Cooperation between expert knowledge and data mining discovered knowledge: Lessons learned

Fernando Alonso , Loïc Martínez , Aurora Pérez , Juan P. Valente

Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo, s/n, 28600 Boadilla del Monte, Madrid, Spain

A B S T R A C T

Expert systems are built from knowledge traditionally elicited from the human expert. It is precisely knowledge elicitation from the expert that is the bottleneck in expert system construction. On the other hand, a data mining system, which automatically extracts knowledge, needs expert guidance on the successive decisions to be made in each of the system phases. In this context, expert knowledge and data mining discovered knowledge can cooperate, maximizing their individual capabilities: data mining discovered knowledge can be used as a complementary source of knowledge for the expert system, whereas expert knowledge can be used to guide the data mining process. This article summarizes different examples of systems where there is cooperation between expert knowledge and data mining discovered knowledge and reports our experience of such cooperation gathered from a medical diagnosis project called Intelligent Interpretation of Isokinetics Data, which we developed. From that experience, a series of lessons were learned throughout project development. Some of these lessons are generally applicable and others pertain exclusively to certain project types.

1. Introduction

Expert knowledge and discovered knowledge are two powerful tools that can be combined. Used together they maximize the qualities that they have separately.

An expert system operates on a knowledge base that contains the knowledge elicited from the expert (EK). This knowledge base is represented by some formalism (rules, frames, Bayesian networks, etc.) and is built by the knowledge engineer from elicited expert knowledge and, later, validated by the expert. Evidently, the system is subject to and limited by the amount of knowledge entered, that is, represented in its knowledge base. And, precisely, the bottleneck in expert system construction is knowledge elicitation, a phase conditioned by countless constraints ranging from the number of available experts, or how much expertise the experts have, to the complexity of the actual knowledge elicitation process.

Recently, automatic knowledge acquisition techniques have attracted a lot of interest as they are potentially a big help for remedying this bottleneck. The knowledge discovery in databases (KDD) process, especially data mining techniques, is used to automatically discover knowledge from data. The knowledge discovered by data mining (DMK) is implicit in the data and can take the shape of patterns or models that fit the data, trends in temporal data, associations among different data features, rules, etc.

The key point is that these two approaches, knowledge elicitation from experts and knowledge discovery from data, complement each other (da Silva, Amorim, Campos, & Brasil, 2002; Daniels & van Dissel, 2002; de la Vega et al., 2010; Weiss, Buckley, Kapoor, & Damgaard, 2003). Applied together, they can be used to build better systems: data mining techniques can be used to support the different tasks involved in expert system (ES) or knowledge-based system (KBS) development (Flior et al., 2010; Mejía-Lavalle & Rodríguez-Ortiz, 1998; Phuong, Phong, Santiprabhob, & Baets, 2001; Wang, Liu, & Cheng, 2004), and expert knowledge can be used to facilitate and improve the results of the different stages of the KDD process (Kusiak & Shah, 2006; Zhang & Figueiredo, 2006).

The aim of this article is to describe the key results of this interaction between EK and DMK, while highlighting the lessons learned over the years from our own experience of these issues in the medical field, presenting a long-term project called I4 (Intelligent Interpretation of Isokinetics Data). This project integrates expert systems and data mining techniques to process isokinetics data. We believe that the results of and the lessons learned from this project are potentially useful for developing systems incorporating EK and DMK.

The remainder of the article is organized as follows. Section 2 describes related work analyzing other applications that present some facet of this type of cooperation. In Section 3 we outline our I4 project. Sections 4–6 describe the three I4 project phases:

expert system development, data mining and symbolic data mining. In Section 7 we summarize the lessons learned. And, finally, Section 8 outlines some conclusions.

2. Cooperation between expert knowledge and discovered knowledge in different fields

Cooperation between EK and DMK has emerged in different fields, like medicine, engineering, finance, etc., and has diverse features. We have grouped different examples of this cooperation into two separate sections. In Section 2.1, we present cases where DM techniques have been used to aid the development of an ES or KBS. In Section 2.2, we present cooperation in the opposite direction, that is, expert system techniques used to improve a knowledge discovery process or system.

2.1. Contribution of discovered knowledge to expert knowledge

Cooperation by applying data mining techniques to an expert system built from heuristic knowledge elicited beforehand from the expert has the goal of optimizing and maximizing the performance of the resulting KBS. This contribution of discovered knowledge to expert knowledge has been designed for different purposes, as we describe below.

- When an expert's heuristic knowledge is the basis for developing an ES and there is a consistent database of solved cases, it is usual practice to check ES robustness by applying DM techniques to the generated database. An example of this type of cooperation is to be found in (Cooke et al., 2000). PERFEX is a rule-based expert system for the automatic interpretation of cardiac SPECT (single photon emission computed tomography) data. This system infers the extent and severity of CAD (coronary artery disease) from perfusion distributions, and outputs a patient report summarizing the condition of the three main arteries and other pertinent information. The overall goal is to assist in the diagnosis of CAD. The expert system presents the resulting diagnostic recommendations in both visual and textual forms in an interactive framework, thereby enhancing overall utility. DM techniques were applied to the patient database containing images and text to validate the system and the confidence (certainty factors) in the heuristic rules of the ES.
- Frequently, KBS users have to answer a great many questions formulated by the system to gather information about the problem. In consultation systems like these, the application of DM techniques to the stored results of previous system executions can optimize future question/answer sequences by reducing the number of questions. This is what the research presented in Bethel, Hall, and Goldgof (2006) did. The developed system is a web-based expert system to match eligible breast cancer patients with open clinical trials, or categorize the reasons why an eligible patient was not put onto the trial. Through interviews with clinicians, implications were discovered that reduced the number of questions/answers required to determine eligibility. The idea is that a physician will immediately know the answers to some questions based on answers to others, that is, there is a clear implication that always holds.
- When the knowledge base is built with generic global knowledge and ES behavior on a specific local problem is unknown (i.e., we do not know whether it will correctly and completely account for all the cases that can occur in that local situation), discovered knowledge can be used to verify whether the rules obtained are representative enough for the local data and if the local data have any new correlation that the knowledge base does not contain. This cooperation between discovered knowledge and expert knowledge was applied in Lama et al. (2006) to treat nosocomial infection. The MERCURIO system was designed to support medical practitioners in the complex task of controlling nosocomial infections. Its knowledge base was built from both the NCCLS (the National Committee for Clinical Laboratory Standards) guidelines and the expert's suggestions. These guidelines are quite general, since they were built considering data regarding many laboratories around the world. However, it is not clear that they can completely and correctly interpret the infections developed inside a particular hospital environment. For this reason, it was necessary to verify if the rules obtained from the NCCLS document were representative of the local hospital infections, and if there were other correlations in the local hospital infection data that are either not considered in the NCCLS document or unknown to the expert microbiologists. To address these problems, data mining was applied to local hospital infection data to generate association rules that show the susceptibility or resistance of a bacterium to different antibiotics. The discovered association rules were transformed into alarm rules that were confirmed by experts and then used for data validation in the system.
- Another type of cooperation of discovered knowledge with expert knowledge takes place when the application domain changes too quickly, and it is very hard or even impossible to timely update the KBS developed by the expert. Discovered knowledge is applied for this purpose, as it can be used to update the system. An example of this type of cooperation is described in WuJing (2001) in a project concerned with optimizing a mobile network. The problem with using a mobile network is that it changes very often, and its management software also upgrades very often. So the knowledge in the expert system must be upgraded accordingly. Traditional knowledge extraction methods cannot conform to the fast changing environment, and DM techniques have been applied to update the knowledge, using the Operation and Maintenance Center repository as a source.

