

Discovering Similar Patterns in Time Series

Juan P. Caraça-Valente
Universidad Politécnica de Madrid
Campus de Montegancedo, s/n
28660 Boadilla del Monte, Spain
(+) 34 91 336 73 89

jpvalente@fi.upm.es

Ignacio López-Chavarrías
Universidad Politécnica de Madrid
Campus de Montegancedo, s/n
28660 Boadilla del Monte, Spain
(+) 34 91 336 73 89

nacho@pegaso.ls.fi.upm.es

ABSTRACT

In this paper, we describe the process of discovering underlying knowledge in a set of isokinetic tests, using a new algorithm to find similar patterns in a set of temporal series. 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. Then, Data Mining methods based on the discovery of sequential patterns in time series by means of which to find similarities and differences among exercises were applied to the processed information to characterise injuries of those patients. The results obtained were applied in two environments: one for the blind and another for elite athletes.

Categories and Subject Descriptors

H.2.8: Information Systems. Database Management. Database Applications. Data Mining

General Terms

Knowledge Discovery, Data Mining, Time Series.

1. INTRODUCTION

Historically, medicine has been a crucial domain for computer science applications. Many well-known models and methods have been designed to solve medical problems. The sheer volume of the data for collection, the need for better procedures and techniques to assist physicians in their work and, generally, the need for subtle or routine processes to be automated are some of the reasons behind this state of affairs.

Data Mining is not an exception. A better comprehension of medical data is a must for improving medical effectiveness. Moreover, there is a shortage of good models for certain injuries that could reduce the rate of wrong decisions being made in a host of fields.

Permission to make digital or hard copies of part or all of this work or personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers, or to redistribute to lists, requires prior specific permission and/or a fee.

KDD 2000, Boston, MA USA
© ACM 2000 1-58113-233-6/00/08 ...\$5.00

In this paper, we describe some of the results of R&D in the field of physiotherapy and, more specifically, in muscle function assessment based on isokinetic data. Physicians collect these data using a mechanical instrument called an isokinetics machine. This machine can be described as a piece of apparatus (shown in figure 1) on which patients perform strength exercises (in this case, knee extensions and flexions), which has the peculiarity of limiting the range of movement and the intensity of the effort at a constant speed (which explains the term isokinetic). Data concerning the strength employed by the patient throughout the exercise is recorded and stored by the machine so that the physician can analyse and visualise the results using specialised computer software.



Figure 1. Diagram of isokinetics machine use

The information supplied by an isokinetics machine has a lot of potential uses: muscular diagnosis and rehabilitation, injury prevention, training evaluation and planning, etc. However, the software built into these systems and the isokinetic-based diagnosis techniques themselves still have some significant handicaps, due to the lack of expert knowledge to analyze the massive data flow, that have detracted from the success of this field.

The I4¹ line of projects is aimed at conducting a more comprehensive analysis of the data output. This involves

¹ I4 stands for Intelligent Interpretation of Isokinetic Information and has been developed in conjunction with the Spanish National Centre for Sports Research and Sciences (hereinafter referred to as CNICD) and the School of Physiotherapy of the Spanish National Organisation for the Blind (hereinafter referred to as EFONCE).

intelligent understanding and interpretation of the strength curves obtained through the isokinetic exercises on which the assessments are based. This paper will focus on the Data Mining techniques that have been developed to analyze strength curves and characterize possible injuries.

1.1 Discovering sequential patterns in time series

The analysis of time-ordered data sets is essential in many fields, including engineering, medicine or the business world. The future behaviour of a particular variable can be predicted by studying how it has behaved up to a given time. Similarly, determining what other values have behaved similarly can be an aid for deciding on the actions to be taken either to conserve or drastically modify the present evolution. Data mining techniques are very useful tools for doing this.

There is a growing need to search databases for data time series that resemble a particular one. For example, it could be a matter of finding companies with a similar growth pattern or discovering products with similar sales patterns. One important question is to decide what similarity means. The simplest solution is to calculate any sort of distance, like the Euclidean or Manhattan distance, between two time series, which are considered similar if the above distance is less than a given threshold value. In an attempt at reducing the time taken to calculate the distance, some authors, like [2] and [12], have proposed the use of the Fourier transform to transfer the series from the time to the frequency domain, using only the first coefficients to filter the dissimilar series. [2] divides the queries concerning similarity between time series into two categories: total comparison, that is, the sequences for comparison have the same length, or partial comparison, which involves merely checking whether a sequence appears as part of another.

With regard to total comparison, most of the papers written so far have focused on the search for a particular sequence within a set of time series or on searching for all the sequences that are similar to a given one. [2] or [12] propose indexing all the time series, using some sort of spatial access method, like the R* trees proposed by [7], as an indexing method. This index will contain the first coefficients of the Fourier transform of the series. Thus, it will suffice to run through the sequences to be close together in the R* tree rather than through all the sequences to find the similar ones. In [24] and [8], wavelets [23] are used as a means of moving from the time to the frequency domain, although both use the Euclidean distance as a method for evaluating similarity.

