Computational Intelligence, Volume 22, Number 3/4, 2006

CASE-BASED REASONING AND KNOWLEDGE DISCOVERY IN MEDICAL APPLICATIONS WITH TIME SERIES

PETER FUNK AND NING XIONG

Department of Computer Science and Electronics, Mälardalen University, SE 721 23 Västerås, Sweden

This paper discusses the role and integration of knowledge discovery (KD) in case-based reasoning (CBR) systems. The general view is that KD is complementary to the task of knowledge retaining and it can be treated as a separate process outside the traditional CBR cycle. Unlike knowledge retaining that is mostly related to case-specific experience, KD aims at the elicitation of new knowledge that is more general and valuable for improving the different CBR substeps. KD for CBR is exemplified by a real application scenario in medicine in which time series of patterns are to be analyzed and classified. As single pattern cannot convey sufficient information in the application, sequences of patterns are more adequate. Hence it is advantageous if sequences of patterns and their co-occurrence with categories can be discovered. Evaluation with cases containing series classified into a number of categories and injected with indicator sequences shows that the approach is able to identify these key sequences. In a clinical application and a case library that is representative of the real world, these key sequences would improve the classification ability and may spawn clinical research to explain the co-occurrence between certain sequences and classes.

Key words: case-based reasoning, time series, knowledge discovery, medical applications, Bayesian theorem, retrieve knowledge.

1. INTRODUCTION

Clinicians use both explicit knowledge obtained from guidelines and regulations, and implicit knowledge based on their own experience and that of the patients and other clinicians, and experience in the form of past cases (Montani and Bellazzi 2001). Implicit knowledge may be knowledge that is known by many or all clinicians. But because medical case libraries also include case outcomes, they may be a valuable source of implicit knowledge not previously recognized by clinicians. A large proportion of medical research is directed toward discovering new co-occurrences (e.g., von Schéele 1999).

The main contribution of knowledge discovery (KD) is that it makes possible the recognition of previously unknown and potentially useful information. It may be defined as the nontrivial process of identifying novel, valid, and potentially useful data patterns, and ideally, to also understand these data patterns (Fayyad, Piatetsky-Shapiro, and Smyth 1996). KD is related to machine learning, statistics, databases, and data visualization, and uses a variety of techniques such as statistical techniques, decision trees, decision nets, clustering techniques, and neural nets. Some case-based reasoning (CBR) systems contain a proportionately large part that could be labeled as KD, others contain no KD and the learning process is purely based on new experience in the form of new cases stored in the case library during case retaining. Typically the retain step adds a new case or may modify some existing cases in the case library and usually contains a number of substeps of which the learning of knowledge based on the new case is one substep (Aamodt and Plaza 1994).

A limitation of learning included in the CBR cycle is that it is always associated with a specific case newly solved and thereby fails to discover more general domain knowledge relevant for performing CBR tasks. The general domain knowledge in support of various CBR steps can be classified into five different knowledge containers (Richter 1995), which are outside the scope of the case library as shown in Figure 1. It follows that we need to introduce an independent module to elicit knowledge for these knowledge containers apart from learning by means of case retaining. The KD module depicted in Figure 1 serves this purpose and it is considered as a separate background task outside the CBR cycle. Any existing knowledge including the case library constitutes the input source for KD and the new knowledge discovered is then delivered back to knowledge containers as well as the case

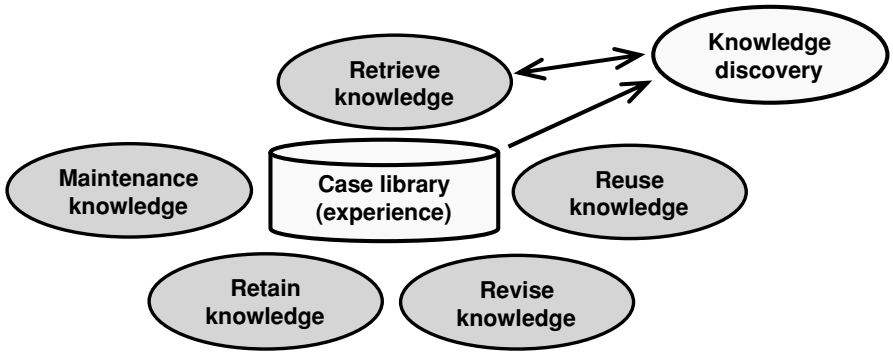

FIGURE 1. Case library, knowledge containers, and the knowledge discovery module.

library (e.g., as prototype or stereotype cases). Such bidirectional information exchange is very beneficial in giving birth to an integration that, on one hand, facilitates knowledge acquisition and learning, and on the other hand, makes new knowledge available to enhance some steps in the CBR cycle.

