Introduction: Lessons Learned from Data Mining Applications and Collaborative Problem Solving

Nada Lavrač (nada.lavrac@ijs.si)
Institute Jožef Stefan, Jamova 39, 1000 Ljubljana, Slovenia
Nova Gorica Polytechnic, Vipavska 13, 5000 Nova Gorica, Slovenia

Hiroshi Motoda (motoda@sanken.osaka-u.ac.jp)
Osaka University, 8-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan

Tom Fawcett (tom.fawcett@hp.com)
Hewlett-Packard Labs, 1501 Page Mill Rd., Palo Alto, CA 94304, USA

Robert Holte (holte@cs.ualberta.ca)
Computing Science Department, University of Alberta, Edmonton, Alberta Canada T6G 2E8

Pat Langley (langley@csli.stanford.edu)
Computational Learning Laboratory, Center for the Study of Language & Information, Stanford University, Stanford, CA 94305, USA

Pieter Adriaans (pietera@science.uva.nl)
Institute for Language Logic and Computation, Plantage Muidergracht 24, 1018 TV, Amsterdam, The Netherlands

April 11, 2004

Abstract. This introductory paper to the special issue on Data Mining Lessons Learned presents lessons from data mining applications, including experience from science, business, and knowledge management in a collaborative data mining setting.

Keywords: data mining, machine learning, scientific discovery, lessons learned, applications, collaborative data mining, knowledge management, future data mining challenges

1. Introduction

This paper reports on experiences gained from a wide variety of applications of machine learning, data mining and scientific discovery. Lessons are drawn from both successes and from failures, from the engineering of representations for practical problems, and from expert evaluations of solutions.

In drawing lessons from two different types of data mining and machine learning applications, fielded commercial applications and applications in scientific discovery, we focus on lessons that are new or have been under-emphasized in earlier articles (Langley & Simon, 1995; Brodley & Smyth, 1995; Fayyad, Piatetsky-Shapiro, & Smyth, 1996;

Saitta & Neri, 1998). However, we also outline the most important lessons reported previously which should not be forgotten, as they still hold true:

- the developer must formulate the problem in a way that is amenable to existing learning algorithms,
- success often hinges on engineering the problem representation;
- it is often crucial to collect or select training data carefully, and
- the preceding steps are more crucial to success than decisions about which learning algorithm to use.

In addition to the lessons learned from the applications of machine learning and scientific discovery, the paper also presents knowledge management lessons learned in a collaborative setting involving distributed data mining teams.

This paper is organized as follows. Section 2 summarizes two invited talks at the workshop on *Data Mining Lessons Learned* (Lavrač et al., 2002) organized at the *Nineteenth International Conference on Machine Learning* (ICML-2002) in Sydney in July of 2002. These invited talks concerned lessons learned from applications of machine learning and the computational discovery of scientific knowledge. Section 3 discusses business lessons learned from large industrial projects. Section 4 discusses lessons learned from collaborative research and development projects, in terms of the data mining process, the organization of teamwork and knowledge management. The paper concludes by outlining some trends and directions for further work.

2. Lessons from applications of machine learning and scientific discovery

Many lessons have been extracted from commercial applications of machine learning and data mining (Langley & Simon, 1995; Brodley & Smyth, 1995; Fayyad, Piatetsky-Shapiro, & Smyth, 1996; Saitta & Neri, 1998). A novel contribution of this section is to combine some of that experience with work on the machine-aided discovery of scientific knowledge. As Langley (2000) has noted, there have been successful applications of the latter kind in many scientific areas, including astronomy, biology, chemistry, ecology, graph theory, and metallurgy, and the lessons enumerated in the introduction are as valid for such discovery tasks as for commercial applications. Successful developers

of scientific discovery systems spend significant time formulating (and reformulating) their problem, in engineering the problem representation, and in collecting and manipulating the data. Likewise, the lessons that we discuss below apply equally to both commercial and scientific applications.

2.1. Machine Learning Works

The first and most important lesson to be drawn from attempts to apply machine learning to commercial and scientific problems is that they often succeed, producing performance systems or knowledge bases that had no previous counterparts or that improve on existing ones. The examples that follow are representative of a much larger body of successful applications.

Loan application screening. In the 1980s, American Express (UK) used statistical methods to divide loan applications into three categories: those that should definitely be accepted, those that should definitely be rejected, and those which required a human expert to judge (Langley & Simon, 1995; Michie, 1989). The human experts could correctly predict if an applicant would, or would not, default on the loan in only about 50% of the cases. Machine learning produced rules that were much more accurate—correctly predicting default in 70% of the cases—and that were immediately put into use.

Printing press control. During rotogravure printing, grooves sometimes develop on the printing cylinder, ruining the final product. This phenomenon is known as *banding*, and when it happens production must be halted and the cylinder repaired or replaced before printing can be resumed. The causes of banding are imperfectly understood, even by experts. The printing company R.R. Donnelly hired a consultant for advice on how to reduce its banding problems, and at the same time used machine learning to create rules for determining the process parameters (e.g., the viscosity of the ink) to reduce banding (Evans & Fisher, 2002). The learned rules were superior to the consultant's advice in that they were more specific to the plant where the training data was collected and they filled gaps in the consultant's advice and thus were more complete. In fact, one learned rule contradicted the consultant's advice and proved to be correct. The learned rules have been in everyday use in the Donnelly plant in Gallatin, Tennessee, for over a decade and have reduced the number of banding occurrences from 538 (in 1989) to 26 (in 1998).

Telephone technician dispatch. When a customer reports a telephone problem to Bell Atlantic, the company must decide what type of technician to dispatch to resolve the issue. Starting in 1991, this decision was made using a hand-crafted expert system, but in 1999 it was replaced by another set of rules created with machine learning (Provost & Danyluk, 1999; Danyluk, Provost, & Carr, 2002). The learned rules save Bell Atlantic more than ten million dollars per year because they make fewer erroneous decisions. In addition, the original expert system had reached a stage in its evolution where it could not be maintained cost effectively. Because the learned system was built by training it on examples, it is easy to maintain and to adapt to regional differences and changing cost structures.

Laws of metallic behavior. A central process in iron and steel manufacture involves removing impurities from molten slag, but metallurgists' knowledge of the laws that govern it are incomplete. To improve this situation, D. Sleeman and his colleagues developed DAVICCAND, an interactive discovery system that lets its user select pairs of numeric variables to relate and focus attention on some of the data, while it searches for numeric laws that relate variables within a given region. This procedure produced two new discoveries about metallurgy. The first involves the quantitative relation between the amount of oxide in slag and its sulfur capacity under different temperature ranges. The second contribution concerned improved estimates for the oxide content of slags that contain TiO_2 and FeO, along with the conclusion that FeO has quite different basicity values for sulphur and phosphorus slags. These results were deemed important enough to appear in a respected metallurgical journal (Mitchell et al., 1997).

