Combining expert knowledge and data mining in a medical diagnosis domain

Fernando Alonso\textsuperscript{a,}\textsuperscript{*}, Juan P. Caraça-Valente\textsuperscript{a}, Angel L. González\textsuperscript{b}, César Montes\textsuperscript{b}

\textsuperscript{a}Department of Languages and Systems, Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo, s/n, 28660 Boadilla del Monte, Madrid, Spain

\textsuperscript{b}Department of Artificial Intelligence, Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo, s/n, 28660 Boadilla del Monte, Madrid, Spain

Abstract

The medical diagnosis system described here uses underlying knowledge in the isokinetic domain, obtained by combining the expertise of a physician specialised in isokinetic techniques and data mining techniques applied to a set of existing data. An isokinetic machine is basically a physical support on which patients exercise one of their joints, in this case the knee, according to different ranges of movement and at a constant speed. The data on muscle strength supplied by the machine are processed by an expert system that has built-in knowledge elicited from an expert in isokinetics. It cleans and pre-processes the data and conducts an intelligent analysis of the parameters and morphology of the isokinetic curves. Data mining methods based on the discovery of sequential patterns in time series and the fast Fourier transform, which identifies similarities and differences among exercises, were applied to the processed information to characterise injuries and discover reference patterns specific to populations. The results obtained were applied in two environments: one for the blind and another for elite athletes. © 2002 Elsevier Science Ltd. All rights reserved.

Keywords: Knowledge discovery; Data mining techniques; Expert knowledge

1. Introduction

This paper shows the results of the I4 project\textsuperscript{1} (Intelligent Interpretation of Isokinetic Information). It describes a medical diagnosis system in the field of physiotherapy and, more specifically, muscle function assessment based on isokinetic machine data, using an expert system and data mining techniques. An isokinetic machine can be described as apparatus on which patients perform strength exercises. This machine has the peculiarity of limiting the range of movement and the intensity of effort at a constant speed (which explains the term isokinetic). Data concerning the strength exerted by the patient throughout the exercise are recorded and stored in the machine so that physicians can visually analyse the results using specialised computer software.

The information supplied by an isokinetics machine has a lot of potential uses (López-Illescas, 1993): muscular diagnosis and rehabilitation, injury prevention, training evaluation and planning, etc. However, the software built into these systems, and even the isokinetic-based diagnosis techniques themselves, still have some significant handicaps that have detracted from the success of this field:

- Standard software provides only an analogical representation of the massive data flow output by these systems. The physician is left to analyse this with no further help. This is not an easy task, as it depends almost exclusively on the personal experience of the therapist.
- Novice therapists find it enormously difficult to interpret and understand the output graphs.
- Decisions are guided by the therapist’s instinct, as there are no models that can be used as a reference for most of the common injuries. Moreover, the few simple models that do exist have merely been stated by experts and are not founded on rigorous data analysis. However, there is a huge amount of stored information (performed tests) that has not been analysed to improve the procedure as a whole.

Due to the above-mentioned problems, system design should combine both practitioner expertise and knowledge that can be discovered within the data. Three objectives

\textsuperscript{*} Corresponding author. Tel.: +34-91-3522546; fax: +34-91-3526388. E-mail address: falonso@fi.upm.es (F. Alonso).

\textsuperscript{1} I4 was developed in conjunction with the Spanish National Centre for Sports Research and Sciences and ONCE (Spanish National Organisation for the Blind).
were therefore defined:

(a) Equip the isokinetic system with a knowledge-based system (KBS) that would perform an intelligent analysis of the strength curves output by the isokinetic test on which the assessments are based, modelling the knowledge of the expert who works the machine. This would be a valuable aid for the examining physician in detecting possible injuries.