Actually, the idea of a separate KD module came from our studies of medical CBR applications with time series. Because the major interest therein lies on the transition of behavior over time rather than on the absolute values in the sequence, original time series data fail to capture the nature of the problem and thus cannot be utilized directly in case matching and retrieval. Acquisition of structured knowledge therefore becomes imperative to better characterize the time-series cases and to perform meaningful similarity matching for retrieval of relevant cases. We focus on discovering knowledge for case retrieval in the context here and that is why only one knowledge container (retrieve knowledge) in Figure 1 exhibits connections with the KD module. But in a general sense any knowledge container can be coupled with KD and receives newly acquired knowledge.

The aim of this article is to advocate an independent KD process to be integrated with CBR and demonstrate the significance of doing so, especially in light of common medical applications with time-series data. The role of KD as supplementary to case retaining is further discussed in Section 2. Followed in Section 3, a concrete medical scenario is tackled in detail, where indicative sequences for judging patient stress categories are identified from cases of series of breath patterns. The identified indicative sequences can be utilized as key features for depicting original data series and also provide valuable information for similarity matching and case retrieval.

Cases containing time series have been explored in medical applications in particular and also in some industrial cases. The value of CBR in medical applications has also been investigated and confirmed in a number of research projects. Successful CBR applications in medical classifications include case-based object recognition (Perner and Bühring 2004); the Auguste project (Marling and Whitehouse 2001); the CARE-PARTNER system (Bichindaritz, Kansu, and Sullivan 1998); the MNAOMIA system, a CBR system able to create hypotheses in the area of eating disorders (Bichindaritz 1995; Bichindaritz 1996), and Schmidt and Gierl's unnamed system for time-series analysis and prediction of kidney function (Schmidt and Gierl 2001). For a more extensive overview of state-of-the-art of medical CBR systems (see Nilsson and Sollenborn 2004).

2. KNOWLEDGE DISCOVERY AS COMPLEMENTARY TO KNOWLEDGE RETAINING

As stated in the preceding section, experiences gained by CBR itself are always associated with specific cases even if generalized in the form of classes and prototypes. We thereby take

the approach of treating KD as a separate process, integrated into and extending the CBR cycle. In the CBR systems we have studied, it is easy to distinguish between learning based on new experience (new cases and their indexing) and KD. Our extended definition of KD in CBR systems is that it is learning that is not naturally associated with a new specific case. While the CBR cycle is often directly triggered by a new problem, the KD process is often naturally suitable as a separate background task. The case library as a whole and additional knowledge are the input to the KD process and the result is delivered back to the system. In an industrial CBR system diagnosing faults in industrial robots on the basis of sound patterns, the KD process was performed manually, discovering new features and general knowledge that is used in the system to improve matching results (Olsson, Funk, and Xiong 2004). The discovery of new features is now considered for automation in an off-line approach using all cases as the source for the discovery process. For the identification of breathing dysfunctions to diagnose stress, a CBR system is used to classify individual breathing cycles (Nilsson et al. 2006). The KD process in this application is too complex to be performed manually but experts in the field are convinced that KD would identify new and valuable co-occurrences in categories of patient's measurement data, and in particular in the time series of classified breathing patterns. Such new knowledge would increase the usefulness of the CBR system in the diagnosis of patients and would encourage research as experts can use the new knowledge to improve and refine existing models.

Case-based reasoning uses domain knowledge to retrieve relevant cases from the case library. In complex applications, as many medical applications are, a large body of domain knowledge is often needed to enable the system to identify and retrieve the relevant cases.

If a system performs weakly in retrieving appropriate cases, either the retrieval knowledge is insufficient or the cases do not include all the essential features necessary to retrieve the most relevant cases. If the retrieval knowledge is insufficient, it may be necessary to optimize the weighting or if a more complex domain, optimize the similarity functions. If, however, essential features are missing or concealed in relations between other features, considerable effort may be needed to identify them. For some systems the separation of the KD makes the CBR system more transparent and reduces the complexity of research and implementation. It may even make it easier to apply CBR to applications for which it has not been suitable previously because of their complex nature.