2.2. Contribution of expert knowledge to discovered knowledge

Expert knowledge supports the extraction of knowledge applying DM techniques and the later validation of this knowledge in a wide variety of manners, as described in the following.

- One of the problems of automatically extracting knowledge is how to validate and check its fitness. In this respect, Holmes and Cunningham describe a typical case of KBS construction where expert knowledge can debug and validate the set of rules gathered from DM (Holmes & Cunninghan, 1993). The "Explora" data mining tool is used to build and maintain the ES. "Explora" is a statistics-based mining program that applies domain knowledge and statistical measurements of the database to identify patterns of attribute values and value combinations that occur more or less frequently in the database. The discovered knowledge is debugged and validated by the expert and built into the ES.
- DM can be applied in personalized applications, typical in e-commerce (dynamic content presentation, personalized ad targeting, individual customer recommendations, etc.). In these types of applications rule discovery methods are applied individually to the transactional data of every user to capture each user's truly personal behavior. EK can be used to globally validate all this locally mined knowledge and discard irrelevant rules. Adomavicius and Tuzhilin deal with the problem of a human expert validating large numbers of locally mined rules with relatively little input from the expert (Adomavicius & Tuzhilin, 2001). This is done by applying different rule

validation operators to cluster similarity-based rules, and filter template-based rules and interestingness-based rules. These operators are supplemented by visualization operators, statistical analysis operators and browsing operators.

- Since a data mining process involves multiple stages, a knowledge discovery worker generally faces a confusing array of choices when presented with a data set to mine. Decisions include for example choosing between C4.5, naïve Bayes or neural networks; deciding whether to use discretization and, if so, which method; deciding whether to subsample or to prune; and deciding how to take into account the costs of misclassification. Expert knowledge can be used to guide the user in all those decisions. An intelligent discovery assistant that helps a data miner to explore the space of valid DM processes is presented in Bernstein, Provost, and Hill (2005). First, it provides users with systematic enumerations of valid DM processes. This way, users do not miss important, potentially fruitful options. Second, it also ranks these valid processes by different criteria to help users choose between the options. A combination of several subjective heuristic functions is used for ranking. The main criteria are speed and accuracy, but any other factors of interest to the user, like cost, sensitivity, comprehensibility, etc., and combinations thereof, could also apply.
- Most existing data mining algorithms are data driven and do not fully exploit domain knowledge and decision makers' intuitions. Thus data mining is expected to perform better with than without prior knowledge. This is the case of the Knowledge-Based News Miner (KBNMiner) presented in Hong and Han (2002), which focuses on the effect that news information can have on the prediction of interest rates. KBNMiner is designed to use a prior knowledge base, representing expert knowledge, as a foundation on which to probe and collect news from the Web using text mining techniques. This news information together with the data stored in a financial database will be applied to a neural network model for interest rate predictions.
- Traditional data mining is merely a data-driven process and often overlooks valuable information such as existing knowledge, expert experience or context and real constraints. Then the output results cannot be directly applied to support business decision making. Companies need more interpretable and acceptable models that cannot be mined without the use of EK. For this reason, understanding and utilizing domain knowledge is a critical success factor for data mining projects. How to combine domain knowledge and data mining methods in the knowledge discovery process is an important issue addressed by several authors who have proposed several approaches (Kopanas, Avouris, & Daskalaki, 2002; Lima, Mues, & Baesens, 2009; Peng & Kou, 2008). Huang, Zhang, Zhu, and Shi (2009) propose a new methodology called data mining integrated with domain knowledge, aiming to discover more interesting, more actionable knowledge combining DMK and EK.
- The interaction between the data mining process for knowledge discovery and the application of expert knowledge to verify and debug this knowledge has sometimes led to the construction of intelligent data mining systems including data mining knowledge and expert knowledge. This is the type of system described by Hu and Liu (2006), which consists of three layers: data mining system layer (mines the first-level knowledge), expert system layer (controls the operation of the data mining layer and executes deep mining) and expert user layer (adjusts controlling parameters generating a more meaningful rule set). Combining the user query, expert knowledge and the produced rules, the intelligent data mining system generates an ES and runs the ES on an inference engine to produce the response to the user queries.

3. The I4 ES-DM project

We have experienced the cooperation between EK and DMK in a medical diagnosis system called I4 (Intelligent Interpretation of Isokinetics data) developed over a number of years (1996–2011). This system uses underlying knowledge in the isokinetics domain, gathered by combining the expertise of a physician specialized in isokinetics techniques and data mining techniques applied to a set of existing data.

The assessment of muscle function has been a primary goal of medical and sports scientists for decades. Its main objectives are to evaluate the effects of training, diagnose muscular dysfunctions and assess the effectiveness of rehabilitation programs (Gleeson & Mercer, 1996). The isokinetics machine is one of the most significant muscle evaluation devices and consists of a physical support (see Fig. 1, left) on which patients perform exercises using any of their joints (knee, elbow, ankle, etc.) within different ranges of movement and at a constant speed. In the case of I4, we analyze leg exercises to assess the knee joint.

Each patient session (test) is composed of a set of exercises performed at a fixed constant speed depending on the test protocol in use. The protocol is the design of the medical test as regards issues like number and order of exercises, rest times between exercises and other such details. Each exercise is composed of three repetitions, and each repetition has one extension and one flexion.

Each exercise is characterized by the speed and the leg used. The patient will be in a seated position and the movement is made within a 0–90° flexion/extension arc of the leg. The isokinetics system records the strength applied by the patient and the leg angle every 2/100 s. These data (strength over time) are represented as

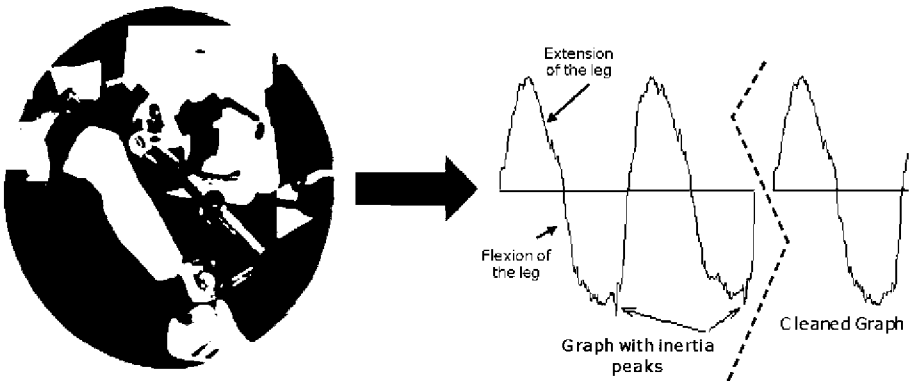

Fig. 1. An isokinetics machine (left), the resulting strength curve (middle) and the curve after inertia peaks cleaning (right).

a sinusoidal curve, containing small peaks and other irregularities (see Fig. 1, middle). The amplitude, total area and irregularities are the main test analysis parameters. The data measured by the isokinetics dynamometer are presented to the examiner by a computer interface. This interface sets out given parameters, which are used to describe the tested muscle function (e.g., maximum strength peak, total effort, etc.).

The software built into these systems does a good job at giving the user access to the data, but there is room for improvement of the software in terms of fully exploiting the massive data flow and also enabling visually impaired physicians to use the system. This is an important point, because one of the initial key goals of the I4 project was to adapt the isokinetics interface for use by visually impaired physicians.

The I4 project was developed in conjunction with the National Centre of Sports Research and Sciences and, in its early phases, the Physiotherapy School of the Spanish National Organization for the Blind, with the aim of building an application capable of analyzing the data output by the isokinetics machine more comprehensively and providing users with more decision support.

