch of computer science: mining scientific data (Fayyad et al. 1996, Fayyad et al. 1996,

WMSED 2006, Kawasaki and Ho 2006). On the one hand, the progress in machine learning and data mining has opened various opportunities for discovery through scientific data. On the other hand, the complexity of scientific data poses many challenging problems for data mining researchers. Importantly, the collaboration between domain experts and computer scientists is always a key factor in successful scientific data mining.

The role of scientific data mining is well recognised. “Given the success of data mining in commercial areas, it didn’t take much time for the scientists and engineers to discover the usefulness of data mining techniques in scientific disciplines. For example, analysis of massive simulation data sets generated by computational simulations of physical and engineering systems is difficult and time consuming using traditional approaches. Indeed, much of the output of computational simulations is simply stored away on disks and is never analysed at all. Availability of suitable data mining techniques can allow engineers and scientists to analyse such data and gain fundamental insights into the underlying mechanisms of the physical processes involved” (Grossman et al. 2001).

On the surface, it may appear that data from one scientific field, say genomics, is very different from another field, such as physics. Despite this diversity, there is much that is common in the mining of scientific data. For example, the techniques used to identify objects in images are very similar, regardless of whether the images came from a remote sensing application, a physics experiment, an astronomy observation, or a medical study. Further, with data mining being applied to new types of data, such as mesh data from scientific simulations, there is the opportunity to apply and extend data mining to new scientific domains.

Scientific data mining is an emerging trend, as illustrated, for example, by the series of annual workshops on scientific data mining held in the U.S. since 1998 (Ho and Nguyen 2002), or by the two medical and chemical databases selected to challenge researchers in the Grant-in-Aid for Scientific Research on Priority Areas (B) active mining, 2001-2005 (Fayyad et al. 1996). The insight from these events is that the directions in advanced machine

learning and data mining described above also are crucial in scientific data mining.

These research directions are widely viewed as theoretically attractive and empirically promising for dealing with complexly structured data, and thus with scientific data mining.

4.4.1 Mining Medical Data

Evidence-based medicine (EBM) applies the scientific method to medical practice, and aims for the ideal that healthcare professionals should make *conscientious, explicit, and judicious use of current best evidence* in their everyday practice. Generally, there are three distinct, but interdependent, areas of EBM. The first is to treat individual patients with acute or chronic pathologies by treatments supported in the most scientifically valid medical literature. The second area is the systematic review of medical literature to evaluate the best studies on specific topics. The third is the medical *movement*, in which advocates work to popularise the methods of EMB and the usefulness of its practice in public forums, in patient communities, in educational institutions, and in the continuing education of practicing professionals (Cios 2000).

Following is an example of the practical implementation of EMB. Viral hepatitis is a disease in which tissue of the liver is inflamed by the infection of hepatitis viruses. As the severity of viral hepatitis increases, so does the potential risk of liver cirrhosis and hepatocellular carcinoma (HCC) – which is the most common type of liver cancer and the fifth most common cancer. While the exact cause of HCC is still unknown, studies on viral hepatitis, especially on hepatitis types B and C, have become essential in medicine. The hepatitis relational temporal database, collected from 1982 to 2001 at Chiba University Hospital in Japan, was recently released to challenge the data mining research community. This database contains results of 983 laboratory tests on 771 patients. It is a large, un-cleansed, temporal relational database consisting of six tables, the biggest of which has 1.6 million records. The doctors posed a number of problems on hepatitis that could be investigated by KDD techniques. For the last five

years we have worked on mining the hepatitis data to solve several problems raised by physicians: for example, can we distinguish hepatitis type B and type C by clinical data; can a patient's fibrosis stage (one of five stages F0, F1, …, F4) be identified without performing a biopsy; and in which stage of viral hepatitis can interferon therapy be effective? In particular, we have developed data mining methods that exploit the most valuable sources: the hepatitis database, the most well-known medical library MEDLINE (119,315 articles on hepatitis), and medical expert knowledge (Ho et al. 2003b).