A KD process integrated in the CBR cycle may also be advantageous because advancements in computing technology will enable more sophisticated approaches to the discovery of knowledge and the return of new general knowledge to the knowledge container. Figure 2 shows the input to and output from the KD process. The cases in the case library may stem

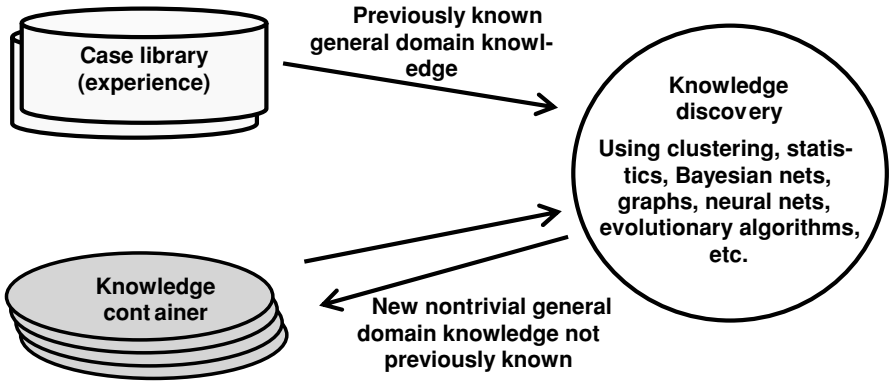

FIGURE 2. Knowledge discovery and its integration with CBR.

from different case bases. In the medical domain it may not normally be acceptable to distribute medical cases between hospitals without a medical reason, but it may be acceptable if they are to be used by a KD process to obtain new knowledge to be made available to all CBR systems of the same kind. In our example the retrieval knowledge is input to the KD process to determine if there are previously undiscovered co-occurrences. In a medical environment the knowledge containers of all the CBR systems may have the same content and share prototypical cases, but in general, they contain different cases. Moreover, it may also be desirable to validate and verify discovered knowledge automatically or manually, for example, if the system is able to reclassify all the cases in the case library as effectively as or more effectively than the system without the discovered knowledge, it may be argued that it is safe to add the discovered knowledge.

The example showing the integration of CBR with KD is from the medical domain, in which the classification of respiratory sinus arrhythmia (RSA) based on sensor readings is becoming increasingly important in physiological/psychophysiological medicine (von Schéele 1999). When clinicians diagnose patients, one important procedure is to classify RSA into twelve or more different patterns, a task being automated with a CBR system (Nilsson et al. 2006). The full patient case is information-rich and cases store the time series of classified breathing patterns, today used manually by clinicians in making a final diagnosis. Time series of breathing patterns are used in the diagnosis process because clinicians know that classification based on a single breathing cycle is not sufficient for the reliable diagnosis of a breathing dysfunction. Classifying complex time series manually is tedious and often requires long experience, with no explicit rules or guidelines available, in the worst case leading to incorrect diagnoses by less experienced clinicians. KD pertinent to CBR in this scenario will be addressed in the following section.

3. AN EXAMPLE OF KNOWLEDGE DISCOVERY IN TIME SERIES

As noted in Nilsson et al. (2006) classification of individual RSA patterns is one of the main procedures used by clinicians to classify RSA dysfunction. Clinicians also emphasize that classification based on a single RSA pattern is not sufficient for a reliable diagnosis of an RSA dysfunction. This is because RSA reflects the net effect of the complex interaction of the many different systems involved.

The pattern classification system (Nilsson et al. 2006) identifies dysfunctional RSA patterns directly from sensor readings. A clinical session usually lasts eighteen minutes divided into a number of different phases (e.g., normal breathing, provoking stress, breathing deeply, etc.), with an average of five to fifteen seconds per respiration period. These phases are handled as individual cases and each phase contains a series of classified RSA patterns (example in Subsection 3.1). A series of breaths contains on average forty to eighty respiration periods (inhalation–exhalation cycles). The pattern classification system eliminates most sensor noise, but there is still the possibility that there may be some misclassifications caused by distortions in sensor reading data. In the following we will only refer to series or breathing patterns and not phases.