The I4 project development has passed through three phases (Fig. 2). The first phase consisted of the development of an expert system for analyzing the isokinetics data and improving the user interface of the system. The second phase consisted of developing a KDD subsystem that used the numerical data generated by the isokinetics machine to discover injury patterns and to create reference models. Finally, the third phase consisted of working with symbols instead of numbers to try to better explain the system output to the user, and the result is a symbols extraction method and a weighted symbolic distance that can be used to perform symbolic isokinetics analysis. These phases are described below, highlighting the lessons learned throughout the process.

4. I4 Phase 1: Expert system development

In the design of the new application, the aim was to build a whole series of new functionalities not developed in the original isokinetics machine software, with the aim of providing better decision-making support. These functionalities targeted preliminary data analysis (data validity, curve morphology, simple comparative analysis) and also included a significant improvement of the isokinetics machine user interface.

To create this decision support system, we opted to design an expert system (Caraça-Valente, Lopez-Chavarrías, & Montes, 2000). To do this, we had access to an expert in the domain with several years of experience running isokinetics tests on top-competition sportspeople and other patients.

The ES works as follows (Fig. 2, left). The data gathered by the isokinetics machine are transformed and formatted before being stored in a test database. Then a cleaning and expert pre-processing component is run to prepare the data for the expert analysis performed by a KBS that contains expert functions and rules. Finally, a visualization module displays the exercises and the expert analysis results for the user.

4.1. Functions

The cleaning and data pre-processing stage had to deal with routine and simple tasks such as removing incomplete exercises, etc. At the same time, though, it had to tackle more complex tasks that required the use of expert knowledge. A series of functions including expert knowledge were created to perform these tasks:

- Firstly, the strength curves are preprocessed in order to eliminate inertia peaks, that is, peaks produced by machine inertia rather than by the patient's actual strength (see Fig. 1, right).
- Then, I4 detects exercise extensions and flexions that are invalid because the patient employed much less effort than in others, or movements that can be considered atypical as their morphology is unlike the others.

The analysis of the strength curves run using the expert functions involves assessing different characteristics of the curve morphology. These characteristics are what the specialist is interested

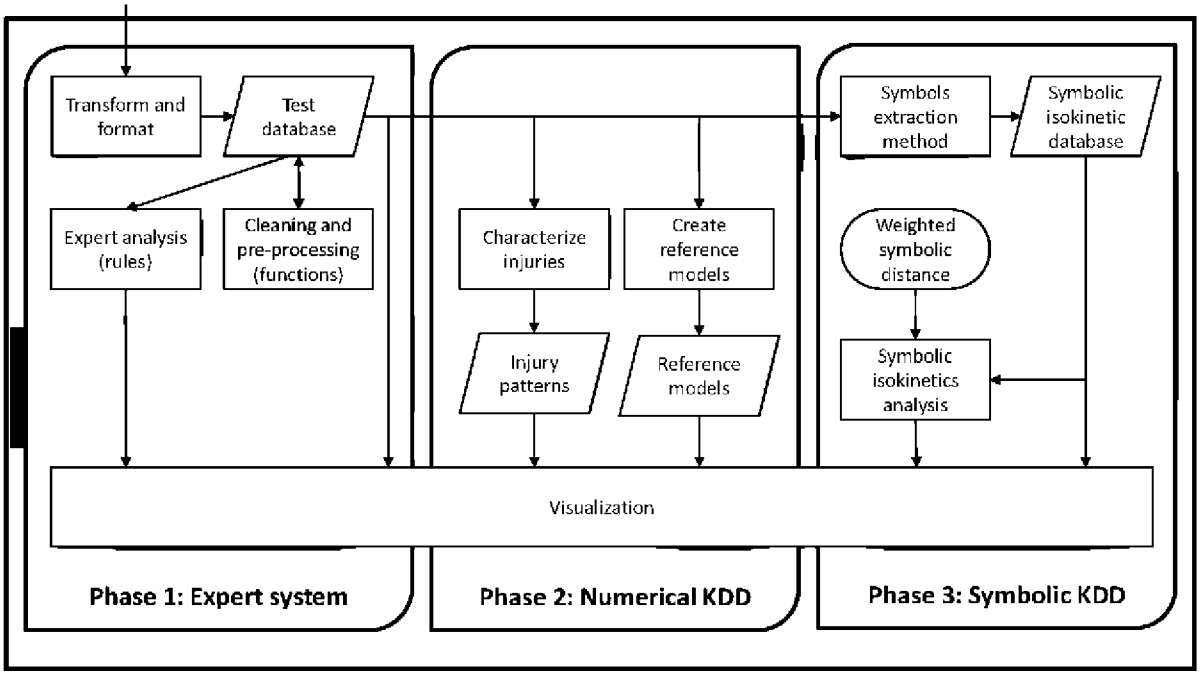

Fig. 2. The I4 main components as developed during three phases.

in, and they constitute the basis for patient assessment. The evaluated aspects are:

- Uniformity: how similar the exercise repetitions are.
- Regularity: whether the curve has a smooth contour or a lot of peaks.
- Acceleration: a qualitative assessment of the time the patient takes to reach the maximum strength value.
- Troughs: prolonged drops and rises in the exercise strength value.
- Shape of the curve: overall assessment of the shape of the curve based on the effort applied at the central angles (around 45°), the flattening of the curve and the angle at which each maximum strength value is reached.

The design and implementation of the expert functions can be described as an iterative and interactive induction process. Given a number of strength curves, the expert evaluated each one and assessed its characteristics. Then, tentative functions were implemented. These functions were applied to a new set of curves, and the results were shown to the expert for evaluation. This evaluation led to some changes in function implementation, and so on. This process ended when the functions provided the correct values in a high percentage of the cases (over 98%). As we will see later, these functions turned out to be useful for pre-processing purposes in phase 2 of the system.

4.2. Rules

Functions represent *procedural knowledge* (especially calculations) very well, but they are not good at representing declarative knowledge, such as heuristic assertions like "If there are many invalid exercises, repeat the tests". Therefore, we decided to add rules to include declarative knowledge to improve the system's decision support functionalities.

The rule-based part of the KBS is responsible for three aspects of isokinetics analysis:

- Protocol validation: This process determines if the protocol has been correctly applied. This is very important since the expertise used for the later parts of the analysis is very sensitive to the way in which the tests are performed. All the exercises must have been completed successfully; the patient must tire to some extent, etc.
- Expert analysis of numerical data: Every numerical feature of the curve (maximum strength value, total effort, gradients of the curve, etc.) is expertly analyzed and findings supplied to the user. There is an individual analysis of each leg and a comparison between both legs.
- Morphological analysis of data, based on the output of the expert functions described above. While the functions can determine the morphology of the strength curves, the rule-based subsystem uses the results output by the functions to analyze the strength curves, compare each leg's morphology and extract conclusions such as, for instance, the presence of some kind of dysfunction.

4.3. I4 Phase 1: Results

In this phase, we developed an intelligent system, built by means of an incremental KBS development process. This system has a modular architecture that consists of: (a) data decoding, testing and storage; (b) intelligent data analysis; (c) data and results display and report generation. To carry out the intelligent analysis of data, we built an expert system that represents procedural expert knowledge by means of functions and declarative expert knowledge by means of rules.

The intelligent system described was designed using an object-oriented ES development methodology (Alonso, Fuertes, Martínez, & Montes, 2000). This methodology is especially suited for this project, as it encourages interaction between software developers and medical/sports assessors. Exercises and tests are the most significant classes of the object model in relation to data interpretation.

This system is particularly useful for specialists, as it makes their job easier. In addition, as there are very few specialists in isokinetics assessment of muscle strength data, this system is extremely valuable as an instrument for disseminating isokinetics technology and encouraging non-expert medical practitioners to enter this field.