Reaction pathways in chemistry. A recurring problem in chemistry involves determining the sequence of steps, known as the *reaction pathway*, for a given chemical reaction. Because the great number of possible pathways makes it possible that scientists will overlook viable alternatives, R. Valdés-Pérez developed MECHEM, a system that carries out a search through the space of reaction pathways that can account for a set of chemical reactants and products provided by its user. This approach has produced several novel reaction pathways that have appeared in the chemical literature. For example, Valdés-Pérez (1994) reports a new explanation for the catalytic reaction $ethane + H_2 \rightarrow 2\, methane$, which chemists had thought was largely solved, whereas a later application generated new results on acrylic acid. MECHEM produces pathways in a notation familiar to chemists, and users can influence its behavior by invoking alternative domain constraints.

2.2. Applications generate new scientific directions

The second lesson is that applications often generate challenging new questions and directions for scientific enquiry within machine learning. Typically, these issues are not fully investigated in the course of developing the applied system, and much remains for the scientific community to unravel. If these issues are properly communicated in the scientific literature, the scientific community benefits tremendously (Provost, 2003). For example, the following research issues all arose in a single application effort (Kubat, Holte, & Matwin, 1998), and all are still open questions:

- What are the best methods for learning and evaluating performance when classes are highly imbalanced?
- What are the best methods for learning and evaluating performance when data occurs in batches that differ systematically from one another?
- What are appropriate methods for evaluating performance when the training data are hand-picked, rather than sampled randomly from the natural population ?
- What are the best methods for learning and evaluating performance when objects are described and can be classified at multiple levels of granularity?
- What are the most appropriate ways of developing, tuning, and evaluating a system when very little training data is available?

The last point deserves some amplification, because the data mining community takes as its starting point that we are deluged with data, and the fundamental problem is dealing with the overwhelming quantity of observations. While this may be true for some business applications, for the scientific applications we have examined (Langley, 2000), this is the exception rather than the rule.

For example, in one project on ecosystem modeling (Saito et al., 2001), some 14,000 observations could be extracted from satellite images for a few variables, but only 303 data points were available for other variables that were needed to determine the relations of interest. In another effort that involved modeling gene regulation, there were thousands of measurements, since DNA microarrays can estimate expression levels for many genes at the same time, but there were only 20 distinct samples measured over five time steps, which provided very few constraints on candidate models. Thus, the frequently heard

rhetoric about the massive data sets generated by satellite imagery and microarray technology is misleading at best.

2.3. The deployment context is all important

The *deployment context* of a machine learning application refers to the actual, final use to which the learned knowledge/system will be put, and the real-world context and conditions in which that use will take place. This context defines three important aspects of the machine learning task. First, it determines the exact role that the learned system will play; for example, will the system make final decisions? Is it allowed to defer to a human? Is it ranking alternatives rather than selecting among them? Second, it specifies the scope and distribution of the inputs to which the learned system will be applied. Finally, it defines the performance criteria and standards by which the learned system will be judged.

For example, rarely in a commercial application will classification accuracy be the performance criterion. Misclassifications have costs, and performance is judged on the basis of these costs, not in terms of predictive accuracy (Provost, Fawcett & Kohavi, 1998). Moreover, these costs are usually not known in advance and they may change over time. In both commercial and scientific applications, the deployment context creates learning tasks that differ from ones typically studied by the research community in the following key ways.

Interactive (man-machine) systems. Although the literature on machine learning and discovery emphasizes automated systems, applications often require systems that interact with a domain expert. Even when the goal is to develop a commercial system that will replace a human, an expert can usually provide crucial guidance in defining the search space and evaluating results. In scientific applications, domain experts have no desire to be replaced, and instead are eager for computational tools that can make their own data analyses more productive. Both contexts point to the need for interactive environments that assist humans in understanding data while letting them remain in control.[1] This has been the fundamental premise of several learning systems, including Structured Induction (Shapiro, 1987) and Protos (Porter, Bareiss, & Holte, 1990), and it has been a key element in both commercial (Evans & Fisher, 2002) and scientific (Mitchell et al., 1997) applications.

[1] The data mining community has also developed such interactive environments, but they are designed for use by professional data miners, not those who use the knowledge.

Comprehensible models. An interactive approach makes it imperative that the learned model be comprehensible to the human expert. Even for non-interactive systems, the resulting model is seldom accepted and deployed unless it can be understood by people with decision-making power. In many scientific domains, there are well established notations for expressing knowledge. For example, structural models are used in organic chemistry, systems of numeric equations are common in Earth science, and qualitative causal models are typical in microbiology. A minimum requirement for comprehensibility is that the learned model be expressed in a notation that is familiar to the human expert. Unfortunately, few scientific notations bear much relation to the formalisms popular in machine learning and data mining. Other factors can also accept comprehension and acceptability. For instance, Pazzani et al. (2001) found that medical doctors were more likely to accept induced rules that did not violate preconceived notions (e.g., that Alzheimer's is more likely in older patients).

Explanatory models. In many cases, learned models must be more than comprehensible, in the sense of being readable; they must also "make sense" to human experts. Some fields of science desire more than descriptive laws, and instead aim for explanatory accounts of observations in terms of interactions among hypothesized components and processes. The MECHEM system (Valdés-Pérez, 1994) has used this approach to discover new reaction pathways in physical chemistry. But even the induction of predictive models can benefit by exploiting domain knowledge to ensure that the models it produces make sense and that the conclusions it draws can be explained satisfactorily. The need for explanatory models was the key driving force behind PROTOS (Porter, Bareiss, & Holte, 1990) and the system of Clark and Matwin (1993).

3. Business lessons learned from large industrial projects

Unlike most scientific investigation, companies operate in highly competitive environments. Part of the competitive advantage of an organization is embodied in the knowledge it has of its environment (Halliman, 2001; Berry & Linoff, 1997). Much of this knowledge can be modeled, allowing the organization to select those actions that maximize its profit, or minimize its loss.

Data mining in industry can be defined as the effort to generate actionable models through automated analysis of their databases. In order to be useful for industry, data mining must have a financial

justification. It must contribute to the central goals of the company by, for example, reducing costs, increasing profits, improving customer satisfaction or improving the quality of service. Formulated in this way it is clear that data mining in industry is quite an ambitious effort (Adriaans, 2002a) as it is mainly the improvement in the return of investment that counts.

From this perspective it is clear that data mining cannot be successfully undertaken in all companies. Many companies simply lack sufficient data, and their business depends almost completely on informal human knowledge they have about their clients, products and services. Companies that can benefit from data mining typically have a considerable size and maintain extensive information systems with large databases. Most promising for data mining are organizations that automatically gather large quantities of data without human intervention, such as web providers, call-detail records in telecommunications, companies that process financial or credit transactions, and systems that collect data from large automated systems. Companies of this type operate in very dynamic environments and use data mining as a continuously evolving process to update their models. The infrastructure needed by such companies is completely different from the infrastructure one needs for isolated short-term projects to solve an individual problem.