(b) Characterise injuries. Bearing in mind that a huge number of isokinetic tests had already been performed and stored in a database, we aimed to find any sort of patterns useful for characterising different injuries in terms of isokinetic data. These patterns would be extremely valuable in two ways: as a useful research tool for therapists adding to the knowledge about the isokinetic shapes of common injuries and as reference models to be used for injury classification and, if possible, diagnosis. Data mining techniques based on the discovery of sequential patterns in time series were applied for this purpose.

(c) Create reference models. The third objective involved discovering standard patterns that characterised specific population types taken from the isokinetic data already prepared and stored in the database. For example, the process of evaluating a particular athlete against a standard curve, specific to his or her sport, is a very effective means for assessing athletes’ capacity and potential for the sport they intend to go in for.

2. Overview of the system

Fig. 1 shows the architecture of the I4 system from the viewpoint of its functionalities.

The isokinetics machine includes the LIDO™ Multi-Joint II system, which supplies the data from the isokinetic tests run on patients. Seated patients move their right or left leg within a 0–90° flexion/extension arc (see Fig. 2). The system records the angle, as well as the strength exerted by the patient.

After the isokinetic tests have been run by the LIDO system, they are stored in the LIDO DB. As shown in Fig. 1, the first operation performed by I4 is to decode, transform and format the LIDO DB data into a standard format, correct any inaccurate or incomplete particulars and store the result in the I4 DB. This is the only I4 module that depends on the LIDO isokinetics system.

These transformed data are stored in a separate database for later processing. The Visualisation Module can display stored exercises either individually or jointly as graphs. So, this module can be used to analyse an individual exercise, to compare any pairs of exercises or even to compare an exercise with a pattern or model that is representative of a particular population group.

The Data Cleaning and Pre-processing Module automatically processes the data stored in the database in order to correct and remove any inconsistencies between the isokinetic tests. These data are processed on the basis of the
expert knowledge stored in the KBS. After this operation, the data are ready for analysis by the KBS and by KDD (knowledge discovery in databases).

The Intelligent Analysis Module is the core of the KBS. Its job is to interpret the isokinetic data of the database using the system expert knowledge. It aims to provide an assessment of the numerical parameters and morphology of the isokinetic curves. The Report Generator Module is responsible for editing and printing the reports that describe the KBS inference process on the processed exercises.

The above three modules—Data Cleaning and Pre-processing, Intelligent Data Analysis and Report Generation—are what make up the system’s KBS architecture.

At the same time, the Characterise Injuries Module processes the exercises in the database that are indicative of any sort of injury. Its goal is to discover sequential models in time series in order to detect patterns repeated across more than one isokinetic exercise. These can then be used to class any injuries. The Visualisation Module discussed above can display the output models.

Finally, the Create Reference Models Module processes all the exercises proper to a given patient population and gets a reference curve for the group. To do this, it is necessary to assure that the curves of the patient population are similar and to discard any atypical curves.

The fast Fourier transform is used to convert the curves into a frequency domain, thus leading to a more efficient search for similarities. The Visualisation Module can display the resulting reference models.

The above two modules (Characterise Injuries and Create Reference Models), together with the Cleaning and Pre-processing Module are the basic components of the KDD architecture.

3. Expert knowledge

The KBS provides an expert knowledge-based analysis of the isokinetic curves in relation to both their numerical parameters and their morphology. It lightens the workload of experienced physicians in isokinetics and provides support for non-specialists. However, the knowledge of this expert system is vital for the data cleaning and pre-processing phases in the KDD process.

Expert knowledge is formalised in the system using three different structures: functions, rules and isokinetic models. Each of these structures, and therefore the knowledge they contain, play a different, though complementary, role in expert analysis. Functions are used to extract conclusions or facts that can be calculated mathematically from the data. Rules model expert heuristics that deal with symbolic parameters or deductive processes. Finally, the isokinetic models include structural knowledge, which is extremely useful as a reference for decision making.