Our framework consists of four steps (Kawasaki and Ho 2006):

1. Create different transactional databases for various hepatitis problems with the proposed temporal relations extraction (TRE) algorithm.
2. Use D2MS and learning methods, such as CBA (http://www.comp.nus.edu.sg/~dm2), C5.0, etc., to find rules from the transactional database.
3. Exploit MEDLINE for background or domain knowledge in order to support the knowledge evaluation.
4. Analyse the findings with (or by) physicians

The merit of our framework is that it gives us rather refined output from the original data by combining different views: the combinations of complexly structured temporal sequences are transformed into a set of simple representations within a medical context. These enable us to apply many types of learning algorithms and their output conveys meaning to the physicians. In addition, the knowledge obtained from MEDLINE provides a key to focus on the search space and gives supportive and confident background on learned results, which prevents us from considering unlikely patterns.

Below are examples of discovered rules which were judged to be potentially new and useful for solving two problems:

> R#2 (HCV): "TTT in high state with peaks" AFTER "ZTT in high state with peaks" (support count = 86, conf. = 0.73)
>
> R#5 (HBV): "GOT in very high state with peaks" ENDS "GPT in extreme high state with peaks" (support count = 41, conf. = 0.71)

R#10 (NonLC): "GPT in very high state with peaks" AFTER "TTT in high state with peaks" AND "GOT in very high state with peaks" ENDS "GPT in very high with peaks" AND "GOT in very high state with peaks" AFTER "TTT in high state with peaks" (support count = 10, conf. = .80).

R#8 (LC): "GPT in very high state with peaks" AFTER "TTT in very high state with peaks" AND "GPT in very high state with peaks" BEFORE "TTT in high state with peaks" AND "GOT in very high state with peaks" AFTER "TTT in high state with peaks", (support count = 8, conf. = .80)

4.4.2 Mining Genomic and Proteomic Data

Recent developments in molecular biology have given the scientific community a large amount of data about the sequences and structures of several thousand proteins. This information can effectively be used for medical and biological research only if one can extract functional insights from the sequence and structural data. Bioinformatics methods are among the most powerful technologies available in life sciences today (Baclawski and Niu 2006, Baldi and Brunak 2001, Bourne and Weissig 2003, Rashidi and Buehler 2000, Wang et al. 2004). We will show how computational methods can perform some tasks that are expensive and tedious to do during experiments (Pham et al. 2005a, b, c).

We focus on the problem of protein-protein interactions (PPI). Most proteins in cell are considered not to be independent individuals. They could interact permanently or transiently with the others to function in many biological processes or biochemical events. There are three major ongoing trends in PPI. The first is predicting and classifying whether a pair of proteins is interacting or not. The second is determining the features of PPI as, e.g., biological or biochemical or physiological features. The last is inferring the biological functions of interacting protein partners and of the PPI networks as well.

We addressed the first task of PPI study, that of predicting protein interaction. In our work on PPI, we used multiple genomic/proteomic databases and applied a PPI prediction of protein domain-domain interactions (DDI) with inductive logic

programming (ILP). ILP is one of the most effective classification techniques; it allows integrating diverse data types in terms of predicates. The output rules of ILP were also considered and used to discover the relations between DDI and other genomic and proteomic information. With the predicted domain-domain interactions, we have developed a novel approach which combines Bayesian networks and ILP (called probabilistic ILP) to infer protein-protein interaction networks (Tran et al 2005, Nguyen and Ho 2006b, c).

4.4.3 Mining Materials Science Data

We addressed the following problem of mining materials science data in terms of two processes (Ho et al. 2004).

In the forward process, a researcher postulates a molecular structure or a material formula and then wants to predict what properties that structure or formula will have. The inverse process is just the opposite: Researchers enter the properties they are looking for, and the system gives them a molecular structure or formula that is likely to have those properties. The inverse process cannot begin until the forward model is completed because the former depends on information in the model.