3.1. Criteria for success

Most business people (marketing managers, sales representatives, quality assurance managers, security officers, and so forth) who work in industry are only interested in data mining insofar as it helps them do their job better. They are uninterested in technical details and do not want to be concerned with integration issues. A successful data mining application has to be integrated seamlessly into a marketing application, a CRM tool, a service management environment, an inventory system or a prognostic and health management tool. Bringing an algorithm that is successful in the laboratory, even on real life data, to an effective data mining application in industry can be a very long process. Issues like cost effectiveness, manageability, maintainability, software integration, ergonomics and business process re-engineering come into play.

Introducing a data mining application into an organization is not essentially very different from any other software application project;

the same basic principles hold.[2] For every software project one has to form a project team that acts as an agent of change in an organization. If t is the size of the team and d is the duration then $(t \cdot d)$ is a good indication of the costs involved. Now one can evaluate the risks of a software application project along the following four dimensions:

- Duration d of the project,
- Size s of the project in person months,
- Internal Complexity C_i of the project, and
- External Complexity C_e of the project.

C_i is a measure for the amount of learning and mutual communication that is involved inside the team. C_e is a comparable measure for the impact that the project has on the organization as a whole. How many employees are affected and how fundamentally is their work environment changing? Especially the measure C_i can be estimated quite accurately. Let l be the average time needed to learn the necessary skills for the project for each team member. Let c be the fraction of time each team member needs to communicate with one of their direct colleagues and let m be the fraction of the total number of team members a team member needs to communicate with. Then we have:

$$C_i = (1 - \frac{l}{d})^{-1} \cdot (1 - c \cdot m \cdot (\frac{s}{d} - 1))^{-1}$$

Here the first term $(1 - \frac{l}{d})^{-1}$ is an indication of the loss of efficiency as a result of time spent learning on the job. The second term $(1 - c \cdot m \cdot (\frac{s}{d} - 1))^{-1}$ is an indication of the loss of efficiency as a result of internal communication in the team. This analysis makes clear that there are some conditions under which any team will collapse: as soon as l approaches d or when $c \cdot m \cdot (\frac{s}{d} - 1)$ approaches 1. In the first case all the time is spent doing nothing but learning; in the second case all time is consumed in communications. Seasoned project leaders will recognize the situations. People in top management invariably tend to make the following mistake. They formulate a strict deadline, so d is fixed and then take $t = \frac{s}{d}$ to be the size of the team necessary to finish the project where in fact it is $t = \frac{s}{d} \cdot C_i \cdot C_e$.

The importance of this analysis for data mining lies in the fact that for data mining projects the values of C_i and C_e are somewhat

[2] This specific form of metrics was used to manage about 200 projects in Syllogic between the years 1990 and 2000. It is not fundamentally different from other techniques (Garmus & Herron, 2001).

higher than for normal projects whereas it is not uncommon for top management to expect a return on investment within 6 months. Data mining projects typically have a higher complexity than most other software projects, so l is usually inherently high. Data mining projects are knowledge intensive. Remember that data mining can only be deployed successfully when it generates insights that are substantially deeper than what the company already knows about its business. Designers and analysts have to learn a good deal about the situation of the client before they can assess the viability of a data mining approach. This implies that successful data mining projects in industry preferably should be done by consultants who already have a deep understanding of the client's business. The same holds for the employees of the company. They should have a feel for the possibilities of data mining[3]. For the same reasons internal communication in the data mining team has to be quite intensive. This creates higher values of c and m. Finally, the deployment of data mining results in an organization can have considerable consequences for the structure of business processes. This creates a higher value for C_e.

From this analysis we can draw several pieces of advice:

- Data mining projects should be carried out by small teams with a strong internal integration and a loose management style.

- Pilot projects with a steep learning curve are of vital importance.

- A clear problem owner should be identified who is responsible for the project. Preferably this is not a technical analyst or a consultant but someone with direct business responsibility, e.g., someone in a sales or marketing environment. This will benefit the external integration.

- The positive return on investment should be realized within 6 to 12 months.

- Since the roll-out of the results a data mining application mostly involves larger groups of people and is technically less complex, it should be a separate and more strictly managed project.

- The whole project should have the support of the top management of the company.

[3] An interesting observation from tool vendors is that clients often discover new unexpected application areas of their software of which the vendor is unaware. The learning curve of experienced users is very steep.

Before starting a long term data mining project it is wise to carry out one or more small pilot projects with a relatively small data set and a small team consisting of a data mining expert and a domain expert with some technical support. In order to use data mining techniques in an industrial setting the following conditions (Chapman et al., 2000; Adriaans & Zantinge, 1996) have to be satisfied.

1. The data must be available. This point may seem trivial, but data are not always available and ready for data mining. Data sets may be scattered over an organization, stored in legacy database systems, probably in different formats. Sometimes manual codification and editing is necessary. In some cases, legal constraints may prevent sensitive data from being shared with a third party. In some cases, the decision to start a data mining project may even be the first time the company has committed to gathering historical data in a systematic way. Such impediments may cause a substantial delay before tangible results are obtained.

2. The data must be relevant, adequate and clean. By *relevant* we mean that the data must be sufficient to support analysis and for drawing conclusions for the domain of interest. By *adequate* we mean that all the attributes that we need are available and filled to an acceptable degree. By *clean* we mean free of noise and errors. Most infrastructure databases in a company were not designed with data mining applications in mind, so only data that have direct relevance for the day-to-day operation are stored. To test these criteria, it is very useful to engage in a small preliminary pilot project to analyze a data sample. Companies often over-estimate the quality of their data. A pilot study should make clear what data quality issues exist, from which an estimate can be made of how much time and effort must be spent in data pre-processing and cleaning. It is very useful to warn clients about this.

3. There must be a well-defined problem. Data mining should be a goal-directed activity. Data mining companies often get datasets from clients with the directive that they simply "find any interesting patterns" in them. But there are always patterns in data; without knowing what to look for the data miner cannot judge their value. The client must provide a well-defined goal, without which there is no measure for success and no way to assess the value of data mining results. In some cases, a data miner may be faced with an open-ended assignment; for example, to do a feasibility study for a bank to investigate the possible uses of data mining in the organization. This is a valid undertaking,

but the best way to approach it is to discuss possible applications (for example, detecting fraud, improving quality of customer service, or reducing mail costs) with the management. A series of tests should be done on these possible application areas. The entire problem definition process may take the form of a "negotiation" between data miner and client.