4.4. I4 Phase 1: Lessons learned

From the development of the phase 1 of the project, we learned the following lessons, which can be generalized to other expert systems.

4.4.1. Advantage of using more than one formalism to represent knowledge

The I4 expert system uses two types of expert knowledge representations: functions and rules. Each representation accounts more intuitively for the different types of knowledge elicited from the expert and is more appropriate for some tasks. Expert functions were used to represent procedural knowledge and were extremely useful for calculating different numerical characteristics related to the curve morphology, and for filtering and pre-processing the data. Rules were used to represent declarative knowledge: they are useful for representing fine-grained knowledge, like heuristics, and were used for protocol validation and in-depth data analysis. The knowledge comprised in each formalism can be used separately to provide conclusions on particular issues, but they can also be used together to provide more general conclusions.

4.4.2. A KBS lightens the workload of experts and provides support for non-specialists

The ES built for the I4 system was highly successful in two different ways. First, it is a very useful tool that is an effective aid for experienced physicians in their daily work. Second, it enables novice users to use and understand the results of an isokinetics machine. Like the isokinetics machine, many devices run proprietary software and have an interface that provides set information, making the results that they offer very hard to interpret. Consequently, only very experienced physicians are able to operate them confidently. KBSs can help expert users to do their job and be a vital aid for non-specialist users, overcoming the above difficulties.

4.4.3. A KBS can enable access for users with disabilities

A KBS can enable access for users with disabilities, as it can gather and analyze information presented in different formats, and provide a summary with text-based information that can be easily transformed to several modalities (speech, tactile dots in Braille, graphs, etc.). In our case, the I4 system offered a text-based user interface for numerical data that was, at the time, only used in graphical format. The text-based summary information provided by the system included the shape and other visual characteristics of the curves that blind users could then use to access and correctly interpret isokinetics data. This enabled blind physicians to perform a new task that they were unable to perform before the development of the I4 system. On the one hand, they could analyze the isokinetics data without external help and, on the other, they had an intelligent system that provided support for many tasks and decisions.

4.4.4. The development of a KBS can improve the performance of expert knowledge

During the interviews with the knowledge engineer, the expert had to formulate and explain her knowledge. This led to an improvement in the structure and, consequently, the efficiency of her knowledge. The improvement in expert knowledge performance is especially significant in underexplored domains, about which little is known.

4.4.5. In problems in which privacy, safety or security issues have to be considered, special care is needed when dealing with the users and experts, and when using the data

In some domains, such as the medical domain, data confidentiality is highly important and it is necessary to guarantee compliance with data protection regulations. A less obvious, but equally important, point is to mind the interaction with users, be they physicians or patients, and be especially sensitive to how they are treated. For example, system outputs should be provided more than ever as decision support, as the physician is the only one that can establish a diagnosis. Therefore, both the knowledge engineer and the applications developer should be especially careful in this type of domains.

4.4.6. Shortage of experts calls for exhaustive data processing

One option in domains where there are few experts and/or a shortage of knowledge available in the public domain is to resort to an exhaustive analysis of the stored data. In our case, the system was based on the knowledge of just one expert, because there was a shortage of domain experts and the literature on isokinetics data analysis was sparse. Taking into account that the expert had run and stored a great many isokinetics tests, we resorted to an in-depth analysis of these data for the purpose of learning from the test cases to counterbalance this handicap.

4.4.7. The expert systems approach is not enough when expertise is limited

Isokinetics is a poorly explored domain, where there is neither public knowledge nor a great many experts. I4 was very successful at performing its tasks, but an expert system is evidently limited to the amount of expert knowledge there is available. In our case, some tasks, such as early injury detection or population modeling, were beyond the expertise of the available expert and as a result beyond the scope of I4. This pointed to the need to apply some other approach that would supplement the work already completed. As there were a great many isokinetic tests run on sportspeople and other patients and that had been stored but never processed or used in any way, we examined the option of exploiting the potential of the stored data by applying data mining techniques to extract implicit knowledge from the data. Therefore, in the case of limited expertise, the expert systems approach is not enough and something else is needed in order to achieve a more successful system. In cases where a huge amount of data is available, KDD and DM techniques are a very promising option.

5. I4 Phase 2: Data mining

To exploit the knowledge stored in all the available isokinetics exercises we decided to build a KDD system (Alonso, Caraça-Valente, Martínez, & Montes, 2003; Alonso, Martínez, Pérez, Santamaría, & Valente, 2006) in order to extend and validate the ES that we had built during the first phase of the I4 project. The KDD subsystem has two modules (Fig. 2, middle): one for detecting patterns (significant irregularities) in isokinetics curves and another for generating characteristic models of certain population groups. Both modules use the cleaning and expert pre-processing component of the KBS for data preparation tasks.

5.1. Data cleaning and pre-processing

The data to be collected and selected must be relevant, adequate and clean. This step is more crucial to success than decisions about which learning algorithm to use. Data preparation was an extremely pressing question in I4 due to the poor quality of the stored data: incomplete or poorly run tests, missing data in some tests, etc. Part of the ES built was used to clean the data, as it already analyzed tests for completeness and quality. This EK and DMK cooperation was very useful as it provided a very powerful mechanism for cleaning and debugging data.

5.2. Discovering similar patterns in time series

Some knee-related injuries and dysfunctions appear as characteristic patterns in the isokinetics curves. However, there is no expert knowledge about these patterns. But one of the most important potential applications of DM methods is precisely to find patterns in data. In the case of time series (like the isokinetics curves), pattern recognition means detecting representative parts of the series for characterization.

Therefore, we developed a DM algorithm for finding significant patterns (time subsequences of undetermined length) that are likely to characterize a set of non-uniform time series, even though important characteristics of these patterns, like length or position within the time series, are unknown. Major changes had to be made to existing algorithms in order to consider variable-length patterns and pattern similarity.

5.3. Creating reference models for population groups

Another common task involved in assessing isokinetics exercises is to compare a patient's test against a reference model created beforehand. These models represent the average profile of a group of patients sharing common characteristics.

We developed a DM algorithm to create reference models from a set of isokinetics curves, representing a particular group of individuals. A key aspect of the algorithm is the selection of the individuals to be used to generate the reference model and the rejection of what are considered to be outliers that, if included, would distort the model. Often, there is clearly a majority subgroup of similar exercises, which represent the standard profile that the user is looking for, and a disperse set of groups of one or two exercises that are outliers. The exercises in the majority group are used to create the reference model unifying all the exercises' common characteristics while the outliers have to be identified and discarded.

This used to be a difficult manual process done by an expert in isokinetics. We implemented a semi-automatic mechanism to take charge of the run-of-the-mill tasks, leaving the important decisions to the expert. To do this, an automatic clustering process is enacted, and the clusters are then shown for the expert to decide which set or sets of exercises the model should be based on.

At the end of the process, the system outputs the model that represents the given population, which can be used to perform several important tasks. For example, we can compare an isokinetics exercise or a set of isokinetics exercises for a patient with the models stored in the database to determine which group the patient belongs or should belong to and identify what sport he or she should go in for, what his or her weaknesses are with a view to improvement or how he or she is likely to progress in the future.

5.4. I4 Phase 2: Evaluation and results

The goal of this evaluation step was to detect and to correct any significant deviation of the partial results output during DM from the results expected by the practitioners at an early stage of development. The discovered knowledge was difficult to evaluate because there was little sound background knowledge about most of the populations under study. Furthermore, even acknowledged experts experience great difficulty in assessing the quality of a model.

Therefore, the evaluation process focused on, first, verifying whether the output patterns and models were representative and to what extent and, second, validating their fitness for achieving the selected goals: pattern-based injury detection and model-based population characterization. For a more thorough discussion of each of the evaluation tasks performed in phase 2 of I4, see Alonso, Caraça-Valente, Martínez, and Montes (2005). Here, we will just outline the results of the evaluations of injury detection and reference models creation.