Our goal was to find optimised structures of PtRu nano-clusters (a promising catalyst for use in fuel cells) by combining data mining methods and *ab initio* calculations on generated structures of PtRu nano-clusters. In fuel cell systems that use H2 and O2 gas as fuel, CO molecules are known to deactivate the catalytic function of the Pt bimetal catalysts. This deactivation process is called CO poison. A weaker binding of the CO molecule on a PtRu cluster may lead to a more efficient catalyst for fuel cells. Thus, finding the structure of PtRu nanoclusters that minimises the CO adsorption energy is a significant task in nanocatalyst design.

Our method consists of two phases: one is to generate a database of the structures of PtRu bimetal nanoclusters (with a size smaller than 1nm), and the other is to find in this database the optimised structure of the PtRu bimetal nanocluster which has the lowest CO adsorption energy.

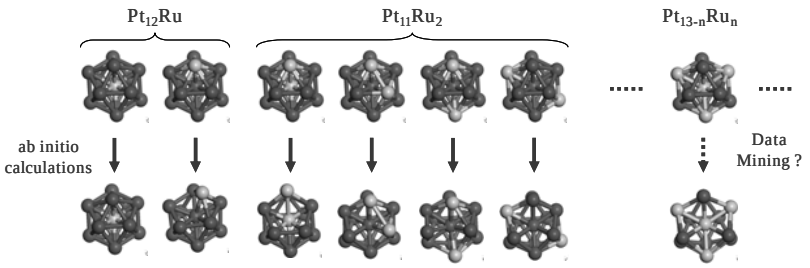

Fig. 4.5 Construction of optimised structures of PtRu bimetal clusters

In short, in this section we have shown various examples of successful or on-going projects on mining scientific data, as well as the recent research directions in machine learning/data mining that offer advanced techniques to deal with complexly structured data; in this way we have tried to clarify some of the different opportunities and challenges in the very promising field of scientific data mining.

4.5 Experiences of Data Mining in Telecommunications

One of the most challenging commercial applications of data mining relates to telecommunications systems and operators. We will outline some experiences with such applications at the National Institute of Telecommunications in Warsaw, Poland. They illustrate some further challenges beyond those perceived by data mining experts as the ten most challenging problems in this field (Section 4.3).

Existing data repositories in the telecommunications industry are huge, often measured in terabytes. The large amount of existing data necessitates the use of some guiding principles or information in order to organise the data analysis process. Here we use the term “data analysis” as a general term that can be understood as simple data analysis, exploration, and data mining. Such combination is necessary in solving complex industrial problems. Moreover a generalised interactive approach must be applied: the data analysis process is performed repetitively, using several interactions not only of data mining (between analysts and the data mining system), but also between analysts and other stakeholders

of the results (managers, the decision makers). This generalised interactive approach does not actually correspond to pure data mining according to its usual definition. However, the point is that for complex problems (at least in industry), the generalised interactive approach is the only way to obtain useful results in practice.

There are diverse classes of knowledge that are actually utilised in the process of data analysis:

- knowledge about problem being solved;
- detailed knowledge about the sources of data (in telecommunications, there are often more than 10 different data sources);
- knowledge about the preprocessing methods and attributes being used for data mining;
- knowledge about specific data mining methods and algorithms;
- transformation of the output of the algorithms into useful knowledge that is understandable and usable by the stakeholders of the results (a customised report, for example, or the generation of inputs for other information systems).

While all these types of knowledge might be utilised in the process of data mining, the actual success of this process depends on a good specification of its goals, including the types and forms of knowledge to be created. In the interactive approach, knowledge is created in a process of interrelated steps, feedbacks, and recursions, forming a network of interconnected steps rather than the one-dimensional chain suggested by some authors. The essential question relates to the type of the final output and to the final form for representing the created knowledge. Is it a formal form of knowledge representation or rather, do we need a textual description supplemented with diverse graphs? Such questions must be answered for each case. Below we present a review of the possible forms of knowledge representation in data mining:

1. For simple data analysis, the most popular and simplest way of presenting data is a report including:
   a. Statistics: tables summarizing the results of statistical analysis

   b. Diverse graphic forms of data representation, such as bar charts, maps, etc.
   c. Textual summaries and conclusions
   d. Customised combinations of the above (selected data, graphs and texts)
2. Output of multidimensional data analysis: graphs, multidimensional bar charts, etc.
3. Output of data mining algorithms together with visualisation, including:
   a. Decision trees
   b. Decision rules, etc.