Partial comparison, that is, the search for subseries that are repeated throughout a particular series, has gained in importance recently. [3] developed an A priori algorithm-based technique to discover sequential patterns. [22] address the problem of recognising frequent episodes in event collections, where episode is defined as a collection of events that occur at given intervals of time in a given partial order. The A priori property is also used in [19]. In this case, however, it is used to prune infrequent patterns. In this paper, patterns of any length are detected, while the pruning of infrequent patterns succeeds in providing a very high efficiency.

Despite the growing interest in the study of time series, there are fields of research in which there is still much work to be done.

The above-mentioned papers search for similar data curves or sequential patterns in one series. However, it may be necessary to search for patterns that are repeated not in one but in several curves in some fields. The frequency of occurrence of the subsequence in the different series is an important factor in this problem. Only subsequences that are repeated in enough data series can be considered patterns. This search for each possible pattern in all the series under consideration amounts to an important change with regard to the above algorithms. The medical domain is a clear example of the application of this problem. A huge amount of data series that measure different variables over time are stored in medical analyses. If sequential patterns that are repeated in different graphs belonging to patients with the same complaint are discovered, the above results can then be used as an aid for detecting such complaints.

The algorithm presented in [19] has proved to be useful for the above-mentioned problem. This algorithm only discovers patterns that are repeated exactly in different series, which means that it is not suitable for searching for similar patterns in more than one time series. This has obliged us to design a new algorithm, based on Han's, to tackle this problem.

This algorithm is capable of detecting similar sequential patterns in a set of time series. The algorithm uses some ideas, like the use of the A priori property to prune the search tree but adds the use of the Euclidean distance, by means of which it manages to level out the insignificant differences between patterns possibly occurring in different series.

2. KNOWLEDGE DISCOVERY PROCESS

2.1 Raw Data: Isokinetic Tests

Isokinetic systems were conceived to analyse the muscular fitness of patients who are members of any population group. During a standard session, patients must perform a set of exercises, for example, ten seconds extending and flexing their leg with the machine moving at a constant speed of 90°/s. The number of exercises and speed at which they must be performed are determined by the isokinetic protocol currently in use [21]. This protocol defines the number of exercises to be performed and at what speeds. Obviously, the protocol must always be the same, if we are to be able to analyse some patients against others, as differences in the number, order and speed of the exercises would significantly alter the isokinetic curves obtained. All of the exercises performed in a session are called an isokinetic test and, in this case, this is composed of 6 exercises for each leg at given constant speed (60, 180 or 300°/s). Each exercise is characterised by the speed and the leg used (left or right), and the angle² and the strength applied every 2/100 s by the patient is recorded. The graphic representation of these data (strength over time) resembles a sinusoidal curve, containing a lot of small peaks and other irregularities, as shown in the previous section. The amplitude, total area and irregularities are the main parameters in the analysis of the tests.

In the case with which we are concerned, isokinetic tests have been used to assess the physical capacity and injuries of top

² The angle is recorded because minor deviations from constant speed are very significant.

competition athletes since the early 90s. An extensive collection of tests and exercises has been gathered since then, albeit unmethodically. Every time an isokinetic test is run, a series of isokinetic system database registers are created. These were kept in this database until the hard disk was full and were then dumped to diskettes. The tests were saved on different diskettes, depending on the sport in question, and in the internally coded format of the isokinetic system, which uses the imperial system of measures. Therefore, most of the data were stored and coded across a set of diskettes. The sex of the patient was not defined in any of the tests and personal particulars were missing in many, because they are of no interest for muscle analysis. Additionally, some tests do not include all the exercises defined in the protocol for a variety of reasons. Hence, we had a set of heterodox, unclassified data files in different format, which were, partly, incomplete.

On the positive side, the quality of the strength data within each exercise was unquestionable, as the protocols had been respected in the huge majority of cases, the isokinetic system used was of proven quality and the personnel who operate it had been properly trained.

This collection of data was composed of close to 1580 isokinetic tests (this number grew throughout the project life cycle). In order to calculate the volume of information, it is important to take into account that the tests are formed by the personal particulars of the patient and 6 isokinetic exercises. Each exercise is a series of from 350 to 600 triplets of real numbers (strength, angle, time). All this amounted to just over 103 Mbytes of information at the start.

2.2 Data Preparation

A series of tasks, detailed in this section, had to be carried out before the available data set could be used (not only for KD tasks but also for expert analysis itself). Although the planning of objectives and milestones was very important in this domain, it was very difficult to establish a long-term plan from the start, as it was hard to foresee the final outcome of some of the phases. Instead we set short-term milestones which, when achieved, were used as a basis for planning the next steps. Figure 2 shows a diagram of the main data preparation tasks. Each of these is described in detail below.