The ability of experienced clinicians to identify RSA series is based on their experience. They are able to explain them once they recognize such a series, but the knowledge is not so explicit that they are able to describe such a series in advance. It should also be noted that the complexity of the systems reflected in RSA, their behavior, and their interaction are not fully understood and more theoretical work is needed (von Schéele 1999).

For this reason we have developed a method

- that permits the recognition of similarity of RSA sequences;
- that makes possible the identification of new RSA sequences of importance for the diagnosis of patients,
- that detects co-occurrence between RSA sequences and patient status, which may lead to the discovery of new co-occurrences as results of clinical experiments isolating the causal factors.

The method and relevant terminology are explained in the following subsections.

3.1. Important Sequences of RSA patterns

Assume that each dysfunctional RSA pattern is assigned a number between 1 and 9 (assuming for simplicity that there are only nine patterns, there being in reality more than ten different dysfunctional RSA patterns) and that one normal RSA pattern is assigned the number 0. A sequence of classified RSA patterns for a session can be illustrated by a series of successive integers with the length equal to the number of breathing cycles (usually 40–80, in the example below some parts of the series have been omitted)

RSA series: [000300000106000300024005000302003002000070000902000 0]. (1)

For a reliable diagnosis it would be advantageous to be able to identify recurring sequences of importance in the RSA series. Such sequences are exemplified in (2).

Significant sequences: [302], [3002], and [30002]. (2)

The sequences in (2) can be used by clinicians to detect RSA dysfunction, especially if they recur a number of times during a series. If the RSA patterns occur in this particular order (a RSA sequence) this may be a strong indication of a dysfunction, but if they (patterns "3" and "2") occurred in a different order or in a random order, then a clinician may not regard them as an indication of a dysfunction. Hence a way to automatically recognize recurring sequences of possible importance would be of value to clinicians.

3.2. Discovering New Important RSA Sequences in the Case Library

As can be seen from above, the important sequences recognized play a crucial role in the detection of dysfunctional RSA. A number of RSA sequences of importance in the diagnosis process are provided by clinicians with extensive experiences. But there may be RSA sequences not yet discovered by clinicians that may indicate RSA dysfunction. Experts in the field state that the discovery of new sequences is important for improving the reliability of the diagnosis process.

Discovery of RSA sequences potentially indicating dysfunction can be made by analyzing a large number of an RSA series from patients with known diagnosis. Once an RSA sequence occurs frequently in different series, a data-mining tool is able to discover a co-occurrence between the sequences and diagnosis. Such a tool may use clustering methods, statistics, and search techniques, and inspect all cases with this particular sequence, and identify any relation to a specific diagnosis. If such a relation is established, it can be used to aid clinicians in their diagnosis process. An experienced clinician may also be interested in using this discovery tool for discovering useful knowledge to accelerate the progress in clinical research.

Figure 3 shows a model that we are following to find new important sequences. The starting point (the top in Figure 3) is the problem space of all plausible sequences. A procedure with some domain knowledge including a priori known expressions may guide the