To evaluate the injury detection system performance, the pattern detection algorithm was run on several sets of patients. Then new cases (15 common injuries, 5 uninjured and 5 rare injuries) were presented to the system, and the expert and novice physicians. Only novice physicians made mistakes regarding common injuries: they misclassified 2 of the 15 cases. As regards the 5 uninjured cases, the system failed to identify 1, the expert correctly classified all of them and the novices made 2 mistakes. Finally, for the rare injuries, system performance was the best, identifying all 5 cases. The expert made 2 mistakes and could not identify 1 of the cases, so she had only 2 correct responses. On the other hand, the novice physicians made 2 mistakes and could not classify any of the other 3.

To evaluate the reference models creation process first the system was used to create a reference model for a given population. Then new individuals were presented to the system, and the expert and the novice physicians. They had to decide whether the individuals belonged to the population represented by the model. The new individuals were members of the population, non-members and some unclassified individuals. System performance was best. Out of the 30 individuals belonging to the model, it made only 4 mistakes whereas the expert made 9 mistakes and the novice 21. All three correctly classified the 10 non-member individuals. Finally, the system correctly identified the 10 unclassified individuals, whereas the expert and the novice physicians each made two mistakes.

We can conclude that the results obtained with the KDD modules surpassed the expected results. Obviously, these results must be read carefully, as most of the mistakes made by practitioners can be put down to the fact that they were working without sound background knowledge of many diseases and had to deal with a huge amount of data for decision making. In this respect, the I4 system was at an advantage. Anyhow, the same practitioners have found I4 to be an extremely useful tool for their work.

The I4 system integrating the KBS and the DM components is fully operational. The users of the system claimed that it improved the work of physicians in the field of isokinetics. They listed the system's prominent features as follows:

- Physicians who are not isokinetics specialists can use the system thanks to the test interpretation, the patterns and the reference models.
- It is possible to analyze the full isokinetics strength curve. This way, the complex parameters that are of use for interpreting the tests can be inferred more correctly and completely.
- I4 increases the power of isokinetics systems. Population modeling (by sports, specialties, diseases, etc.) is used to detect both coincidences with and slight deviations from group norms. For instance, models are used for the early evaluation of the capabilities of young athletes and to detect what likelihood elite athletes have of suffering certain injuries.
- I4 has provided for the intelligent analysis of the strength curves and has improved evaluation procedures. The isokinetics system features are better exploited thanks to, for example, the automatic extraction of information concerning injuries.
- I4 has provided friendly access for medical practitioners to the isokinetics parameters and an improved graphical presentation of the results of isokinetics tests, making reports easier to understand and more useful.

Additionally, the I4 system is providing more knowledge of the characteristics of athletes' strength. This has implications for the development and evaluation of training and rehabilitation programs. The deployment of the I4 system was a major advance in isokinetics data processing, as it meant that muscle strength measurement systems could be better exploited. These issues make it highly relevant in the field of top-competition sport.

5.5. I4 Phase 2: Lessons learned

The development of this phase 2 provided the following lessons learned.

5.5.1. Knowledge from different sources can cooperate to enhance system functionality

The co-operation between the two knowledge sources, the expert (KBS functions and rules) and the KDD system (patterns and models) is able to output higher level conclusions that would be difficult to achieve using only one knowledge source. The knowledge from these two sources interacts to provide results that would be very difficult to achieve using only one knowledge source. For instance, both the rule-based KBS component and the KDD rely on the expert functions for preparing and filtering the data. These expert functions thus enable both an expert analysis of a strength curve and the process of discovering injury patterns and reference models.

However, the relations between these sources of knowledge are not one way. The injury patterns and reference models can be entered into the KBS for numerical and morphological comparisons. They can provide higher level conclusions, such as a diagnosed dysfunction or an individual's likelihood of excelling at a particular sport. This means that KDD systems can be used as a complement to elicit knowledge from the expert. Even in ideal situations (where there are many experts and a lot of public knowledge), the many previously stored data can be used to discover knowledge that can help to significantly increase existing domain knowledge or simply add to/confirm existing expert knowledge.

5.5.2. Expert participation throughout the entire KDD process is fundamental

Without skilled human supervision, blind use of data mining software will only provide the wrong answer to the wrong question applied to the wrong type of data (Larose, 2004). Therefore, the procedure for applying KDD must include experts throughout the whole process. They should start to get involved during the actual evaluation of the input data and they will be key actors in enriching the results generated by the system. Hence, it is important to define approaches that combine and integrate expert knowledge and knowledge gathered by means of data mining procedures.

In particular, expert participation is especially important for:

- Data preparation (data debugging, filtering and completeness). KDD systems require good quality data to achieve good results. In the I4 system we were able to use expert functions (developed in phase 1) to perform data cleaning and pre-processing. This use of expert knowledge, represented as functions, improved the quality of the results of the KDD system. The knowledge of an expert is vital for the data cleaning and pre-processing phases in the KDD process.
- Evaluate the results of KDD. This is the typical case of cooperation between ES and DM. Expert knowledge is required to validate the results of KDD process before they can be used in the real world. Usually a KDD system produces large sets of results, most of which are quite worthless. Expert knowledge is needed to select the more relevant and applicable results of the KDD process. For example, expert knowledge can be used to select, debug and validate the models and patterns discovered during the DM process. In our case, expert participation was also important because she identified with the project, which overcame the traditional reticence of experts concerning technologies of this sort.
- The participation of an expert during the KDD process is also important to deal with possible inconsistencies between the EK and the DMK. There is no guarantee of positive results when mining data for actionable knowledge (Larose, 2004). That is, there is no guarantee that the knowledge gathered by the discovery algorithms is consistent with the knowledge elicited from the expert. If there are inconsistencies, the expert has to decide which knowledge is more adequate for the system under development.

5.5.3. A KDD system can increase expert domain knowledge

The results of this phase of KDD (injury patterns and reference models) revert to the expert because they enable him or her to advance in his or her knowledge of the domain, further exploring areas that were not accessible before (further investigating what the injury patterns are like and what consequences they have). This lesson learned is applicable in application domains where experience is limited and the existing data could be put to more uses than those provided by expert-elicited knowledge-based systems.

5.5.4. Data preparation is highly domain dependent

Data preparation is so domain dependent that it is very hard to extrapolate the methods used to other domains. During data preparation, some of the key problems are related to:

- Data structure: Problems such as unnecessary fields or missing important data are very frequent.
- Deficient data collection and updating: Many data are still very often collected by hand, and it is even quite common for them to be stored on paper without using computer media. Even when digital storage media are used, we often find that the data are not updated or maintained as they should be. The effects of this are devastating for the KDD process.

These problems mean that KDD activities have to be carried out with few usable data, where data are poor and there is too much noise. This leads to the problem of distinguishing between noise and exceptional cases. In these cases, data preparation activities are of utmost importance and depend on domain knowledge to be carried out successfully.

5.5.5. The actual KDD process can check for outliers in the creation of reference models for population groups

When creating a reference model for a given population, outliers may turn up. Outliers are members of the population, but have special characteristics that make them very unlike their peers. This means that they are exceptions rather than rules for modeling that population. For example, if we had a population of basketball players and it contained a very short player, this player has very different physical characteristics to the other members of the population, characteristics that would not be representative.

These outliers may significantly deviate from the expected behavior model for the above group, thereby detracting from model representativeness. To do this, it is best to remove these outliers before creating the reference model. Traditionally, this is a job usually commended to the domain expert, but it is a painstaking chore that can be prone to errors and omissions. Additionally, an individual is very likely to be characterized as an outlier multidimensionally (that is, depending on the combination of more than one characteristic), making it difficult for anyone to recognize the above case.