2.2.1 Data Analysis and Decoding

The data came from a commercial application (the isokinetic system) that has its own internal data format, which are compressed and coded using special methods. Therefore, the current isokinetic system database could not be used (after loading the information from all the diskettes) to process the data, and each test had to be exported to text format. Apart from this being hard work, there was an added difficulty, as the format of this text file had to be decoded. This was not immediate, as the use of the imperial system of measures caused some confusion.

2.2.2 Creation of the I4 Database

The creation of the I4 database, which was to be organised and store the isokinetic tests run and would allow these and future tests to be used rationally, was a must for the success of the project. In view of the circumstances mentioned before, it is not hard to imagine that the process of database creation was long and

laborious. Once all the isokinetic tests run had been transferred to text files, they had to be entered in the I4 database. The tests were classified neither by patient nor by any other means and contained some incomplete data.

In order to help in this task, an automatic loading tool was designed to transfer all the tests to the I4 database, automating this task to the utmost. This loader is able to detect the exercises performed in one session (exercises that are part of one test) to allocate the same test code to all of them. This is a means of structuring separate exercises. Additionally, the test object is created containing the data common to its exercises.

With regard to incomplete data in tests and exercises, these will be included automatically when they can be calculated and, otherwise, the user is interactively asked for their value. Patient gender is a particularly interesting case, which was not defined in any of the exercises. This datum is very important, as it conditions any assessment of the exercises considerably (there is a discrepancy among the strength values of men and women). The loader was equipped with a name recognition feature to enable this datum to be added to the tests. This allows the sex to be identified depending on the patient's name, and the user is asked to enter this particular only if in doubt.

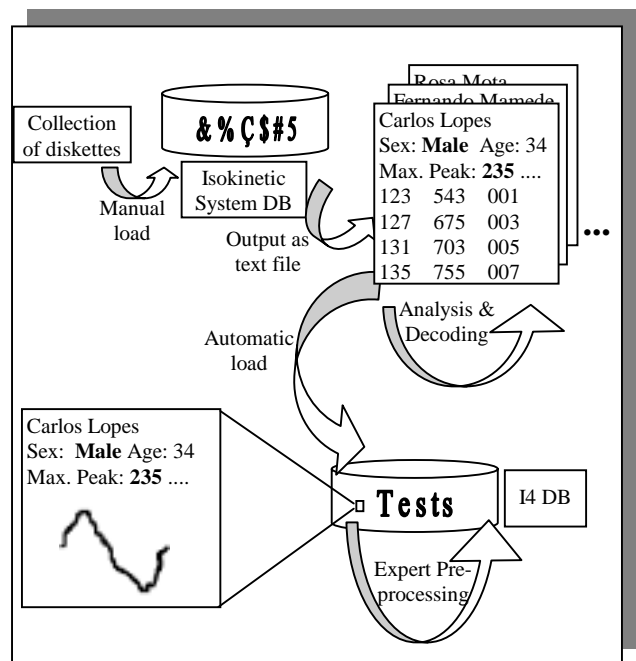


Figure 2. Database construction process and data pre-processing

2.3 Data Cleaning

Once we had a standardised database with organised and complete information, the curves had to be evaluated to identify any that were invalid for assessing patient muscle capacity and to remove any irregularities entered by mechanical factors of the isokinetic system. Expert knowledge was needed to perform these pre-processing tasks. Therefore, we built an expert system for this task (hereinafter mentioned as I4ES) [11] [6]. Two data cleaning tasks are performed by the I4ES:

- Removal of incorrect tests. This pre-processing task has the mission of determining that the isokinetic test protocol has been correctly applied. Any difference in the performance of the exercises has a significant effect on strength data. All the exercises defined in the protocol must have been completed successfully in the correct order. Additionally, the strength values must demonstrate that patients exerted themselves during the exercises and, therefore, tired to some degree. Only if this protocol validation is correct will it be possible to use the data.
- Elimination of incorrect extensions and flexions. Even if the isokinetic protocol has been correctly implemented, some of the extensions and/or flexions within an exercise may be of no use, owing mainly to lack of concentration by the patient at a particular time during the exercise. I4 detects exercise extensions and flexions that are invalid because much less effort was employed by the patient than was in others, as well as movements that can be considered atypical as their morphology is unlike the others.

2.4 Expert Pre-Processing

Having validated all the exercises as a whole and each exercise individually, they have to be filtered to remove irregularities that are not caused by patients. Again expert knowledge had to be used to be able to automatically identify the irregularities caused by the strength employed by patients and any that are due to the isokinetic machine. So, the strength curves are pre-processed in order to eliminate flexion peaks, that is, maximum peaks produced by machine inertia. This is detected when the angle at which the maximum peak is produced deviates a lot from the norm. Figure 3 shows a graph with peaks and the same graph after they have been removed.