3.1. Functions

The analysis of the strength curves involves assessing a range of characteristics of extension/flexion curve morphology. These characteristics provide specialists with vital information and are the main inputs for patient assessment. Part of this analysis involves processes and calculations that can be performed by means of functions. The functions evaluate the curve characteristics and implicitly contain the expert knowledge required for this task.

The aspects evaluated by the functions are (see Fig. 3): uniformity—‘Homogeneidad de las repeticiones’ (how similar repeated extension and flexion exercises are), regularity—‘Regularidad de la gráfica’ (whether the curve has a smooth contour or a lot of peaks), maximum peak
time—‘Velocidad de alcance del pico máximo’ (the time it takes to reach the maximum peak for both extensions and flexions), troughs—‘Hundimientos’ (prolonged drops and rises of the value of the moment of exercise extensions and flexions), shape of the curve—‘Tendencia de la curva’, etc.

The design and implementation of these functions can be described as interactive human induction, that is, given a number of strength curves, the expert evaluated each one and assessed its characteristics (i.e. whether it had inputs, troughs, the shape of the curve, etc.). Then, tentative functions were implemented, whose inputs were strength curves and whose outputs were the same characteristics for the given curves. These functions were applied to a new set of tests, and the results obtained were shown to the expert for evaluation. This evaluation led to some changes in function implementation, and so on. This process ends when the methods provide the correct value in a high percentage of the cases (over 98%). It took 3–5 iterations, depending on the complexity of the interpretation of each characteristic.

3.2. Rules

Obviously, there is a great deal of knowledge (mostly heuristic) whose representation calls for more powerful structures. Fine grain knowledge concerning test validation and analysis is represented in I4 by production rules, such as ‘If there are many invalid exercises, repeat the test’.

There are over 480 rules in the system, outputting conclusions on three concerns in isokinetics analysis:

- **Protocol validation.** This part of the analysis has the mission of determining whether 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. These conclusions are used within the KDD process to remove the tests that do not comply with the minimum requirements for assessment.

- **Numerical analysis of data.** Every numerical feature of the curve (maximum peak, total effort, gradients of the curve, etc.) is expertly analysed and conclusions are presented to the user. There is an individual analysis of each leg and a comparison between both legs.

- **Morphological analysis of data.** The last part of the rule-based subsystem analysis, concerned with the morphological aspects of the data, takes into account the output of the expert functions described in Section 2. The rules cannot evaluate the morphology of the strength curves (this is what the functions do); the rule-based subsystem analyses the morphology of the strength curve of each leg and their comparison and tries to identify dysfunctions.

The set of rules was induced by means of a series of interviews with an expert in isokinetics applied to elite athletes. Other physicians, who cannot be considered experts in isokinetics but have some experience in isokinetics applied to patients in general, were interviewed so as to assure that no bias was introduced by the above factor.

3.3. Isokinetic models

One of the processes most commonly performed to evaluate patient strength is to compare the test results against a standard. This third structure, called isokinetic model, aims to reflect the normal isokinetic values for a given population group. It is composed of a standard isokinetic curve and a set of attributes. The curve is automatically calculated from a set of tests. The user selects a group of tests, usually belonging to patients of similar ages, the same sex, similar sport if any and/or same injury, and the system calculates the reference curve for the group in question. This requires some sort of pre-processing (to discard poor exercises, standardise the curve, etc.), for which purpose functions are used. However, there may be some heterogeneity even among patients of the same group. Some patients will have a series of particularities that make them significantly different from the others. Take a sport like American football, for instance, where players have very different physical characteristics. Here, it would make no sense to create just one model, and separate models would have to be built for each subgroup of players having similar characteristics. Therefore, exercises have to be sifted, and the reference model has to be built using exercises among which there is some uniformity.

The discrete Fourier transform, whose use for comparing time series is widely documented (Agrawal & Strikant, 1995), is used to perform this process efficiently. The discrete Fourier transform of a time series (isokinetic exercise) $S_i$ is given by

$$S_{df} = \frac{1}{\sqrt{n}} \sum_{f=0}^{n-1} s_i \exp(-j2\pi ft/n) f = 0, 1, \ldots, n - 1$$

This technique drastically reduces the number of comparisons to be made, which is a very important factor in this case, since there are a lot of exercises for comparison in the database and comparison efficiency is vital.