A deep understanding of the raw data is a key to successfully reaching the goal in a data mining process. Achieving a satisfactory level of understanding of raw data in a new business environment is often time-consuming and requires extensive interaction with personnel working in many different departments. Raw input data must be preprocessed before the data mining algorithms are applied. The extent and scope of the preprocessing depends on the goals of the knowledge acquisition process and on the method assumed to be used in data mining. Raw data are stored in diverse sources: databases, text files, electronic documents, etc. We can distinguish business databases, with a subdivision into operational and analytical databases. Since databases in market companies are built in order to support business processes, they often contain the most significant information and knowledge about the business, but this knowledge is often hidden or obscure. Contemporary databases are usually components of complex systems, such as ERP (Enterprise Resources Planning), CRM (Customer Relationship Management), etc. Systems of this type focus on operational management. Another group of systems contains analytical systems that provide the aggregated information used to support decision making. Analytical databases are usually implemented as data warehouses. Some analytical databases are specifically built for data mining purposes; in this case, they contain sets of data mining tables.

The experiences in the telecommunication industry concern diverse tasks and, in general, very successful applications of data mining; however, we concentrate here on two problems: *segmen-*

tation of clients and *identification of significant events*. The first of these problems illustrates the complex interaction process and the second one focuses on providing significant information in real time.

An Example of Complex Interaction Process

Segmentation of clients is one of the basic tasks of marketing departments, particularly in telecommunication operations. Segmentation can have many purposes; for example, different marketing strategies are often used for different segments of the market. In the process of segmentation, we create new knowledge that results from a combination of preexisting business knowledge, analytical knowledge, and data mining (Granat 2004, Granat and Wierzbicki 2004). Fig. 4.6 shows the basic components of a segmentation process.

The goal of segmentation is to find segments of clients and provide profiles of each segment in a descriptive form. Each segment should have characteristics that clearly differentiate it from other segments, and should contain a sufficient but not too large number of clients, manageable by business departments.

After defining the goals of segmentation, the database of clients has to be built. Operational databases should be searched for information about clients. The availability of such data and its history should be analysed. At this stage, the decision is made as to which features should be chosen to describe each client. This depends, on one hand, on the goals of the segmentation and, on the other hand, on data that are available in the system. At this stage the analyst can provide feedback for the owner of the databases or the decision maker in the company, concerning suggestions about new data to be collected for future use. If the available data are already sufficient for segmentation purposes, then the preparation of data preprocessing scripts begins. When the scripts are ready, they have to be run in order to load actual and historical data. It should be stressed that the selection of features and the preprocessing of data require a thorough understanding of the diverse characteristics of the data and of future modelling requirements.