In such situations a system can apply automatic techniques to extract outliers as a first step in the reference model definition process. This is what we did in the I4 system, which provides automatic support for the expert to discard all the exercises that are different. This is very helpful for the physician.

5.5.6. KDD is able to build new functionalities into the system

An important lesson to be drawn from attempts to apply machine learning to a scientific problem like I4 is that, if successful, it provides a high-performance system with new, unprecedented functionalities. In the case of I4, the KBS was unable to generate or use reference models. Our KDD subsystem is capable of generating such models and applying them for new purposes, such as, for example, predicting the best sport for a young sportsperson or forecasting a patient's propensity to suffer injuries in the future.

5.5.7. Facing problems in different domains poses new challenging problems for the computer scientist

Another lesson learned is that new applications in new domains often generate challenging new questions and directions for scientific enquiry within computer science. For example, owing to the special features of our problem, major changes had to be made to the state-of-the-art algorithms in order to consider pattern similarity using the Euclidean distance, as the related work either searched for identical patterns in the series or considered only patterns of a fixed given length.

6. I4 Phase 3: Symbolic data mining

Although the results of the I4 data mining system were correct, we found that medical specialists sometimes found the system's conclusions, expressed in numerical terms, to be unsatisfactory and difficult to justify. To improve system usability and the expert's confidence, it was decided to transform the numerical time series into symbolic series so that they could be interpreted in the same way as an expert does. We used expert knowledge in this transformation of data into symbolic series to capture the key concepts from the viewpoint of isokinetics series analysis (Fig. 2, right).

To do this, we designed a symbols alphabet (ISA), a symbols extraction method and a symbolic distance. The *Symbols Extraction Method* (SEM) transforms the numerical isokinetics curves into symbolic sequences represented according to ISA as shown in Fig. 3.

The symbols are labeled with the region to which they belong (extension or flexion) and characterized by their type (for instance,

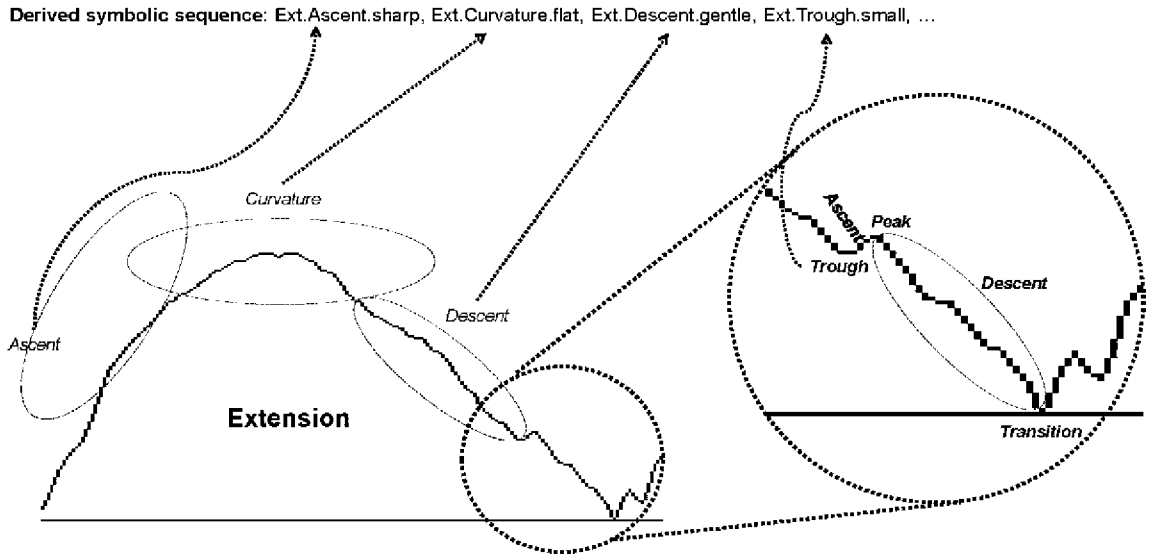

Fig. 3. The symbols of an isokinetics curve.

peaks and troughs will be characterized by their size, curvatures by their shape, etc.) as shown in Fig. 3.

To be able to compare symbolic series, we propose a variation on the Needleman–Wunsch distance (Needleman & Wunsch, 1970). The suggested distance – weighted symbolic distance (WSD) – allocates a variable cost to the *insert, substitute* and *delete* operations. This cost depends on the symbol and symbol type to be inserted or deleted.

Not all the operations or all the symbols in the isokinetics field have the same importance. For example, curvatures are symbols that are part of any repetition, whereas peaks and troughs are circumstantial symbols, usually induced by minor patient injuries and, therefore, may or may not appear. Additionally, a large peak cannot be considered the same as a small peak.

With the help of an isokinetics expert, we defined a *graph structure*, where the majority of the cost of substituting two symbols is determined by the symbol, whereas the symbol type serves to refine that cost. The expert found this structure easier to understand, and it is much easier to introduce changes due to alphabet variations or on any other ground. For a more detailed description see Alonso et al. (2006).

6.1. I4 Phase 3: Evaluation and results

The evaluation of the system at this stage focused on two points: (a) check whether the physiotherapist achieved more efficient results by analyzing symbolic isokinetics sequences than using numerical isokinetics sequences; (b) check whether the results achieved by the system comparing symbolic sequences using the symbolic distance were more significant than comparing their respective numerical sequences.

For point (a) the expert and novice physicians who participated in the I4 project development were given 34 isokinetics tests (including both the numerical sequence and the respective symbolic sequence) for 20 healthy patients, 8 with a common injury (torn ligament) and 6 with an unusual injury (osteochondritis). Both the expert and the novice physicians had to classify each test as showing no injury, torn ligament injury or osteochondritis injury.

We found that the symbolic sequence yielded better results than the numerical sequence for both the expert and the novice physicians, although the results were better in relative terms in the case of novice physicians. The expert correctly classified all the healthy patients and the torn ligament injuries. However, the expert was unable to correctly classify two of the osteochondritis cases using the numerical sequence, which she was able to classify using the symbolic sequence. The novice physician was able to correctly classify two healthy patients and two torn ligament injuries.

For point (b), the I4 system was fed with 28 knee isokinetics tests, each performed by a different sportsperson: 20 had no injury, 5 had a common knee injury (torn ligament), and 3 had an unusual injury (osteochondritis). The system was also given 3 reference models, built by applying the symbol extraction method to the respective numerical models: 1 without injuries and 2 with the above-mentioned injury types. We calculated the distance between each test and the 3 models. The WSD average distance between each symbolic test and its corresponding model was slightly lower than the distance between each numerical test and its model. But when each test was confronted with the models it did not fit, the difference was much bigger for the symbolic distance. The conclusion is that the WSD distance is more discriminative than its numerical counterpart because it focuses more on the singular points (curvatures, peaks and troughs) that define the injury.

This third phase of I4 is still under development, but has already returned two major results. First, we obtained isokinetic symbolic sequences that are more useful for the expert and novice than their respective numerical sequences. Second, the designed WSD for the symbolic sequences that is more discriminative than the Euclidean distance applied to the numerical sequences. Current work focuses on creating reference models in the symbolic domain, applying grammar-guided genetic programming.

6.2. I4 Phase 3: Lessons learned

6.2.1. Benefits of using non-numerical methods

The results of the comparisons revealed a point that we had not noticed earlier. It was related to the fact that the system using numerical methods calculated things differently to how the expert worked within each population. The reason for these differences is that the system, driven by the algorithms it uses, compares the whole curves, whereas the expert focuses on only certain aspects of the curves (such as maximum values, slopes, etc.), aspects that can be associated with semantic criteria. This discovery was

extremely important, because it allowed us to define a new line of work, namely using symbolic sequences. The use of symbolic methods to analyze the isokinetic tests has provided major benefits, not only as regards the efficiency of better detecting injuries and discriminating patterns and reference models, but also for experts in their everyday work as they work better with this type of tests.