Once the user has selected all the patient tests of which the model will be composed, the process for creating a new reference model is as follows:

- Calculate the discrete Fourier transform of all the exercises.
- Class these exercises, using some sort of indexing to rapidly discard all the exercises that deviate from the norm. The system creates a variation of the $R^*$ search tree (Beckman, Kriegel, Schneider, & Seeger, 1990). So, the exercises are classified, and the groups of similar exercises are clearly identified.
• Users generally intend to build a reference model for a particular group. In this case, there is a clear majority group of similar exercises, which represents the standard profile that the user is looking for, and a disperse set of groups of one or two exercises. The former is used to create a reference model.

The first step for creating the actual model is normalization. This step levels out the size of the different isokinetic curves and adjusts the times when patients exert zero strength (switches from flexion to extension and vice versa), as these are singular points that should coincide in time. The second step is to calculate the mean value of the curves point-to-point.

The discrete Fourier transform will be used later to compare an isokinetic exercise for a patient with the models stored in the database. For example, a long jump athlete can be compared with a group of elite long jumpers, a promising athlete can be compared with a set of models to determine which is the best suited discipline, strength dysfunctions can be assessed in apparently normal patients, etc.

3.4. Combining different knowledge representations

The use of three different structures for representing isokinetic expertise is important, for two main reasons:

• The isokinetic domain (like most domains) contains different sorts of knowledge, each of which is better suited to a particular knowledge representation structure;
• The co-operation between the three knowledge representation structures is able to provide higher level conclusions than only one knowledge representation structure.

The models also input facts to the rule-based subsystem for assessing comparisons between a patient and a model. Numerical (e.g. a patient is 20% stronger than the reference group) and morphological comparisons are inputs to the rule-based subsystem that provide a higher level conclusion than a diagnosed dysfunction or any other relevant finding.

4. Data mining

4.1. Pre-processing

Data was collected and prepared before executing the data mining process. These tasks are described very briefly below. The pre-processing phase is described in detail in (Alonso, Lopez-Chavarrias, Caraça-Valente, & Montes, 2001).

• Collection of initial data. This collection of data was composed of close to 1580 isokinetic tests. The tests include the personal particulars of the patient and six isokinetic exercises. Each exercise is a series of 350–600 triplets of real numbers (strength, angle, time). All this amount to just over 103MB of information.

• Preparing the data. This task involved the following actions: data analysis and decoding, creation of the I4 database, expert cleaning of data (removal of incorrect tests, elimination of incorrect extensions and flexions) and expert pre-processing. Expert knowledge had to be used to automatically remove the irregularities in the strength curves caused by isokinetic machine inertia and retain deviations that were due to the strength exerted by patients. Fig. 4 shows a diagram of the data preparation tasks.

4.2. Detecting injury patterns in isokinetic exercises

One of the most important potential applications of data mining algorithms for this sort of time series is to detect parts of the graph that are representative of an irregularity (Povinelli, 1999). As far as isokinetic exercises are concerned, the presence of this sort of deviations could be representative of some kind of injury, and the correct identification of the deviation could be an aid for detecting the injury in time. So, the identification of patterns (portions of data that are repeated in more than one graph) is of vital importance for being able to establish criteria by means of which to class the exercises and, therefore, patients.

Isokinetic exercises have a series of characteristics that cannot be overlooked when designing a pattern identification algorithm. Remember that each datum in an isokinetic exercise is a measurement of strength at a given time. Owing to the special characteristics of the individuals who complete isokinetic exercises, the graphs may have different amplitudes and be distant in time, even if the same pattern is observed. Therefore, some sort of distance has to be used to take into account not only the parts that are exactly repeated but also any that are more or less the same. Another particular to be considered in pattern search is that there is no expert knowledge about the possible patterns and their length.