4. The problem should not be solvable by means of ordinary query or OLAP tools. In order to start a data mining project some kind of data storage with query facilities must be in place. This is a necessary condition for the success of any data mining project. If the problem can be solved with this basic infrastructure alone then there is no need to start an advanced data mining effort. Ideally an organization that begins to implement data mining solutions should already have experimented extensively with more traditional query and reporting tools and, on the basis of these experiments, should have concluded that the traditional solutions do not work or are too labor intensive. Many companies start a data mining project on the basis of the belief that their databases form a real asset. Even if this is true, they might be helped substantially with some clever reporting from a query tool rather than an advanced data mining effort. In some cases we have seen data mining projects called off because traditional reporting tools already produced more interesting information than the organization could cope with.

5. The results must be actionable. This point recurs throughout many of the articles in this special issue. Data mining can usually produce new knowledge, but the results of the data mining process should lead to actions that really can be implemented by the organization to further its financial goals. What constitutes "actionable" varies from organization to organization. In a direct marketing company, for example, one could deploy the results of a data mining effort in a number of ways:

- Making the discovered knowledge accessible to other (less experienced) users via dedicated interfaces to existing software applications.
- Optimizing outbound marketing campaigns. With direct mail one can reach a 20–40% cost reduction.
- Deploying results in other channels, e.g., the call center. On-line data mining results can be used for dialog-control. If we combine content data of a dialogue with back-end data from operational systems we get a very powerful marketing system.

Summary. Given these five criteria, it is clear that not every company will satisfy them. Having a lot of data, even if the quality is good, is certainly not a guarantee for success. A number of data mining projects have failed in the past years because one or more of these criteria were not met. On the other hand, if a data mining project satisfies these criteria it is likely that the project can lead to substantial improvements and cost reductions for the company.

3.2. Data mining as a business

On the basis of these issues it is also clear under what conditions data mining as a business can be successful. Data mining companies vary considerably, but in general a company either sells consultancy, tools or a combination of the two. If one takes point four of the previous section into account then it is clear that a vendor that only offers horizontal data mining tools will always be in competition with vendors of data base management environments, OLAP, reporting and query tools. They have inherently a bigger potential market then data mining tool vendors. For vendors of horizontal database management systems and query tools it is relatively easy to enhance their product with data mining capabilities or to buy a small innovative data mining company in order to get access to data mining expertise. Furthermore, the data mining vendor is probably the last one to enter the client's site and will find a database environment that is already up and running. This is one of the main reasons that it is almost impossible for a company to survive on the basis of sales of horizontal data mining tools only. There is no room in the market for independent vendors of horizontal data mining tools. In this light there are a number of strategies that a data mining company can follow: find a vertical market and specialize, sell to a strong vendor of horizontal solutions, or simply quit the business.

In selling data mining consultancy, it is not easy to find a market that sustains a healthy business in the long run. The problem is that data mining per se deals with finding deep knowledge that is specific to an organization (points 3 and 5). Also, as Kohavi et al. (2004) observe, every problem is different. The client usually knows his business better than the consultant. Only by building up specific expertise concerning the application of data mining techniques in a vertical segment the data mining business can survive.

3.3. Quality of the data

Traditionally most database systems have been designed to fulfill a specific (mostly administrative) need. All the data are gathered with a specific application in mind. Data mining, in contrast, is generally

an open-ended process. Any subset of the data, any combination of attributes or any derived attribute in the database can be of interest at some point in time. This sets much higher qualitative criteria for the data.

Often the quality of the data in commercial databases is very uneven. Attributes that are vital for the business are of high quality. Other attributes are heavily polluted. A lot of companies have made disturbing discoveries about the quality of their data after starting data mining projects. There is a hidden problem of legacy databases in industry. Often it is impossible to cleanse the data in a cost-effective way. It is clear that in such cases an organization needs to install business processes that ensure the production of data of high quality. This may even include a redesign of applications and their underlying databases. Ergonomic application design is becoming important in this respect.

Designing databases with potential data mining applications in mind is very important. This is a matter of detailed analysis of the context in which the applications are to be used. Software ergonomics is becoming more and more important in this respect. In a lot of cases poor design of a database leads to polluted useless data. Industries have made this disappointing discovery over and over in the past years.

4. Lessons learned from collaborative research and development projects

The core of data mining is the extraction of useful patterns or models from data (Hand et al., 2001). However, to reach actionable results from data usually requires a long and non-trivial process (Berry & Linoff, 1997) involving aspects of business and technology (Pyle, 1999), as well as human skill; the human factor is one of the most important success factors, including project management and control. A well defined process is of importance to achieving successful data mining results, particularly if the number of participants involved in carrying out the data mining tasks is large, involving teams of individuals with different expertise, skills, habits and cultural backgrounds.

Many authors have suggested broadly defined process models to perform data mining (Fayyad, Piatetsky-Shapiro, & Smyth, 1996; Adriaans & Zantinge, 1996). The emerging standard data mining process model is the CRoss Industry Standard Process for Data Mining (CRISP-DM) (Chapman et al., 2000). CRISP-DM subdivides a data mining project into the six, interrelated phases of: 1) business understanding, 2) data understanding, 3) data preparation, 4) modeling, 5) evaluation,

and 6) deployment. Like the alternative data mining processes, there are numerous feedback loops connecting the phases in CRISP-DM.

As data mining is multidisciplinary it often requires the expertise of numerous individuals. The business understanding phase requires communication skills to work closely with the data mining client (the organization interested in the data mining results). The modeling phase—which requires the use of statistics or machine learning—can be undertaken largely independently of others, making it possible to perform parts of a data mining process in a remote e-collaboration setting. To ensure that collaborative data mining is successful, well defined collaboration principles and support tools are required (Jorge et al., 2003; Mladenić et al., 2003; Mladenić & Lavrač, 2003).

A data mining project that is collaborative involves more complexity than one that is small and local, but there are benefits to combining expertise. To realize such benefits it is vital that all collaborating parties share their results, either complete or intermediate. For example, in the data preparation phase, any data transformations should be made available; while in the modeling phase, the models should be made available in a standardized format. Information needs to be securely but easily shared using an appropriate e-collaboration knowledge management system. The evaluation phase is important in the data mining process for it is in this phase that the key results—the models—are evaluated against the initial project objectives. When working in a collaborative setting it is important that all models be evaluated fairly and consistently. This is best done by centralizing the model evaluation as much as practical.

4.1. Lessons from managing large data mining projects involving business and academia

In Europe, US and Japan, the collaboration between academic, business and industrial teams has been supported by collaborative Research and Development projects performed at the national or international level. Examples of such collaborative data mining projects are the European project *Data Mining and Decision Support: A European Virtual Enterprise* (SolEuNet) (Mladenić et al., 2003; Mladenić & Lavrač, 2003), the European project *Enabling End-User Datawarehouse Mining* (MiningMart) (Morik & Scholz, 2003), the US project *Evidence Extraction and Link Discovery* (Senator, 2002), and the Japanese project *Active Mining* (Motoda, 2002; Motoda & Washio, 2002).