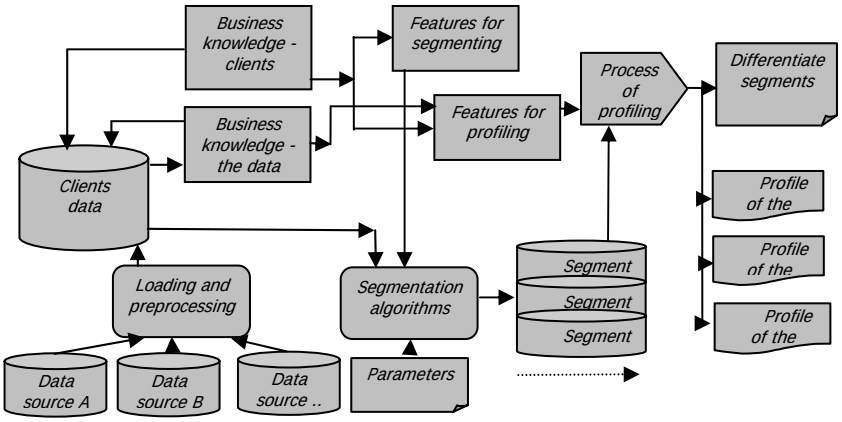

Fig. 4.6 Basic components of a segmentation process

Then we can start working on segmentation algorithms. These algorithms might be, e.g., based on business knowledge about the range of the values of some features, while simple conditional expressions can be used in the segmentation algorithm; or we can apply advanced clustering algorithms that automatically find segments of clients with similar behaviour. Usually, however, even if we are using advanced clustering algorithms, they have to be run many times with diverse parameter sets, diverse sets of features and with various supporting algorithms, such as for outlier detection. Preferably the results should be discussed with a business representative who has good knowledge of the domain. The clustering must be done by an analyst who combines a deep knowledge of the algorithms used with good domain knowledge; this helps in understanding the results of clustering as well as in interactions with the business representative.

The results of the clustering are not presented to business representatives directly. For them, each segment should be characterised by a profile – an easily understandable description of each segment. The profile should contain at least two types of information: a description of the segment and information about the main differences between this segment and others. Again, this process requires several interactions of analysts and business representatives. The description of segments cannot be prepared automatically by an algorithm. Analysts and business representatives usu-

ally possess a large amount of tacit, intuitive knowledge that cannot be directly stored and automatically utilised.

The client profiles actually constitute the knowledge that is created in the process of segmentation. It is represented in the form of written documents with tables, graphs, and texts. Such knowledge is one of the most important assets of the company, which organises its business based on such profiles. The company can assign specific groups of employees to deal with clients belonging to the most important segments, can prepare dedicated marketing campaigns for each segment, etc. We stress that running a clustering algorithm is only one step in the complex process of segmentation that requires the tacit, intuitive knowledge of the large number of people involved in the process.

Already we see in this case that, although good data mining algorithms are important and were successfully applied, more crucial is computer-human and person-to-person interaction, generally, a *human-centred process*. Since the decision maker in the company is, at the same time, the source of tacit knowledge that helps to find relevant segmentation patterns and the target for the knowledge found in data mining, the middleman – the analyst – not only must have good tools for computer-human interaction, but also must be skilled in the face-to-face exchange of tacit knowledge with the decision maker.

4.5.2 Event mining

New opportunities arise from the large amount of data that is stored in various databases. Event mining is one such challenging area of research. In this subsection we will focus on formulating an event mining task that considers observations of the system as well as internal and external events. Figure 4.7 (Granat 2005a, b) shows the interrelations between events and observation of the system that is given in time series and alarms. Sometimes, it is impossible to observe events directly. In such cases the data are stored in databases in the form of time series. This data represents observations of the system in selected points in time. The observations are analysed by the system and alarms are generated by abrupt changes in the values of observations. In the next step,

other algorithms find the events that caused changes in the system.

The following algorithms can be considered:

- For a significant change in an observation, find events that are the reasons for this change;
- Predict future events by analyzing the changes in observations;
- Predict changes in observations after an event occurs.

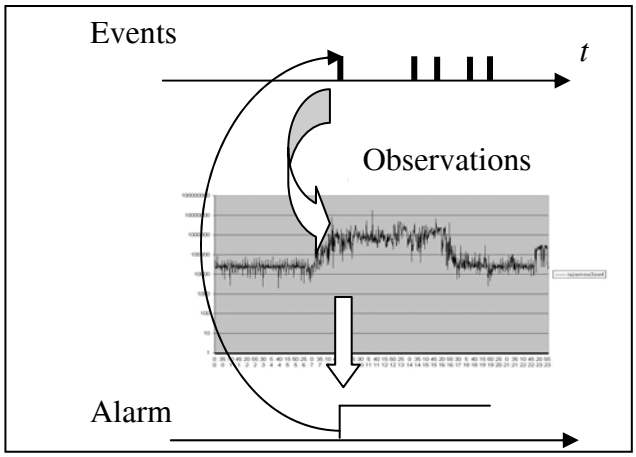

Fig. 4.7 Illustration of event detection

Identification of significant events is a highly complex task that we will only discuss here in general terms. Knowledge is acquired from data mining in order to provide decision support for company managers and other decision makers. One of the requests from the manager of a telecom company was to be instantly informed – e.g., by mobile phone – whenever a significant event occurred either in the network or in the business environment. Data mining methods can be applied for this task, but we must first define what the decision maker means by an *occurrence of an event* and is meant by a *significant event.* Even though a well developed theory of events exists in information science – e.g., based on Petri nets – the decision maker is interested only in events corresponding to his tacit knowledge, events that are significant for him. Even if the significance of events can be ranked by applying decision support methods, new problems