Given a set of isokinetic exercises (time series) \( S = \{S[i], i = 1...n\} \), a pattern \( p \) will be a sequence of numbers that is repeated in enough series \( S_j \in S \). The number of appearances of the pattern is known as frequency, and this frequency, divided by the total number of series \( n \) in \( S \), is called pattern confidence.

A pattern \( p \) of length \( l \) is said to be a frequent pattern if its confidence is greater or equal to a given threshold \( \epsilon \). The goal of this algorithm is to find frequent patterns in the set of series processed by the algorithm. The Euclidean distance is used as a measure of distance. It is defined as

\[
\text{Distance}(S[i], S[j]) = (\sum_{k=1}^{n} (S[i]_k - S[j]_k)^2)^{1/2}
\]

The main problems of searching for what are originally unknown patterns are the exhaustive use of memory and the
time it can take an algorithm to run through all the isokinetic exercises stored in the database. One way of reducing the search space is to use the property known as a priori (Agrawal & Strikant, 1995). This property states that if a pattern is infrequent, that is, if its confidence is not over the threshold value, a pattern that contains this pattern cannot be frequent. This means that rather than checking all the patterns, only patterns containing infrequent sub-patterns have to be inspected, that is, the patterns of length $i$ will be used as filters for the candidates of length $i + 1$.

4.2.1. Method for discovering typical injury patterns

The process of developing a data mining algorithm for identifying patterns that potentially characterise some sort of injury was divided into two phases:

(a) Develop an algorithm that detects similar patterns in exercises (described below).

(b) Develop an algorithm that uses the algorithm developed in point (a) and is capable of detecting any patterns that appear in exercises done by patients with injuries and do not appear in exercises completed by healthy patients (described under Section 4.3).

A pattern search tree was built in order to speed up the pattern-searching algorithm. Each depth level of this tree coincides with the length of each pattern, that is, a branch of depth 2 corresponds to a given pattern of length 2. Rather than storing the pattern and a counter, however, the search tree nodes are composed of three fields (Fig. 5). These fields contain: the sequence of values of which the pattern is composed, the series in which the pattern appears (CA) and the series in which a pattern appears that can be considered similar to the one represented by the node (CAS).

4.2.2. Algorithm for discovering similar patterns

The problem defined in phase (a) of the method of injury identification is set out as follows:

1. Given a collection $S$ of time series, composed of
2. Find all the similar sequences of length \( 0 \leq i \leq \text{max-length} \), whose frequency within \( S \) is greater or equal to \( \epsilon \).

Major changes have to be made to state-of-the-art algorithms in order to consider pattern similarity using the Euclidean distance, as the above algorithms either search for identical patterns in the series or consider only patterns of a given length. In the identical pattern-searching algorithms, each pattern matches a branch of the tree. In the similar pattern-searching algorithm in question, however, a pattern can match several branches, depending on the specified distance \( d \). For example, taking a distance of 1, the patterns \{12, 14, 16, 18\} and \{12, 14, 16, 19\} are considered similar, and this must be taken into account when calculating the frequency of the two patterns.

First, patterns that appear in the time series are built in the same manner as an identical pattern-searching algorithm would do. To calculate its frequency, however, it is not enough just to store the number of times the pattern appears in the series. It is important to find out in which series the pattern appears in order to be able to analyse its similarity with other patterns. The identifiers of the series in which each pattern appears are stored in the CA field of each node. Then the algorithm has to run through the patterns to modify the CA set, taking into account the appearances attributed to similar patterns. For each pattern, all the patterns of the same length in the tree that are at a lesser distance than specified \( d \) are considered similar patterns, and the respective CA sets are updated with the series from the CA set of the other similar node.