All these practice-oriented research efforts in data mining have recognized the need for methods and tools that include a larger part of the problem solving process than data mining. Based on the CRISP-DM

methodology that covers the process from problem definition to delivery of the resulting patterns, the MiningMart project developed methods and tools that include the preprocessing stage and support the construction and use of a database of solutions. The Active Mining project extended the scope to the active acquisition of data, emphasizing the role of a domain expert in all stages of the data mining process, and demonstrated a spiral modeling of knowledge discovery. The project on Evidence Extraction and Link Discovery addressed similar issues but in the context of specific applications such as military decision making and terrorist network discovery. The SolEuNet project aimed at the integration of data mining and decision support processes in a remote collaborative problem solving setting, resulting in numerous pilot data mining and decision support solutions.

Such projects result in the formation of a dynamic network of expert teams with long term experience in data mining, frequently involving partners from academia, business and industry. For the management of geographically distributed research teams, the SolEuNet project investigated a new organizational model—a virtual enterprise model (Camarinha-Matos, Afsarmanesh, & Rabelo, 2000)—as a basis for establishing dynamic links between experienced data mining experts and customers in need of solutions. This model proposes a flexible association of academic institutions and business entities who, although they may having different motivations for this partnership, share the objective of promoting and selling advanced services offered by the pool of partners. The definition upon which a virtual enterprise is modeled is "a temporary aggregation of core competencies and associated resources collaborating to address a specific situation, presumed to be a business opportunity" (Goranson, 1999).

National and international research and development projects can thus be viewed as virtual enterprises emerging from a single business opportunity (a call for project proposals). Successful partner collaboration in such projects shows that strong motivation and well-defined common goals allow individuals to successfully collaborate across organizational boundaries (Moyle, McKenzie & Jorge, 2002). The problem encountered by partners of such projects and networks is that of identity and long term viability (Lavrač & Urbančič, 2003). During the project financing period many joint results are achieved, much information is gathered and disseminated, and many working relationships and workflows are established. However, briefly after the end of the financing period, the gathered information risks becoming inaccurate and workflows risk being dissolved because the established working relationships are not viewed as intellectual capital that should be further cultivated and exploited (Edvinsson & Malone, 1997).

Project partner relationships are usually regulated by a contract. Since some important relationships between partners cannot be formalized contractually, social and communication aspects of the collaboration should be managed carefully. A lesson from the SolEuNet project is that establishing trust is an important goal in cooperative projects, since the possibility of opportunistic behavior of partners cannot be eliminated by formal contracts. Means for trust building include regular communication, sharing of information and knowledge, and stable rules of the game. The principal-agent theory (Furubotn & Richter, 1997) provides many answers to problems arising from informational asymmetries between partners in a business relationship. In the SolEuNet project, this theory provided a better understanding of the gap between business and academic project partners, which was mainly due to different backgrounds and motivations for entering the project partnership.

The partnership model developed in the SolEuNet project aimed to alleviate this problem of partnership discontinuation after the end of the project funding period. Helped by the analysis of the principal-agent theory, different models of academia-business collaboration were proposed in SolEuNet. A lesson learned is that the virtual enterprise model (Camarinha-Matos, Afsarmanesh, & Rabelo, 2000) with multiple marketing agents (Lavrač & Urbančič, 2003) turned out to be the most appropriate model for handling the difficult business relationship between remote e-collaborating data mining partners. In addition to the virtual enterprise organizational model developed, the collaborating data mining teams learned a number of lessons concerning the data mining process, organization of team work and knowledge management, outlined in the rest of this section.

4.2. Incorporating business and collaborative problem solving aspects into CRISP-DM

The collaboration of remote teams is feasible due to recent technological developments. However, as learned in SolEuNet, e-collaboration is non-trivial: while the CRISP-DM data mining phases are well understood, data mining team-work in an e-collaborative setting has only recently been investigated (Jorge et al., 2003; Mladenić et al., 2003). Remote collaborative data mining enables the exploitation of available complementary skills at different geographic locations, and benefits organizational memory. Some aspects of the remote e-collaborative data mining setting are outlined below (Jorge et al., 2003).

Prior to the collaborative data mining phase, a local team gathers information about the business and data mining problems from the client

by following the business understanding and data understanding phases of CRISP-DM. Negotiation terms, privacy and intellectual property issues need to be addressed at this point. The outcomes of this phase include the data mining specification and an initial database. Further steps include appointing a project coordinator, setting up the data base and the knowledge repository, and agreeing upon the communication media for remote collaborative work.

While the database usually resides within a centralized server, the knowledge generated by the data mining experts can be stored in other systems and accessed remotely. The virtual expert team is now formed, consisting of data miners and possibly including domain experts. Each member must be aware of the legal implications of joining the project and must have access to the database, the knowledge repository, and other project resources.

Although most of the work can be done remotely, an initial face-to-face meeting of a significant part of the team with a representative of the client should occur. This meeting should preferably result in a pilot data mining project conducted on samples of data. The goal of this pilot study should be a good understanding of the data mining problem.

Having defined the tasks and the deadlines, remote collaborative data mining proceeds by following the CRISP-DM data preparation and modeling phases. New questions about business and data understanding can be posed directly to the client or through the local team. Team members have online discussions, publish all produced knowledge in the repository, and can use the knowledge produced. Frequent summaries of the developments, both scheduled and spontaneous, are important. At each milestone, intermediate results are delivered to the client as described in the project plan (a CRISP-DM output of the business understanding phase) and agreed in the specification. For each of the intermediate results, tasks in the CRISP-DM evaluation and deployment phases can start, which is accomplished by the local team and the project coordinator. At the end of the project, before the virtual team dissolves, it is useful to gather the lessons learned from this project and to produce an experience documentation report. Afterwards, the knowledge repository becomes read-only but remains accessible to team members.

A lesson learned from reporting in the process of developing a solution is that people usually do not want to discuss their failures or dead ends, even though such information would be useful to others. They prefer to first solve the problem themselves and then report on the success. Collaborating groups should thus be encouraged to share all investigations, whether fruitful or not. Another benefit is that a

great deal can be learned by observing other group's working styles; for example, seeing diverse approaches to problem solving, cultural differences, work habits, etc. Such knowledge can be very useful in future collaborations with experts from different institutions.

One final lesson concerns the cost of collaborative data mining. In our experience, only some parts of the data mining process can be done remotely, and the initial business and data understanding phases are best done through personal communication between the client and the data mining experts. One needs to be aware of the costs and the resulting benefits of a collaborative problem solving approach, considering the criteria for success discussed in Section 3.1. For simple applications, collaboration is usually not cost-effective; however, for complex and difficult applications, such as the ones in e-science where costs are less important than the quality of solutions, a collaborative approach may produce superior results.