3. DATA MINING METHOD

Data mining techniques were required to analyse isokinetic exercises in order to discover new and useful information for later use in a range of applications. Patterns discovered in isokinetic exercises performed by injured patients were very useful, in particular for monitoring injuries, detecting potential injuries early or discovering fraudulent sickness leaves.

3.1 Detecting Injury Patterns

Data mining algorithms can only be applied to the isokinetic exercises stored in the database if they are represented in a suitable format for processing. Therefore, an isokinetic exercise was considered as a sequence of numbers that represent values at a point of time. One of the most important potential applications of data mining algorithms for this sort of time series is to detect any parts of the graph that are representative of any irregularity. As far as isokinetic exercises are concerned, the presence of this sort of alterations could correspond to some kind of injury, and correct identification of the alteration could be an aid for detecting the injury in time. So, the identification of patterns, that is, 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 classify the exercises and, therefore, patients.

Isokinetic exercises have a series of characteristics that cannot be overlooked when designing an algorithm to identify patterns. Remember that each datum in an isokinetic exercise is a

measurement of strength at a particular time. Owing to the special characteristics of the individuals who complete an exercise of this sort, the graphs can have different amplitudes and be removed 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 repeated exactly but also any that are approximately the same.

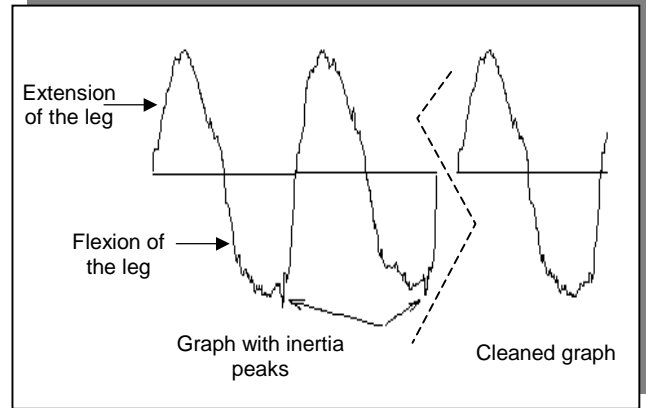


Figure 3. Exercises at 60°/s with inertia peaks and their elimination

Another particular to be considered in the search for patterns is that there is no expert knowledge about the possible patterns and their length. Therefore, all the exercises have to be run through to get patterns of different lengths. The memory consumption and execution time of this process can be very high, and these are both factors to be considered when designing the algorithm.

As mentioned above, an isokinetic exercise will be represented as a series of real numbers. That is, an exercise S_i can be represented as:

$$S_i = \{s_{i1}, s_{i2}, \dots, s_{in}\}, s_{ij} \in \mathcal{R} \forall j$$

A pattern will be a sequence of numbers that is repeated in enough exercises. The number of repetitions is known as frequency, and this frequency, divided by the total number of exercises, is called pattern confidence. Formally, the frequency of a pattern p is defined as

$$\text{frequency}(p) = \left\{ \left\{ S_i \mid 1 \leq i < m \text{ where } p \text{ appears in } S_i \right\} \right\}$$

$$\text{confidence}(p) = \frac{\text{frequency}(p)}{m}$$

where m is the total number of exercises.

The length of a pattern is called period. A pattern p of period l is said to be a frequent periodic pattern of period l if its confidence is greater or equal to a given threshold ϵ . The objective of this algorithm is to find periodic patterns in the isokinetic exercises in the database. The Euclidean distance, defined as

where R and S are two isokinetic exercises, will be used as a

$$D(R, S) = \left(\sum_{i=1}^{i=n} (R_i - S_i)^2 \right)^{1/2}$$

measure of the distance.

As mentioned above, the main problem of searching for what are originally unknown patterns is the exhaustive use of memory and the amount of time it can take an algorithm to run through all the isokinetic exercises stored in the database. One way of making the search space smaller is to use the property known as A priori [1]. This states that if a pattern is not frequent, that is, if its confidence is not over the threshold value, no pattern of a longer period that contains it will be able to 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 period i will be used as filters for the candidates of period $i+1$.

3.2 Data Mining Method

The process of developing a data mining algorithm that detects patterns in isokinetic exercises and identifies patterns that potentially characterise some sort of injury was divided into two phases:

- a) Develop an algorithm that detects patterns in exercises.
- 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.

A pattern search tree was built in order to speed up pattern-searching algorithm operation by applying the A priori property. Each depth level of this tree coincides with the period of each pattern, that is, a branch of depth 2 corresponds with a given pattern of period 2. An example of this sort of trees is given in Figure 4. Each branch of the tree will contain a counter that will specify the number of exercises in which the pattern appears. Further appearances of one pattern within the same exercise will not be taken into account. This counter can be used to prune any branches under the threshold in question, and the pattern will be ignored in the future. The patterns are formed by running through the exercises using a window of the same size as the period under consideration. So, the algorithm will be capable of forming all the patterns of any period, ending when there are no more branches to run through or when the period of the pattern is greater than the length of the longest isokinetic exercise.