6.2.2. Importance of expressing results and reasoning in the expert's terms

Phase 3 provided a key lesson learned: it is not enough for the system operation and results to be understandable for the expert, but it is also necessary for the results to be expressed in a notation that is familiar to the human expert.

In line with the above, even if the results and model presented to the expert are useful, they may have the drawback of not being expressed in the terms used normally in the domain. This is not only a matter of the known problem that algorithms provide numerical results that have to be translated to values that are understandable for the expert, but also that the actual algorithms should, whenever possible, emulate the expert's cognitive structure and reasoning model to achieve better quality and more understandable results.

7. Lessons learned throughout the project

Globally, the following lessons were learned.

7.1. Need for a close and continuous cooperation among the expert, knowledge engineer and KDD analyst

To increase the possibilities of success of a project of this type it is very important to have a close and continuous cooperation among the expert, knowledge engineer and KDD analyst in any knowledge discovery process and project development process. Good coordination among the members of any project is always important, but coordination is vital for project development if any or all of the following aspects occur (as they did in this project): shortage of available expert knowledge, little or no public domain knowledge, computer scientists inexperienced in the domain, etc.

In the case of the I4 project, this close cooperation between the expert and knowledge engineer played a key role in phase 1 in defining the expert functions and rules. In phase 2, cooperation between the expert and the knowledge engineer, plus the KDD analyst, played a key role in the data cleaning and preparation stages, and injury pattern discovery and numerical reference model creation. Finally, in phase 3, cooperation between the expert and the knowledge engineer played a key role in defining the isokinetics symbol alphabet, and all three cooperated on the definition of the SEM and WSD.

7.2. Need of a well-defined integration planning process for EK and DMK

An important success criterion is that the integration of EK and DMK must be a well-defined process. It should be a goal-directed activity. There are always patterns in data, and the data miner cannot judge their value unless he or she knows what to look for. The client must provide a well-defined goal, and the knowledge engineering with the human expert must provide a well-defined process, without which there is no measure for success and no way to assess the value of results. This is, however, often not possible, as the client, users and experts, who are not knowledgeable about the fields of EK and DMK, do not really know what to expect from a development of this kind. In these cases, the knowledge engineer should work in cooperation with the client to help define the goals.

The fact that the I4 project had only one expert and there was no well-founded knowledge in the isokinetics field forced us to define partial goals and not set out a clear and specific schedule for the EK and DMK process from the start. This EK and DMK process opened up new research lines when we found that the symbolic isokinetics representation matched the expert's way of thinking better than the numerical option. All this led to the extension of the research over a sizeable time period.

7.3. The full understanding of the data by the knowledge engineer and KDD analyst is vital for being able to provide novel solutions that help to bring forward the state of the art in the field

Proposing better algorithms, suitable representation structures, etc., are decisions that contribute in a large measure to system success and cannot be taken without a full understanding of the domain. For example, the selection of the representation formalisms and the inference mechanisms are just two of the decisions that largely condition project development. In this type of developments,

Table 1
Lessons learned.

	Lesson learned	Scope of application
4.4.1	Advantage of using more than one formalism to represent knowledge	General
4.4.2	A KBS lightens the workload of experts and provides support for non-specialists	General
4.4.3	A KBS can enable access for users with disabilities	General
4.4.4	The development of a KBS can improve the performance of expert knowledge	Domains with few experts
4.4.5	In problems in which privacy, safety or security issues have to be considered, special care is needed when dealing with the users and experts, and when using the data	Domains where privacy, safety and security are critical
4.4.6	Shortage of experts calls for exhaustive data processing	Domains with few experts
4.4.7	The expert systems approach is not enough when expertise is limited	Domains with few experts
5.5.1	Knowledge from different sources can cooperate to enhance system functionality	General
5.5.2	Expert participation throughout the entire KDD process is fundamental	General
5.5.3	A KDD system can increase expert domain knowledge	General
5.5.4	Data preparation is highly domain dependent	General
5.5.5	The actual KDD process can check for outliers in the creation of reference models for population groups	General
5.5.6	KDD is able to build new functionalities into the system	General
5.5.7	Facing problems in different domains poses new challenging problems for the computer scientist	General
6.2.1	Benefits of using non-numerical methods	Domains with numerical data
6.2.2	Importance of expressing results and reasoning in the expert's terms	General
7.1	Need for a close and continuous cooperation among the expert, knowledge engineer and KDD analyst	General
7.2	Need of a well-defined integration planning process for EK and DM	General
7.3	The full understanding of the data by the knowledge engineer and KDD analyst is vital for being able to provide novel solutions that help to bring forward the state of the art in the field	General

the managers must be very familiar with the domain to be able to make decisions. Due to the very nature of the domains where these projects are to be developed (especially when they are very specialized domains with little public knowledge, etc.), this is not always easy or immediate, meaning that it will require significant effort in the early project phases. Despite contrary pressures, no effort should be spared, as it will be vital for proposing the right representation structures, the best algorithms, etc., and, therefore, critical for project success.

In the case of I4, the use of the three formalisms for representing the knowledge (functions, rules and, in phase 2, reference models) helped to provide a novel conception of the isokinetics system, which emerged as a result of an in-depth analysis of the data by the knowledge engineer and KDD analyst.

8. Summary of the lessons learned

This section summarizes the findings of the research into a table listing the lessons learned described above. Table 1 includes a column labeled scope of application that aims to define the type of domains or circumstances in which each lesson learned can be applied.

Most of the lessons learned are generally applicable, as we consider that they can be applied in almost any domain and/or system type. Other more restricted scopes are listed for other lessons learned, as they are only applicable to domains that meet a series of characteristics.

In particular, lessons L04, L06, L07 are only applicable in domains where there are few experts. If many domain experts were available then a KBS is less likely to improve expert knowledge, and exhaustive data processing is less likely to be required. On the other hand, the expert systems approach is more likely to be good enough for building useful systems on its own.

Similarly, lesson L05 is specific for domains where privacy, safety and security are critical. This lesson is not relevant in other domains.

Finally, lesson L15 applies in domains were there are numerical data that are directly evaluated by human beings. As humans are well suited for pattern detection in data but have limited processing capabilities, experts are likely to develop some symbolic-based simplified model of the numerical data and thus a symbolic approach to DM would be beneficial.

9. Conclusions

Expert knowledge and data mining discovered knowledge do not have to be two separate problem-solving alternatives. They can be used together, complementarily, to develop, validate and maintain a KBS. In this article, we have highlighted some real examples of how this cooperation can be exploited to build better systems and optimize the resulting system performance. In some cases, the cooperation between these two fields can lead to the construction of systems that would not have been built if it were not for the positive effects of that cooperation.

We detailed a real example of this cooperation in the I4 project, which our research group developed over several years. This project started off as a typical expert system development, but later had to incorporate the development of a data mining system intrinsically linked to the expert system. This second part of the project led to the implementation of a system with numerous functionalities that would not have been possible if only one of the paradigms had been used: EK or DMK.

Noteworthy is the fact that we have been able to draw numerous conclusions from the experience acquired in the development of this project, which we set out in Section 7 of this paper as lessons learned.