4.3. Model formalization and visualization using the Predictive Model Markup Language standard

As pointed out in the lessons reported by (Wettschereck, Jorge, & Moyle, 2003), models induced by a team of collaborating data miners should be made available in a standardized format. The emerging standard for the platform- and system-independent representation of data mining models is the Predictive Markup Model Language (PMML). PMML is the result of an ongoing standardization efforts of the Data Mining Group[4], an independent, vendor-led group which develops data mining standards. PMML is intended for the representation of the results of knowledge discovery tasks. The primary purpose of the PMML standard is to separate model generation from model storage in order to enable users to view, post-process, and utilize data mining models independently of the data mining tools that generated the model.

Different data mining methods typically produce syntactically and semantically different models. The current PMML standard (version 2.0) is currently supported by a number of data mining tools. It supports separate document type definitions (DTDs) for decision trees, neural networks, center and density based clusters, general and polynomial regression, Naive Bayes and association and sequence rules. The representation language is XML. All DTDs have certain common elements (such as a header with common information and a data dictionary), but the XML elements describing actual data mining models can differ significantly.

[4] `http://www.dmg.org`

Experience from the SolEuNet project taught us that use of the PMML standard has significant advantages for collaborative data mining. Because it is a declarative representation, it separates analysis results from the details of the systems used to generate them (implementation, platform, operating system, and so forth). It allows convenient exchange of data mining results among systems and researchers. For example, consultants or researchers can produce models, and customers can import models into their own tools. Finally, the use of PMML for standardized model description enables the tools for the visualization of PMML models to be developed independently of particular data mining software employed, providing a great advantage for team-work in a remote collaborative data mining problem setting (Wettschereck, Jorge, & Moyle, 2003).

4.4. Knowledge management for data miners

Every organization possesses a considerable body of knowledge that is important to retain and reuse without relying on the presence of its members (Armistead & Meakins, 2002). Organizational knowledge tends to be tacit, and distributed, so only a small part of it is likely to be acquired and retained.

Since knowledge can only be explicitly kept as information, it is necessary to design effective ways of representing the knowledge as information, collecting it from members of the organization, storing it in an understandable, computationally accessible and flexible way, and finally disseminating it to different audiences. This effort can be referred to as knowledge management (Turban & Aronson, 1998), and is related with the setup of an organizational memory (Dieng, 2000). Collections of definitions of what is and what was knowledge management can be found in the literature (Malhotra, 2001).

Jorge et al. (2003a) point out that the knowledge gathered by a team of data miners throughout its activity is too valuable an asset to be kept volatile, always dependent on those who produced it. In a data mining team, knowledge management needs to focus on particular tangible aspects such as expertise, resources, finished projects, solved problems and products, and on making the relations between them as explicit as possible. To identify the relevant types of knowledge, it is useful to restate the aims of the knowledge management enterprise being described in more detail:

– Linguistic standardization. Key terms and concepts should be fully understood by all data miners. To ensure this, an online glossary is a useful asset.

- Document sharing. Reports may be produced by several authors, in several iterations. A lesson learned is that collaborating teams need a central repository for all communal documents, maintained within a workgroup support system.

- Awareness of common resources. The collaborating teams should collect information about tools authored by their members, about contact persons, and other important resources; a lesson learned is that this can be done through a well-designed Web-based information collection system (Jorge et al., 2003a).

- Collaborative problem solving. A team addressing a particular problem needs to share knowledge. To achieve this goal, a methodology and tools for handling problems remotely and collaboratively need to be proposed (Jorge et al., 2003a; Mladenić et al., 2003; Mladenić & Lavrač, 2003).

- Reuse. Summary information about completed projects and solved problems is frequently provided in different formats, using different terms. Information collection templates that unify seemingly different types of knowledge have to be defined. A set of common descriptors, useful for data mining projects, is proposed in (Jorge et al., 2003a).

- Dissemination. A common Web site, serving as the image to the outside world, can also be used as a portal for the integration of services and for the internal presentation of organizational knowledge.

5. Conclusions and directions for further work

Reported lessons point to some obvious conclusions about directions for additional work in machine learning and scientific discovery. First, researchers should explore methods that generate knowledge in established domain formalisms rather than focusing entirely on those invented by the machine learning community. They should also employ standards (e.g., PMML) for model sharing, use and visualization. We also need increased concern with methods that produce good models from small data sets, whether through incorporation of domain knowledge or statistical techniques for variance reduction, and with methods that generate explanatory models to complement the existing emphasis on purely predictive ones. Finally, the field should expand its

efforts on interactive environments for learning and discovery, rather than continuing its emphasis on automated methods.

These recommendations do not contradict earlier lessons drawn from successful applications. Developers should still think carefully about how to formulate their problems, engineer the representations, manipulate their data and algorithms, and interpret their results. But they do suggest that, despite some impressive successes, we still require research that will produce a broader base of computational methods for discovery and learning. These will be crucial for the next generation of applications in machine learning and scientific discovery.

In addition to the above directions drawn from applications of machine learning and scientific discovery, there are a number of developments in the data mining research that illustrate in which direction the applications are moving. We mention a few challenges observed already (Adriaans, 2000; Lavrač, 2001; Mitchell, 1997) and add some of the recent challenges.

Analysis of data mining lessons learned. Technical literature reports only on successful machine learning techniques and data mining applications. To increase our understanding of machine learning techniques and their limitations, it is crucial to analyze successful and unsuccessful applications. Published accounts rarely discuss the steps leading to success, failed attempts, or critical representation choices made; and rarely do scientific papers include expert evaluations of achieved results. Challenge competitions, such as the KDD Cup, COIL and PTE challenges, provide one of the means for deeper analysis of successes and failures. Specialized workshops and journal special issues may further pave a way towards publishing successful paths to solutions as well as failures or dead ends, which also provide valuable input for machine learning and data mining research.

Analysis of comprehensibility. It is often claimed that for many applications comprehensibility is the main factor if the results of learning are to be accepted by the experts. Despite these claims and some initial investigations of intelligibility criteria for symbolic machine learning (such as the standard Occam's razor and minimal description length criteria) there are few research results concerning the intelligibility evaluation by humans. Pazzani (2000) makes this point as well, and he points out that much of what researchers commonly assume about comprehensibility is unfounded or contradictory.

From batch to on-line. Traditional data mining solutions are batch oriented. Large collections of data are stored in a data warehouse and

once a week or once per month a set of data mining algorithms is processed to see if any interesting patterns emerge. With applications such as fraud detection and production control, this has serious drawbacks. If there is a flaw in the production process a company should take immediate action. The same holds for fraud detection. There is a tendency to apply data mining techniques directly to production databases. In this case one does not have the benefits of an optimized data warehouse architecture, and this calls for new solutions.