For example, if a frequency of 10 were established for the tree shown in Figure 4, branch A would be pruned, because pattern {14, 16, 18} was only found in five exercises, which does not meet the minimum requirement. Any pattern {14, 16, 18, x_i , ..., x_j } will be ignored in the future.

Below we will present two pattern-searching algorithms, which we applied in the research. One considers identical sequences as patterns and the other adds the Euclidean distance to compare patterns. The difference between the two is significant and has a very considerable impact on algorithm performance. In the algorithm that does not consider distance, each pattern corresponds to one and only one branch of the search tree. Therefore, the pattern search in the tree ends when the respective branch is found. This is not possible in the algorithm that takes into account distance, as a pattern has more than one correspondence in the search tree. For example, pattern {12, 14, 16, 18} and pattern {12, 14, 16, 19} would be considered as similar at distance 1, and the tree should increase the counter of the respective branches for both patterns.

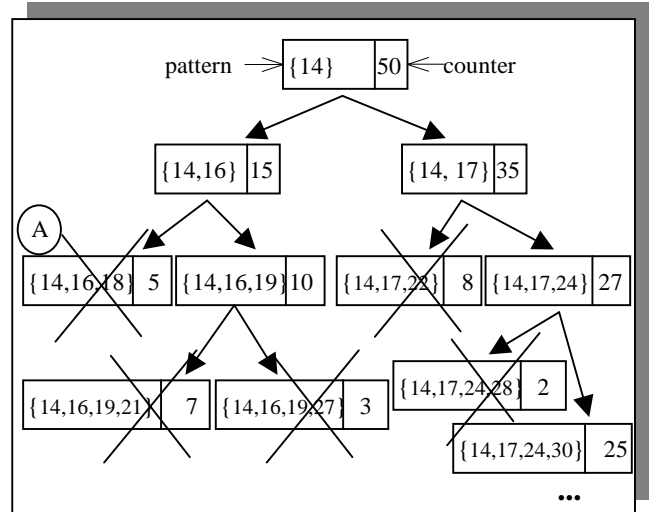


Figure 4. Pruning tree

3.2.1 Algorithm 1. Exact pattern search

The algorithm runs through the exercises in the database and forms patterns of different periods, which are inserted in the search tree and their counter is increased. The process is iterative, starting with patterns of period 1 up to the patterns with the maximum period, corresponding to the length of the longest exercise. This is a very similar process to the one presented in [19].

The search tree must be pruned before moving on to a longer pattern. In accordance with the A priori property described above, all the branches corresponding to patterns whose confidence is not over the threshold will be pruned. This means that the algorithm ignores patterns that cannot possibly be frequent.

Basically, the algorithm starts with the first exercise in the injuries database and forms patterns of period 1 with a window of the same size. As they are formed, they are inserted in the search tree, and the counter is set at one. This is meant to reflect that the above pattern appears in at least one series. This window moves through the exercise forming all the patterns. The same process is carried out on the other exercises in the database. The counter will merely be increased for exercises already entered in the search tree, and any not entered for any previous exercise are inserted with the counter at 1.

Having finished the search for all the patterns of period 1, any patterns that cannot possibly be frequent will have to be pruned. Any patterns of less than the desired frequency will not be run through again, and their branch will be pruned. The process will now be repeated with patterns of period 2, except that patterns which are part of a pruned branch will not be inserted. This rules out having to run through the tree to enter a pattern that is not going to be frequent. The algorithm ends when the window is longer than the maximum period in the tree.

At the end of algorithm execution, we will get all the patterns that have the required confidence. Although it can be useful to find patterns that are exactly repeated in the isokinetic exercises, these patterns are not sufficient in most applications. The differences in the values measured in most isokinetic exercises can cause two similar patterns not to appear as such when the algorithm is executed, losing a lot of useful patterns that should be identified.

This is why we had to develop another algorithm that enhances the potential of the one described above, using the notion of distance among patterns for this purpose.

3.2.2 Algorithm 2. Similar pattern search using Euclidean distance

After adding the pattern search distance, the algorithm becomes more complicated and has to be completely modified. Just one run through the tree could form all the patterns of a given period before. However, the number of appearances of a pattern can now change owing to other similar patterns appearing. Therefore, the tree has to be run through again to add the actual number of times a pattern appears, that is, not only the exact appearances but also occurrences of other similar patterns.