Special care is to be taken in the pruning phase not to prune patterns, which, although not frequent themselves, play a role in making another pattern frequent. If this sort of patterns were pruned, the algorithm would not be complete, that is, would not find all the possible patterns. Only patterns that are infrequent and whose minimum distance from the other patterns is further than the required distance will be pruned. Having completed the tree pruning phase, the next level of the tree is generated using the longest patterns. The algorithm steps are described in detail below.

**Step 1**
Create patterns of length \( i \) (starting with length 1) and enter the series in which the above pattern was detected in the respective CA set. This step is the same as in identical pattern-searching algorithms (Caracã-Valente & Lopez-Chavarrias, 2000; Han, Dong, & Yin, 1998).

**Step 2**
Calculate the distance between pairs of patterns for all the patterns of length \( i \), provided the calculated distance is less than defined by the user (Maximum Distance), and update the CA sets of the nodes of the two patterns by adding the series belonging to the CA set of the other pattern. That is, for two similar patterns \( p_1 \) and \( p_2 \), \( \text{CAS}(p_1) = \text{CAS}(p_1) \cup \text{CA}(p_2) \).

**Step 3**
Prune the respective nodes. Only the patterns that are infrequent will be pruned, that is, whose confidence is less than specified (\( \epsilon \)), and whose minimum distance to the other patterns is greater than defined by the user (Maximum Distance). This means that infrequent patterns that could be part of longer patterns and could possibly raise the confidence of other patterns in the future are not pruned.

**Step 4**
Return to Step 1 if not all the nodes of the branch of the tree have been pruned and patterns of length \( i + 1 \) can be built.

### 4.3. Application to detecting injuries

The above algorithm was applied directly to detect irregularities in graphs that could possibly identify the presence of injuries. All the series of the databases were used for this purpose, and series that indicated a given injury were separated from any that did not. The above algorithm was simply applied to the two sets, getting two collections of patterns. Any patterns that appeared in the graphs of injured patients and did not appear in the graphs of healthy patients were identified as possible injury patterns and, after assessment by the expert, used to indicate an injury. The graphs obtained from future isokinetic tests can then be searched for these patterns to check whether or not there is an injury.

A short example of the application of the similar pattern-searching algorithm to injury detection is shown below, including real data. The exercises used in this test are as shown in Fig. 6. These exercises belong to two patients with cartilage problems in the analysed knee.

Fig. 7 shows the longer patterns detected by the system in the above exercises. The patterns detected by this method are now being evaluated by the expert in isokinetics, who is comparing them with new patients who have the same injury to test their validity. Algorithm performance tests have also been performed using real and random data, altering the parameters of confidence and pattern similarity with very satisfactory results. These tests found that algorithm performance will never compromise the results obtained in this sort of domains.

### 5. Evaluation and application of the discovered knowledge

It was not easy to evaluate I4 because of its hybrid approach and the lack of experience in the domain. Therefore, an evaluation process had to be designed that covered not only the overall performance of the system but also the adequacy of the models for representing the
populations for which they had been created and their fitness for achieving the proposed goals: pattern-based injury detection and model-based population characterisation. The sources of knowledge were confined to the experts, the cases database, the preliminary models and everyday practice. A five-step long-term procedure was used for both evaluation goals. An outline of the process is given below. A more detailed description is given in (Alonso et al., 2001).

(1) Subjective appraisal of the results by the expert. This step refers mainly to the data mining component, and its aim was to get a rough idea of the possible quality of the results in order to plan new data mining tasks, if required. The method used was to give the injury patterns to the experts and ask them for a documented opinion about their quality. If they felt they were sound, they were asked to provide at least five examples that confirmed the pattern; whereas if they did not agree with the result, they had to come up with the same number of negative examples. Thanks to exercises of this sort, we were able to identify some weaknesses in the DM methods. The first outcome was related not to the DM methods, but to the detection process as a whole. The planned procedure involved comparing new cases against the injury patterns discovered by the system. If the case matched a pattern, then an injury had been detected. The problem with operations of this sort is that not every single injury has a pattern, so a few cases of injury were classed as perfectly normal. At this early stage, one might think that a lot of knowledge still remains to be discovered so some margin of error is logical, but this is a short-minded reasoning. The more critical approach is to think that the above injury patterns will never be discovered or will even be discarded by the system, because they belong to rather atypical or infrequent cases. This led us to change the whole pattern application procedure, generating reference patterns both for injuries and for normal cases.