From single table to multi-relational. Data mining algorithms such as decision trees and association rules presuppose that data are propositional and stored in a single table. Most data mining applications operate on a single flattened table in which the semantic structure given by the data schema of the original application is lost. Because of this transformation, the mining process is less effective: it might miss patterns that could easily be detected if the original structure was still available. At the same time, it might find patterns that are trivial artifacts of the flattening process. Currently research in inductive logic programming and relational data mining is focusing on creating variants of data mining algorithms that can operate directly on the original relational data (Džeroski & Lavrač, 2001).

Text, Web mining, automated ontology construction and the semantic Web. Text and Web mining deal with unstructured data: documents and information available on the Web. The main applications in both areas are aimed at creating a better understanding of the content of documents and a better understanding of Web users dealing with documents or services. Current applications of text and web mining include document search based on the content, automatic document summarization, document authorship detection, identification of plagiarism of documents, web-log analysis, user profiling, etc. The most popular text mining application is document categorization which aims at classifying documents into pre-defined taxonomies/categories based on their content. A challenging new ares are automatic construction of document hierarchies, and automated ontology construction in the context of the emerging semantic Web (Maedche, 2002).

From general data mining tools to specialized learners and data libraries for e-science. Particular problem areas have particular characteristics and requirements, and not all learning algorithms are capable of dealing with these. This is a reason for starting to build specialized learners for different types of applications. In addition, libraries of "cleaned" data, background knowledge and previously learned knowledge should

be stored for further learning in selected problem areas. Notice that such libraries are currently being established for selected problem areas such as molecular biology. This approach will lead to the reusability of components and to extended example sets, achieved also through systematic query answering and experimentation (active mining).

Active mining. Most current data mining algorithms do a great job in classification. They detect that a certain production process is in an error state, or that certain transactions might be fraudulent. In such cases a company would like to take direct automated action. In system management, for example, one might want to increase the paging space of a certain server or redirect queries to a different data base server. If a system detects credit card fraud the owner of the card must be warned as soon as possible. This desire leads to the merging of data mining technology with agent technology. Such active mining may require learning from local datasets, referential datasets and case bases collected and maintained by the world's best experts in the area, as well as data that is publicly available through the Web. This type of "continuous global" learning will require human intervention, as well as *learning agents* for permanent learning (through theory revision) from updated world-wide data, and *query agents* that will (through dynamic abductive querying) be able to access additional information from the Web via query answering. Query answering may be invoked either by experts or by automatically extracting answers from Web resources, possibly invoking learning and active experimentation.

From single medium to multimedia. Current data mining algorithms work on data that is stored in tables of a database, textual databases and the information on the Web. One also would like to be able to mine databases containing images, sounds, speech, music and movies. More advanced techniques need to be developed for this. The combination of mining techniques operating on different media will lead to fascinating new applications, e.g., forensic solutions that mine a database with photos, movies, speech fragments and emails of suspects.

Industrial needs. The development of a separate data mining industry induces a shift in focus of the research. The basic data mining technology developed in the past 15 years is sufficiently powerful to support industrial needs. There is no direct need for a vast investment for research into incremental improvement of data mining algorithms. On the other hand, a number of issues that would be useful for industrial applications have been generally neglected by the data mining research community. Among these are the maintenance of data mining models;

feedback on and incremental improvement of models; the ability to deploy models on a regular basis more efficiently with less people; fast construction and verification of models; using models for simulation; and the ability to create more robust models.

Acknowledgements

Nada Lavrač would like to acknowledge the support for this research by the Slovenian Ministry of Education, Science and Sport and the European project Data Mining and Decision Support for Business Competitiveness: A European Virtual Enterprise (IST-1999-11495). Hiroshi Motoda would like to acknowledge the support for this research by the Japanese Ministry of Education, Culture, Sports, Science and Technology. Pat Langley would like to acknowledge the support for this research by NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation. Robert Holte would like to acknowledge the support for this research by the Natural Sciences and Engineering Research Council of Canada and by the Alberta Ingenuity Centre for Machine Learning. Pieter Adriaans would like to thank Marcel Holzheimer from Data Distilleries.

References

Adriaans, P. (2002). Backgrounds and general trends. In J. Meij (ed.) *Dealing with the Dataflood, Mining Data, Text and Multimedia*, 16–25, STT Beweton, The Hague, Netherlands.

Adriaans, P. (2002a). Production control. In Klösgen, W. & Zytkow, J.M. (eds.) *Handbook of Data Mining and Knowledge Discovery*, Oxford University Press.

Adriaans, P. & Zantinge, D. (1996). *Data Mining*. Addison-Wesley.

Armistead C. & Meakins, M. (2002). A framework for practising knowledge management. *Long Range Planning*, 35(1): 49–71.

Berry, M.J.A. & Linoff, G.S. (1997). *Data Mining Techniques: For Marketing, Sales, and Customer Support*. John Wiley and Sons.

Brodley, C.E. & Smyth, P. (1995). The process of applying machine learning algorithms. In *Proceedings of the ICML-95 Workshop on Applying Machine Learning in Practice*.

Camarinha-Matos, L.M., Afsarmanesh, H. & Rabelo, R. (eds.) (2000). *E-Business and Virtual Enterprise: Managing Business-to-Business Cooperation*. Kluwer.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. CRISP-DM consortium.

Clark P. & Matwin, S. (1993). Using qualitative models to guide inductive learning. In *Proceedings of the Tenth International Conference on Machine Learning*, 49–56.

Danyluk, A., Provost, F. & Carr, B. (2002). Telecommunications network diagnosis. In Klösgen, W. & Zytkow, J.M. (eds.) *Handbook of Data Mining and Knowledge Discovery*. Oxford University Press.

Dieng, R. (2000). Guest Editor's Introduction: Knowledge Management and the Internet. *IEEE Intelligent Systems*, 15(3): 14–17.

Džeroski, S. & Lavrač, N. (eds.) (2001). *Relational Data Mining*. Springer.

Edvinsson, L. & Malone, M.S. (1997). *Intellectual Capital*. Harper Business.

Evans, B. & Fisher, D. (2002). Using decision tree induction to minimize process delays in the printing industry. In Klösgen, W. & Zytkow, J.M. (eds.) *Handbook of Data Mining and Knowledge Discovery*. Oxford University Press.

Fayyad, U.M., Piatetsky-Shapiro, G. & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine* 17: 37–54.

Fayyad, U.M., Piatetsky-Shapiro, G. & Smyth, P. (1996a). From data mining to knowledge discovery: An overview. In Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. & Uthurusamy, R. (eds.) *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press.

Furubotn, E.G. & Richter, R. (1997). *Institutions and Economic Theoy: The Contribution of the New Institutional Economics*. The University of Michigan Press.

Garmus, D. & Herron, D. (2001). *Function point analysis: measurement practices for successful software projects*. Addison-Wesley Longman Publishing Co., Inc.