Also special care has to be taken 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. Therefore, the pruning condition must be modified. Any patterns that are not frequent and whose minimum distance from the other patterns is further than the required distance will be pruned. When the patterns are presented, only the frequent ones will be displayed, although not all the patterns in the tree are frequent owing to the above pruning condition.

The form chosen for the new algorithm was to use a variant on the above, which retained both the number of series in which a pattern appeared and a reference to the series in which it appeared. The series are run through using a window, which forms patterns, alters the counter and enters the series in which they appear. Before moving on to the pruning phase, the patterns must be run through again, this time to count the appearances attributed to similar patterns. The presence of the series entered by each pattern is helpful in this respect. For each pattern, all the patterns in the tree that are at a lesser distance than the threshold d are searched, and as they are considered similar patterns, the counters of both must be increased. Figure 5 shows a formal description of this algorithm.

The algorithm is applied directly to detect irregularities in the graphs that possibly identify the presence of injuries. All the series of the databases are used for this purpose, and series that indicated a given injury are 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 injury patterns and, after assessment by the expert, were used to indicate an injury. The graphs obtained from subsequent isokinetic tests will then be able to be searched for these patterns to check whether or not there is an injury.

Input:

- ϵ : minimum confidence
- m : number of time series
- $s[m:i]$: collection of time series

```

d: minimum distance
Code:
Begin {
l= 0;
While ((the tree has unpruned branches) or (l>i))
{
l = l + 1;
for (i=0 to i=m)
{
run through s[i] forming patterns of length = l;
insert the pattern in the tree;
}
calculate distance(l);
prune branches;
}
run through the tree forming patterns;
}end
insert pattern in the search tree
input:
tree: search tree
i: number of the time series that aims to insert pattern
pattern: pattern to be inserted
begin {
if the pattern was not inserted
insert it, place i in the collection of appearances;
}end
calculate distance
input:
tree: search tree
l: length
d: minimum distance
begin {
For each pattern p of length=l
For each pattern q of the same length
if distance(p,q)<d
unite the collection of p and the collection
of q and place it in the new collection
Keep the same distance for the other patterns
}end
}end
prune branches
input:
tree: search tree
begin {
Prune each branch:
(conf(p)< $\epsilon$ ) AND (minimum distance < d)
}end

```

Figure 5. Pattern Search Algorithm using Euclidean distance

Figure 6 shows three real examples of isokinetic exercises stored in the database. Possible patterns can best be viewed by representing the exercises alongside each other. The algorithm that does not include the Euclidean distance is incapable of finding any pattern in these exercises. However, depending on the application to which they are put, several similarities can be found. For example, using $\epsilon=0.8$ and a threshold of 50 for the distance, algorithm 2 finds the pattern shown in Figure 7.

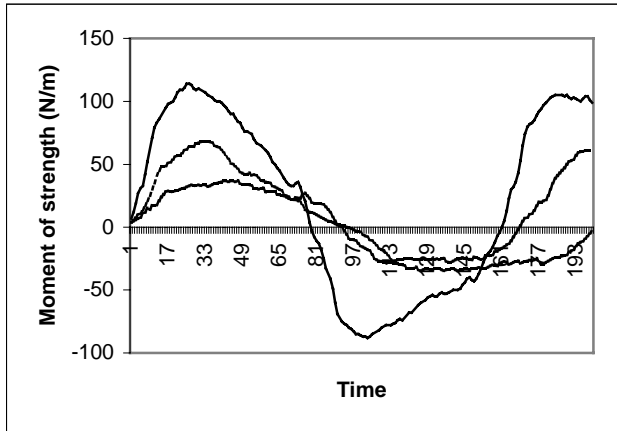


Figure 6. Isokinetic exercises

The pattern shown in Figure 7 is a magnification of a region common to the three graphs, specifically the region between the times 100 and 161. It is important to note that although the pattern appears more or less in the same region of the three exercises in this example, this is not generally the case. Additionally, the pattern does not necessarily have to appear in all the database exercises, but it must appear in enough to comply with the required confidence.

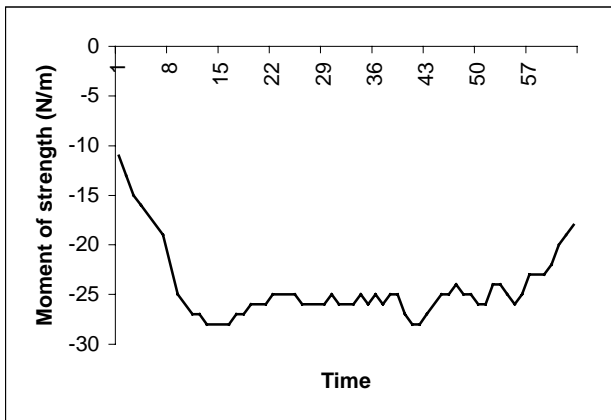


Figure 7. Magnification of the pattern found

4. EVALUATION AND USE OF THE KNOWLEDGE

The I4 project has produced two applications -the initial versions of these are described in [5] - which are quite similar, except for the interface. One, called ES for Isokinetics Interpretation (ISOCIN), was designed for use by sight-impaired physicians, so the interface includes complete voice synthesis of every piece of information presented to the user. That is, it includes information on how to use the system, the options open at any time and, of course, the isokinetic data and their interpretation. The system can be used without a display. ISOCIN is currently being used by blind physicians at EFONCE to analyse injuries and assess their evolution, adapting the physiotherapy administered and rehabilitation process.