(2) Statistical tests comparing the results with previously known cases. All we had to go on for this task were the cases themselves and the preliminary models. These tests had, therefore, some limitations. The main goal was to determine the best values for the mining parameters. We used cross validation across five scenarios with different combinations of parameters extracted from the database.

(3) Turing Test-based validation tests, in which the effectiveness of the system was compared against the expert. The information supplied to both the system and the experts was exactly the same (an isokinetic test). This test was repeated for 25 cases in the injury problem (15 with common injuries, five with rare injuries and five with no injury at all). The examples had not been used during DM, and the comparison was made between our expert, a less experienced specialist and the system. Both the system and the expert identified the 15 common injuries, while the novice omitted two. The expert identified the five cases with no injury, while the system found one that matched neither
the normal nor the injury patterns, and the novice found three of the five possible injuries. Finally, for the five cases of rare injuries, the system was unable to find any match to normal or injury patterns (which was the expected behaviour).

(4) Continuous daily evaluation with real-life cases. This is a corrective stage and will continue throughout the research project life cycle. I4 has produced two applications, which are similar, except for the interface. One, called ES for Isokinetics Interpretation (ISOCIN), was designed for use by visually impaired physicians. The other application, an ES for Interpreting Isokinetics in Sport (ISODEPOR), is being used at the National High Performance Centre to evaluate the muscle strength of Spanish top-competition athletes.

(5) Evaluation of satisfaction. This is an important part of any evaluation process, although it is often overlooked. Its goal is to gain an understanding of the feelings of the practitioners when the new technology is transferred to everyday practice. Regarding I4, the members of the centre’s staff claim that this system has improved the performance of physicians working in the field of isokinetics (Alonso et al., 2001).

6. Conclusions

The development of an ES and its later refinement is mainly based on eliciting and entering experts’ knowledge of the subject into the system. Thanks to data mining techniques, a more efficient and objective process for developing an ES can be applied to complement the above, provided enough data are available from which new knowledge can be discovered.

The I4 project described here is an example of this approach applied to the expert processing of isokinetic data. Initially, the expert knowledge of the isokinetic physician was entered into the system so as to conduct an intelligent analysis of the numerical parameters and morphology of the strength curves output by the isokinetic tests. Considering the volume of tests, data mining techniques, involving processing time series to discover patterns, injuries and reference models, were then applied. The above expert system plays an active role in some of the data preparation and cleaning tasks of this knowledge discovery process (eliminating incorrect tests, exercises or parts of them and eliminating inertia peaks from the strength curves). After evaluation and validation by the expert, this new knowledge was entered into the expert system, which performed better and was more efficient than the one directly elicited from the expert.

The expert participation in the KDD process is not new. Indeed, it is important in most of these processes to have an expert who is familiar with the data to deal with data cleaning, pre-processing and evaluation. There are also several examples of co-operation in these two areas, for example, building a KBS from the discovered rules (Lee, Kim, & Rhee, 2001; Tsumoto, 1999). However, the originality of the proposal described in this paper 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.

It is noteworthy that the participation of experts throughout the entire KDD process was fundamental, as a vital source of knowledge and also to get them to identify with the project, thus overcoming their traditional reticence concerning technologies of this sort.

Acknowledgments

We would like to thank África López-Illescas and Ignacio López-Chavarrías for their co-operation in the I4 project. The I4 project was partly funded by CICYT project no. TIC98-0248-C02-01.

References