Goranson, H.T. (1999). *The Agile Virtual Enterprise: Cases, Metrics, Tools*. Quorum Books.

Halliman, C. (2001). *Business Intelligence Using Smart Techniques*. Information Uncover.

Hand, D.J., Mannila, H. & Smyth, P. (2001). *Principles of Data Mining*. MIT Press.

Jorge, A., Moyle, S., Blockeel, H. & Voss, A. (2003). Data mining process and collaboration principles. In Mladenić, D., Lavrač, N., Bohanec, M. & Moyle, S. (eds.) *Data Mining and Decision Support: Integration and Collaboration*, 63–78. Kluwer.

Jorge, A. Bojadžiev, D., Mladenić, D., Štěpánková, O., Palouš, J., Alves, M.A., Petrak, J. & Flach, P. (2003a). Internet support to collaboration: A knowledge management and organizational memory view. In Mladenić, D., Lavrač, N., Bohanec, M. & Moyle, S. (eds.) *Data Mining and Decision Support: Integration and Collaboration*, 247–259. Kluwer.

Kohavi, R., Mason, L., Parekh, R. & Zheng, Z. (2004). Lessons and challenges from mining retail e-commerce data. *Machine Learning*, this issue, Kluwer.

Maedche, A. 2002 *Ontology Learning for the Semantic Web.* Kluwer Academic publishers.

Kubat, M. Holte, R.C. & Matwin, S. (1998). Machine learning for the detection of oil spills in satellite radar images. *Machine Learning* 30(2–3): 195–216.

Langley, P. (2000). The computational support of scientific discovery. *International Journal of Human-Computer Studies* 53: 393–410.

Langley, P. & Simon, H.A. (1995). Applications of machine learning and rule induction. *Communications of the ACM* 38(11): 54–64.

Lavrač, N. (2001) Computational logic and machine learning: A roadmap for inductive logic programming. *Computational Logic*, Special issue: Computational logic roadmap, 47–73.

Lavrač, N., Motoda, H. & Fawcett, T. (eds.) (2002). *Proceedings of the First International Workshop on Data Mining Lessons Learned*, DMLL-2002,

held in conjunction with ICML-2002, Sydney, July 2002. Available at: http://www.hpl.hp.com/personal/Tom_Fawcett/DMLL-2002/Proceedings.html

Lavrač, N. & Urbančič, T. (2003). Mind the gap: Academia-business partnership models and e-collaboration lessons learned. In Mladenić, D., Lavrač, N., Bohanec, M. & Moyle, S. (eds.) *Data Mining and Decision Support: Integration and Collaboration*, 261–269. Kluwer.

Malhotra, Y. (2001). *Knowledge Management for the New World of Business* http://www.kmnetwork.com/whatis.htm.

Michie, D. (1989). Problems of computer-aided concept formation. In Quinlan, J.R. (ed.) *Applications of Expert Systems*, volume 2: 310–333. Addison-Wesley.

Mitchell, F., Sleeman, D., Duffy, J.A., Ingram, M.D. & Young, R.W. (1997). Optical basicity of metallurgical slags: A new computer-based system for data visualisation and analysis. *Ironmaking and Steelmaking* 24: 306–320.

Mitchell, T. (1997). Does machine learning really work? *AI Magazine* 18(3): 11–20.

Mladenić, D., Lavrač, N., Bohanec, M. & Moyle, S. (eds.) (2003). *Data Mining and Decision Support: Integration and Collaboration*, Kluwer.

Mladenić, D. & Lavrač, N. (eds.) (2003). *Data Mining and Decision Support: A European Virtual Enterprise*. DZS Publishers.

Morik, K. & Scholz, M. (2003). The MiningMart approach to knowledge discovery in databases. In Zhong, N. & Liu, J. (eds.) *Handbook of Intelligent IT*, IOS Press.

Motoda, H. (ed.) (2002). *Active Mining: New Directions of Data Mining*. IOS Press.

Motoda, H. & Washio, T. (eds.) (2002). *Proceedings of the First International Workshop on Active Mining*, AM2002, held in conjunction with IEEE ICDM-2002, Maebashi, December 2002.

Moyle, S., McKenzie, J. & Jorge, A. (2003). Collaboration in a data mining virtual organization. In Mladenić, D., Lavrač, N., Bohanec, M. & Moyle, S. (eds.) *Data Mining and Decision Support: Integration and Collaboration*, 49–62. Kluwer, 2003.

Pazzani, M. (2000). Knowledge discovery from data? *IEEE Intelligent Systems* March/April 2000. pp. 10–13.

Pazzani, M.J., Mani, S. & Shankle, W.R. (2001). Acceptance of rules generated by machine learning among medical experts. *Methods of Information in Medicine* 40: 380–385.

Porter, B.W., Bareiss, R. & Holte, R.C. (1990). Concept learning and heuristic classification in weak theory domains. *Artificial Intelligence* 45(1-2): 229–263.

Provost, F. (2003). The role of applications in the science of machine learning. Invited talk at the *Twentieth International Conference on Machine Learning*.

Provost, F. & Danyluk, A. (1999). Problem definition, data cleaning and evaluation: A classifier learning case study. *Informatica* 23: 123–136.

Provost, F., Fawcett, T. & Kohavi, R. (1998). The case against accuracy estimation for comparing induction algorithms. In *Proceedings of the Fifteenth International Conference on Machine Learning*. 445–453.

Pyle, D. (1999). *Data Preparation for Data Mining*. Morgan Kaufmann.

Saito, K., Langley, P., Grenager, T., Potter, C., Torregrosa, A. & Klooster, S.A. (2001). Computational revision of quantitative scientific models. In *Proceedings of the Fourth International Conference on Discovery Science*, 336–349.

Saitta, L. & Neri, F. (1998). Learning in the 'real world'. *Machine Learning* 30(2–3): 133–164.

Senator, T. (2002). *Evidence Extraction and Link Discovery Program*. Speech at DARPATech 2002. Transcript available: http://www.darpa.mil/DARPATech2002/presentations/iao_pdf/speeches/SENATOR.pdf

Shapiro, A.D. (1987). *Structured Induction in Expert Systems*. Addison-Wesley.

Turban, E. & Aronson, J. (1998). *Decision Support Systems and Intelligent Systems*, Fifth Edition. Prentice Hall.

Valdés-Pérez, R. (1994). Human/computer interactive elucidation of reaction mechanisms: Application to catalyzed hydrogenolysis of ethane. *Catalysis Letters* 28:79–87.

Wettschereck, D., Jorge, A. & Moyle, S. (2003). Data mining and decision support integration through the Predictive Model Markup Language standard and visualization. In Mladenić, D., Lavrač, N., Bohanec, M. & Moyle, S. (eds.) *Data Mining and Decision Support: Integration and Collaboration*, 119–130. Kluwer.