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.

Before and after the deployment of these applications, an important evaluation phase was performed (and continues nowadays). It was not easy to evaluate the results due to the dearth of prior knowledge concerning the behaviour of most of the populations under study. This is an obstacle for evaluating the quality of the results, as it rules out a comparison with well-known widely accepted models. To cope with this situation, three evaluation tasks were performed:

1. Turing Test-based validation tests, in which the effectiveness of the discovered knowledge was compared against the expert using the KBS. This task was planned to get a neat idea of the strength of the results when applied in everyday practice.
2. Continuous daily evaluation with real-life cases. This is a corrective stage and will continue throughout the research project life cycle.
3. Evaluation of satisfaction. This is an important part of any applied KD process evaluation, which is often overlooked. Its goal is to gain an understanding of the feelings of the practitioners when the new technology derived from the DM is transferred to everyday practice. The information obtained here is indispensable for defining new tasks or research lines and for getting accurate data about the potential of technology of this sort within the target domain.

The Turing test step of the evaluation process was to use the KBS to compare the effectiveness of the discovered knowledge against the expert's.

The information supplied to both the system and the experts was exactly the same (an isokinetic test) in order to assure that the test was meaningful. This test was repeated for 25 occurrences in the injury problem (15 with common injuries, 5 with rare injuries and 5 without injuries 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 2. The expert identified the five cases without injuries, 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). Meanwhile, the expert found 2 normal patterns, 2 injury patterns and could not make a decision on the other, and the novice found two injuries and was unable to make a decision on the others. It is important to note that the experts were unable to use any information apart from the isokinetic data

Regarding the daily evaluation with real-life cases, so far no significant errors have been reported by the physicians.

Finally, regarding the user feelings about the new technology and user satisfaction, members of the centre's staff claim that this system has improved the work of physicians working in the field of isokinetics and list the system's prominent features as follows:

- Physicians who are not specialists in isokinetics can use and exploit the system, thanks to the help provided by the intelligent interpretation, patterns and models.

- Analysis of the full isokinetic strength curve is possible, from which the complex or specific strength parameters that are of use for interpreting the tests can be inferred more correctly and completely. The isokinetic system features are better exploited thanks to the automatic extraction of improved information concerning injuries.
- Furthermore, the I4 system will provide more knowledge of the characteristics of athletes' strength, which has implications for the development and evaluation of training and rehabilitation programs. These issues make it highly relevant in the field of top-competition sport.

5. CONCLUSIONS

The development of an ES and its later refinement is mainly based on eliciting and entering experts' present knowledge of the subject into the system. At later stages, more domain knowledge of the problem was added to the system on the basis of the experience gained by the experts in the meantime.

However, all this knowledge was not able to cope with every problem we faced. Thanks to data mining techniques, a more efficient and objective process can be applied that is complementary to the above for developing an ES, provided enough data are available from which new knowledge can be discovered.

The I4 project, 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 in order to conduct an intelligent analysis of the numerical parameters and morphology of the strength curves produced by the isokinetic tests. Later and considering the volume of tests there were, data mining techniques were applied, based on time series processing to discover patterns characterizing injuries. This new knowledge, evaluated and validated by the expert, was entered into the system and produced a system that performed better and was more efficient than the one directly elicited from the expert. This has meant that it has been able to be applied at the Spanish National Centre for Sports Research and at the School of Physiotherapy of the Spanish National Organisation for the Blind.

With regard to the application of a methodological KDD process for extracting the knowledge from the data, it is noteworthy that the data preparation was so domain dependent that we were unable to set out a general-purpose methodological approach that could be extrapolated to any other system.

The same conclusion is applicable to the selection and design of the DM techniques best suited to the problem. Domain dependency is total, with regard to both the data in question and the planned use to which they are to be put. Therefore, it is impossible to generalise a mechanism of selection of mining techniques.

6. ACKNOWLEDGMENTS

We would like to thank África López-Illescas, Fernando Alonso and Cesar Montes for their co-operation in the I4 project and useful comments, and Óscar Marbán, Agustin Santamaría y José Carlos-Roca for their implementation work. The I4 project was partly funded by CICYT project no. TIC98-0248-C02-01. This

paper was written in co-operation with CETTICO (Centre of Computing and Communications Technology Transfer).