arise in such an approach: how to effectively combine decision support methods with data mining and knowledge acquisition,[4] and how to select decision support methods that respond well to the tacit knowledge of the decision maker (some related examples and approaches are discussed in Chapters 2 and 10 of this book). We have included this short indication of some of the issues related to the identification of significant events only to further justify the following conclusions from actual experience in applying data mining in telecom companies.

Exchanging Tacit Knowledge

The statement with which we started this chapter – that the critical problem in the development of knowledge-based systems is capturing knowledge from the experts – is not only supported by this experience, but even extended and modified. The critical problem is not only *capturing*, but also *exchanging* (including both *capturing* and also *conveying back*) knowledge with experts and decision makers. This knowledge exchange concerns tacit, intuitive knowledge, which is mostly preverbal (see Wierzbicki and Nakamori 2006). Thus it is, by definition, very difficult to verbalise – nevertheless it can be conveyed by metaphors, goal setting, prioritisation and ranking, but above all by face-to-face meetings, all methods typically used in practical management. Today the practice of successful data mining responds to this problem by exchanging tacit knowledge face-to-face between the data mining analyst and the decision maker or actual user of the acquired knowledge. The challenge is to advance further, to support such an exchange more directly with the computer.

Thus, we can conclude that the ten challenging problems identified by data mining experts and listed in Section 4.3 incom-

[4] For a very long time, the fields of decision analysis and support and of machine intelligence have been developing in close parallel, sometimes sharing researchers (such as H. Simon), but still they are distinct disciplines that do not necessarily mix well. On the other hand, there is no doubt that many successful applications of data mining use the acquired knowledge for decision support.

pletely characterise challenges in the field, perhaps because they represent a specific hermeneutic horizon (see Chapter 17) or a paradigmatic concern only with machine learning. Successful applications of data mining indicate that at least two additional problems appear as challenges:

1. *Human centred knowledge acquisition,* focusing on interaction methods between users or decision makers, analysts, and computerised data mining systems, in particular on tacit knowledge exchange;
2. *Inclusion of and interaction with user preferences* in data mining, including combination of decision support methods with data mining and knowledge acquisition, event mining, identification and ranking, etc.

4.6 Conclusions

We began this chapter with a short introduction of the basic idea, roles, and relations of knowledge acquisition, machine learning, and knowledge discovery as a way of making implicit or hidden knowledge explicit, then outlined the general requirements for making discoveries that are more than gimmicks.

After new trends of learning techniques were briefly explained, the use of those ideas for knowledge discovery practices and results in scientific databases was introduced. It was shown that, although artificial intelligence techniques have a solid theoretical background and have made great progress in the last decades, individual techniques alone are not enough for extracting expected results from real-world databases. In the solution of real world problems, the process of knowledge discovery consists of a number of steps or sub-tasks which require interaction between AI-based computational methods and human reviewers. The development of knowledge discovery methods, thus, includes not only efficient and effective data mining algorithms, but also the visualisation of the data and mined results, as well as an integrated system framework. Since what people want is generally implicit, and what can be expected from raw data depends on its context, there is no universal solution in this area.

Even though we can surely expect to see and to benefit from the continuing development of methods, it is most important to incorporate human involvement into the knowledge acquisition process, in addition to gaining insights into the domain, data, available techniques, and design of an appropriate process. Successful applications of data mining indicate that the issue of incorporating human involvement and decisions might be even more fundamental, leading to the challenging problems of *human centred knowledge acquisition* and an *inclusion of and interaction with user preferences in data mining.*