As regards the cooperation between EK and DMK during project development, it is worth mentioning that there were several types of cooperation in the I4 system:

- Expert functions remove incorrect tests, eliminate incorrect extensions and flexions and remove noise before applying the numerical DM for pattern discovery. They also check that the medical protocols for the tests were correctly applied.
- Expert knowledge was used to select and validate the patterns discovered by the numerical DM system. Once the candidate patterns had been discovered, the expert selected and validated the relevant patterns.
- Expert knowledge was used for guidance at the beginning of the reference model generation by semi-automatically selecting the population to be used.
- In the symbolic stage, expert knowledge was used to generate the vocabulary, gather symbolic data from the numerical data and define the weights used in the symbolic distance.

The originality of the I4 system lies in the fact that an ES was built that directly intervenes in the KDD process. Once this process is complete, the discovered knowledge can be fed back to the ES.

From our experience in the I4 project and from the other examples described, we can conclude that, no matter what the direction, the cooperation between these two disciplines helps to build superior and better validated systems containing more, higher quality knowledge.

References

Adomavicius, G., & Tuzhilin, A. (2001). Expert-driven validation of rule-based user model in personalization applications. *Data Mining and Knowledge Discovery, 5*, 33–58.

Alonso, F., Caraça-Valente, J. P., Martínez, L., & Montes, C. (2003). Discovering similar patterns for characterising time series in a medical domain. *Knowledge and Information Systems, 5*(2), 183–200.

Alonso, F., Caraça-Valente, J. P., Martínez, L., & Montes, C. (2005). Discovering patterns and reference models in the medical domain of isokinetics. In M. Kantardzic & J. Zurada (Eds.), *Next generation of data-mining applications* (pp. 393–414). Wiley-IEEE Press.

Alonso, F., Martínez, L., Pérez, A., Santamaría, A., & Valente, J. P. (2006). Symbol extraction method and symbolic distance for analyzing medical time series. In *7th International symposium on biological and medical data analysis ISBMDA* (pp. 311–322). LNCS 4345.

Alonso, F., Fuertes, J. L., Martínez, L., & Montes, C. (2000). An incremental solution for developing knowledge-based software: Its application to an expert system for isokinetics interpretation. *Expert Systems with Applications, 18*(3), 165–184.

Bernstein, A., Provost, F., & Hill, S. (2005). Toward intelligent assistance for a data mining process: An ontology-based approach for cost-sensitive classification. *IEEE Transactions on Knowledge and Data Engineering, 17*(4), 503–518.

Bethel, C. L., Hall, L. O., & Goldgof, D. (2006). Mining for implications in medical data. In *Proceedings of the 18th international conference on pattern recognition 1* (pp 1212-1215). Los Alamitos, CA: IEEE Computer Society.

Caraça-Valente, J., Lopez-Chavarrías, I., & Montes, C. (2000). Functions, rules and models: Three complementary techniques for analyzing strength data. In *ACM symposium on applied computing*. SAC, Como, Italy.

Cooke, C. D., Santana, C. A., Morris, T. I., DeBraal, L., Ordonez, C., Omiecinski, E., et al. (2000). Validating expert system rule confidences using data mining of myocardial perfusion SPECT databases. *IEEE Computers in Cardiology, 27*, 785–788.

da Silva, I. G. L., Amorim, B. P., Campos, P. G., & Brasil, L. M. (2002). Integration of data mining and hybrid expert system. In *Proceedings Florida AI Research Society Officers* (pp. 267–271). AAAI.

Daniels, H., & van Dissel, H. (2002). Risk management based on expert rules and data-mining: A case study in insurance. In *Proceedings of the 10th European conference on information systems* (pp. 1589–1594). Gdańsk, Poland.

de la Vega, E., Sandoval, G., Garcia, M., Nunez, G., Al-Kinani, A., Holy, R. W., et al. (2010). Integrating data mining and expert knowledge for an artificial lift advisory system. In *SPE intelligent energy conference and exhibition*. Utrecht, The Netherlands.

Flior, E., Anaya, T., Moody, C., Beheshti, M., Jianchao, H., & Kowalski, K. (2010). A knowledge-based system implementation of intrusion detection rules. In *Proceedings of the seventh international conference on information technology: New generations* (pp. 738–742). Las Vegas, NV.

Gleeson, N. P., & Mercer, T. H. (1996). The utility of isokinetics dynamometry in the assessment of human muscle function. *Sports Medicine, 21*(1), 18–34.

Holmes, G., & Cunninghan, S. J. (1993). Using data mining to support the construction and maintenance of expert systems. In: *Proceedings of the first New Zealand international two-stream conference on artificial neural networks and expert systems* (pp. 156–159). Los Alamitos, CA: IEEE Computer Society Press.

Hong, T., & Han, I. (2002). Knowledge-based data mining of news information on the internet using cognitive maps and neural networks. *Expert Systems with Applications, 23*, 1–8.

Hu, J., & Liu, Y. (2006). Designing and realization of intelligent data mining system based on expert knowledge. In *Proceedings of the IEEE international conference on management of innovation and technology* (pp. 380–383). Singapore.

Huang, A., Zhang, L., Zhu, Z., & Shi, Y. (2009). Data mining integrated with domain knowledge. *Communications in Computer and Information Science, 35*(5), 184–187.

Kopanas, I., Avouris, N. M., & Daskalaki, S. (2002). The role of domain knowledge in a large scale data mining project. In I. P. Vlahavas & C. D. Spyropoulos (Eds.), *SETN 2002. LNAI 2308* (pp. 288–299). Berlin Heidelberg: Springer-Verlag.

Kusiak, A., & Shah, S. (2006). Data-mining-based system for prediction of water chemistry faults. *IEEE Transactions on Industrial Electronics, 53*(2), 593–602.

Lama, E., Mello, P., Nanetti, A., Riguzzi, F., Storari, S., & Valastro, G. (2006). Artificial intelligence techniques for monitoring dangerous infections. *IEEE Transactions on Information Technology in Biomedicine, 10*(1), 143–155.

Larose, D. T. (2004). *Discovering knowledge in data: An introduction to data mining.* Hoboken, NJ: John Wiley and Sons.

Lima, E., Mues, C., & Baesens, B. (2009). Domain knowledge integration in data mining using decision tables: Case studies in churn prediction. *Journal of the Operational Research Society, 60*, 1096–1106. doi:10.1057/jors.2008.161.

Mejía-Lavalle, M., & Rodríguez-Ortiz, G. (1998). Obtaining expert system rules using data mining tools from a power generation database. *Expert Systems with Applications, 14*(1), 37–42.

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequences of two proteins. *Journal of Molecular Biology, 48*(3), 443–453.

Peng, Y., & Kou, G. (2008). A domain knowledge-driven framework for multi-criteria optimization based data mining methods. In *4th International conference on networked computing and advanced information management* (pp. 46–49).

Phuong, N. H., Phong, L. L., Santiprabhob, P., & Baets, B. D. (2001). Approach to generating rules for expert systems using rough set theory. In *9th Joint international conference IFSA world congress and 20th NAFIPS* (pp. 877–882). Vancouver, BC.

Wang, X., Liu, H. Q. P., & Cheng, Y. (2004). A self-learning expert system for diagnosis in traditional Chinese medicine. *Expert Systems with Applications, 26*(4), 557–566.

Weiss, S. M., Buckley, S. J., Kapoor, S., & Damgaard, S. (2003). Knowledge-based data mining. In *Proceedings of the 9th international conference on knowledge discovery and data mining* (pp. 456–461). ACM, Washington, DC, USA.

WuJing, J. (2001). Using expert system and KDD in optimization of mobile network. In *Proceedings of the international conference on info-tech and info-net 2* (pp. 240–245). IEEE, Piscataway, NJ.

Zhang, J., & Figueiredo, R. J. (2006). Application classification through monitoring and learning of resource consumption patterns. In *20th International parallel and distributed processing symposium.* doi: 10.1109/IPDPS.2006.1639378.