7. REFERENCES

- [1] Agrawal, R., Faloutsos, C. and Swami, A., "Efficient similarity search in sequence databases," *Proc. Foundations of Data Organisations and Algorithms (FODO Conference)*, (1993).
- [2] Agrawal, R., and Strikant, R., "Mining sequential patterns," *Proc. 1995 Int. Conf. Very Large Data Bases*, 487-499, Santiago, Chile, (September 1994).
- [3] Agrawal, R., and Strikant, R., "Fast algorithms for mining association rules," *Proc. eleventh international conference on data engineering*, IEEE Computer Society Press, 3-14, Taipei, Taiwan, (1995).
- [4] Alonso Amo, F., Barreiro, J., Fuertes, J.L., Martínez, L. and Montes, C., "Incremental Prototyping Approach to Software Development in Knowledge Engineering," *Proc. 5th International Conference on Information Systems Analysis and Systemics*, Orlando, (1998a).
- [5] Alonso, F., Barreiro, JM., Valente, JP. and Montes, C., "Interpretation of Strength Data," *Proc. 11th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, LNAI no. 1415, vol. I, Benicassim, Spain, (1998).
- [6] Alonso, F., Caraça-Valente, JP., López-Chavarrías, I., and Montes, C., "Knowledge discovery in time series using expert knowledge," *Medical Data Mining And Knowledge Discovery*, Springer-Verlag, (to appear in 2000).
- [7] Beckman, N., Kriegel, H.-P., Schneider, R. and Seeger, B., "The R*-tree: an efficient and robust method for points and rectangles," *ACM SIGMOD*, 322-331, Atlantic City, NJ, (1990).
- [8] Bin Xia, B., "Similarity search in time series data sets," Thesis submitted for the degree of Master of Science, Simon Fraser University, Canada, (1997).
- [9] Berndt, D. and Clifford, J., "Finding patterns in time series in advances," *U. Fayyad et al. (eds.) Knowledge Discovery and Data Mining*, MIT Press, (1996).
- [10] Brachman, R., and Anand, T., "The process of knowledge discovery in databases: A human centered approach," *AKDDM*, MIT Press, 37-59, (1996).
- [11] Caraça-Valente, JP, López-Chavarrías, I., and Montes, C., "Functions, rules and models: three complementary techniques for analyzing strength data," *Proc. of the symposium of the ACM for Applied Computing*, (2000).
- [12] Faloutsos, C., Ranganathan, M., and Manolopoulos, Y., "Fast subsequence matching in time series databases," *Proc. Third International Conference on*

- Knowledge Discovery and Data Mining*, Newport Beach, California. AAAI Press, Menlo Park, California, 24-30, (1994).
- [13] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthursamy, R., *Advances in knowledge discovery and data mining*, AAAI Press, Menlo Park, California, (1996).
- [14] Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P., "Knowledge Discovery and Data Mining: Towards a Unifying Framework," *Proceedings of KDD-96*, AAAI Press, 82-88, (1996).
- [15] García, C. and Zanoletty, D., "ISOCIN. Sistema experto para la interpretación de isocinéticos," *Technical report*, Facultad de Informática de Madrid, Madrid, Spain, (1998).
- [16] Gleeson, N.P., and Mercer, T. H., "The utility of isokinetic dynamometry in the assessment of human muscle function," *Sports Medicine*, 21(1), (1996).
- [17] Gómez, A., Juristo, N. and Pazos, J., *Ingeniería del conocimiento. Diseño y construcción de sistemas expertos*, Ceura, Madrid, Spain, (1997).
- [18] Gupta, U., *Validating and Verifying Knowledge Based Systems*, IEEE Computer Society Press, (1991).
- [19] Han, J., Dong, G., and Yin Y., "Efficient mining of partial periodic patterns in time series database," *Proc. Fourth International Conference on Knowledge Discovery and Data Mining*, AAAI Press, Menlo Park, 214-218. (1998).
- [20] Kannus, P., "Isokinetic evaluation of muscular performance: implications for muscle testing and rehabilitation," *Int. J. Sports Medicine* 15(Suppl 1), (1994).
- [21] López-Illescas, A., "Estudio del balance muscular con técnicas isocinéticas," *Proc. I Curso de Avances en Rehabilitación*, Seville, Spain, (1993).
- [22] Mannila H., Toivonen, H., and Verkamo, A., "Discovering frequent episodes in sequences," *Proc. 1st Int. Conf. Knowledge Discovery and Data Mining*, 210-215, Montreal, Canada, (1995).
- [23] Polikar, R., "The engineers' ultimate guide to wavelet analysis- the wavelet tutorial", <http://www.public.iastate.edu/~rpolikar/WAVELETS/Wttutorial.html>, (1996)
- [24] Povinelli, R., "Times series data mining: identifying temporal patterns for characterization and prediction of time series," Thesis. Milwaukee, (1999).