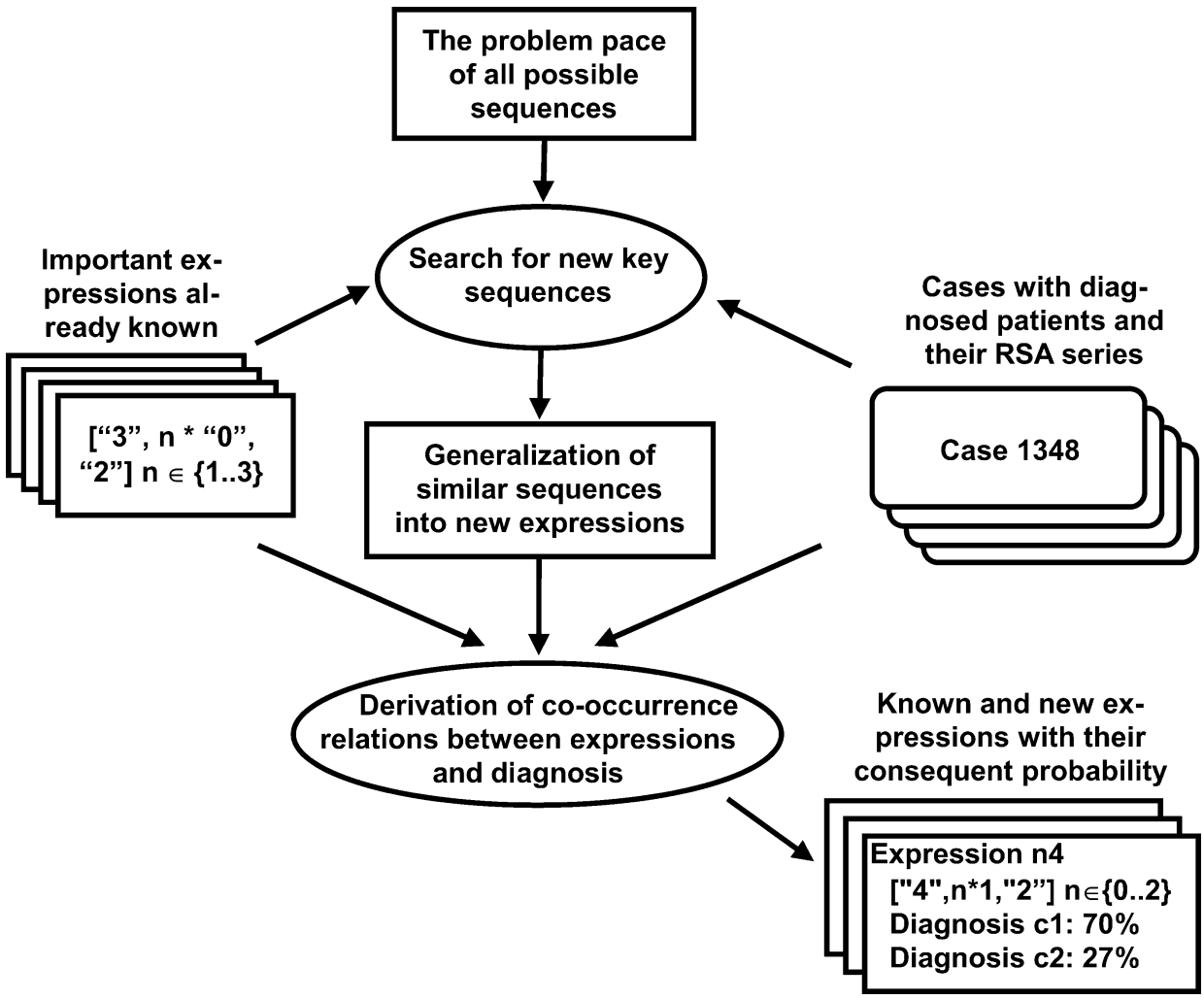

FIGURE 3. Discovery of new expressions and their co-occurrences.

generation of new sequences during the search process. The search is purported to find new key sequences (see Subsection 3.2.2) from all possible ones that are indicative in predicting a class in diagnosis. Each sequence generated is evaluated (see Subsection 3.2.1) in terms of the co-occurrence relationship between it and probable consequent classes based upon RSA series that belong to patients that have been diagnosed before. Clinicians classify patients in fourteen different classes. Because a sequence can appear in cases belonging to different patient classes all the cases containing the sequence have to be taken into account. The percentages of these cases in different classes then reflect the strength of the co-occurrence relationship of interest. Only those sequences exhibiting adequately strong co-occurrence (a threshold may be set by an expert in the field) are accepted as the goals of the search procedure. Following the search is the step of generalization, which intends to merge similar key sequences into expressions. A sequence is regarded as the simplest form of an expression. Creating expressions from sequences will be further explained in Subsection 3.4. Finally the co-occurrences of expressions are assessed with respect to consequent classes based upon the case library. The co-occurrences are also determined numerically for expressions already known because this will supplement more exact knowledge to the clinician's experience. The expressions shown in the right bottom corner in Figure 3 are described by the two most probable patient diagnoses and their corresponding probabilities.

3.2.1. Evaluation of a Sequence. Given a sequence s there may be a set of probable consequent classes $\{C_1, C_2, \ldots, C_k\}$. The strength of the co-occurrence between sequence s and class C_i $(i = 1, \ldots, k)$ can be measured by the probability, $p(C_i \mid s)$, of C_i conditioned upon s. Sequence s is considered as discriminative in predicting outcomes as long as it has strong co-occurrence with either of possible outcomes. The discriminating power of s is defined as the maximum of the strengths of its relations with probable consequents. Formally this definition of discriminating power PD is expressed as:

$$PD(s) = \max_{i=1\cdots k} P(C_i \mid s). \tag{3}$$

The conditional probabilities in (3) can be derived according to the Bayesian theorem as

$$P(C_i \mid s) = \frac{P(s \mid C_i)P(C_i)}{P(s)}. \tag{4}$$

As the probability $P(s)$ is generally obtainable by

$$P(s) = P(s \mid C_i)P(C_i) + P(s \mid \bar{C}_i)P(\bar{C}_i), \tag{5}$$

equation (4) for probability assessment can be rewritten as

$$P(C_i \mid s) = \frac{P(s \mid C_i)P(C_i)}{P(s \mid C_i)P(C_i) + P(s \mid \bar{C}_i)P(\bar{C}_i)}. \tag{6}$$

Our aim here is to yield the conditional probability $P(C_i \mid s)$ in terms of equation (6). As $P(C_i)$ is a priori probability of occurrence of C_i, which can be acquired or approximated from experiences in the domain, the only items that remain to be resolved are the probabilities of s in cases having class C_i and in cases not belonging to class C_i, respectively. Fortunately such probability values can be easily estimated by resorting to the case library. For instance we use the frequency of appearances of sequence s in class C_i samples as approximation of $P(s \mid C_i)$, thus we have

$$P(s \mid C_i) \approx \frac{N(C_i, s)}{N(C_i)}, \tag{7}$$

where $N(C_i)$ denotes the number of cases having class C_i in the case library and $N(C_i, s)$ is the number of cases having both class C_i and sequence s. Likewise the probability $P(s \mid \bar{C}_i)$ is approximated by

$$P(s \mid \bar{C}_i) \approx \frac{N(\bar{C}_i, s)}{N(\bar{C}_i)}, \tag{8}$$

with $N(\bar{C}_i)$ denoting the number of cases not having class C_i and $N(\bar{C}_i, s)$ being the number of cases containing sequence s but not belonging to class C_i.

The denominator in (6) has to stay far above zero to enable reliable probability assessment using the estimates in (7) and (8). Hence it is crucial to acquire an adequate amount of samples containing s in the case library. The more such cases available the more reliable could the derived probability assessment appear. For this reason we refer the quantity $N(s) = N(C_i, s) + N(\bar{C}_i, s)$ as evaluation base of sequence s in this paper.

Especially, should the prior probability $P(C_i)$ be assessed based on the appearance frequency of samples of C_i in the case library, equation (6) for calculation of the conditional probability is simplified to

$$P(C_i \mid s) \approx \frac{N(C_i, s)}{N(s)}. \tag{9}$$

Herein the evaluation base $N(s)$ clearly indicates the number of samples upon which the required conditional probability value is estimated in (9).

Two criteria have to be fulfilled for a sequence to be assessed as a key one upon case samples. The first is sufficient discriminating power to predict possible outcomes. The second is adequate evaluation base to ensure reliability of the probabilities assessed. In real applications it is up to related experts or clinicians to ultimately decide these thresholds that are highly domain dependent. The threshold for discriminating power may reflect the minimum probability value that suffices to predict a potential outcome in a specific domain. The threshold for the evaluation base indicates the minimum amount of samples required to fairly estimate the conditional probabilities of interest. Finally only those sequences that pass the thresholds for both evaluation base and discriminating power are recommended for being evaluated as key ones.

3.2.2. Search Algorithm for Finding Key Sequences. Finding the set of key sequences entails systematic exploration of a state space in which each state represents a sequence of patterns. Connection between two states signifies an operator between them for transition, i.e., addition or removal of a single pattern in time series. The search space for a scenario with three patterns is illustrated in Figure 4, where an arc connects two states if one can be created by extending the sequence of the other with a successive pattern.

The search starts from a null sequence and new sequences are created successively by adding a single pattern to parent nodes for expansion. The newly created nodes are then evaluated according to the evaluation bases and discriminating powers of the associated sequences. The results of evaluation determine the way to treat each child node in one of the following three situations:

i) If the evaluation base of the sequence is under a threshold required for conveying reliable probability assessment, terminate the expansion at this node. The reason is that the child nodes will have even smaller evaluation bases by appearing in fewer cases than their parent node;
ii) If the evaluation base and discriminating power are both above their respective thresholds, terminate the expansion at the node and store the state of this node as a key sequence. It is not necessary to continue on that node for expansion because all its child nodes merely contain redundant information with respect to the key sequence found;

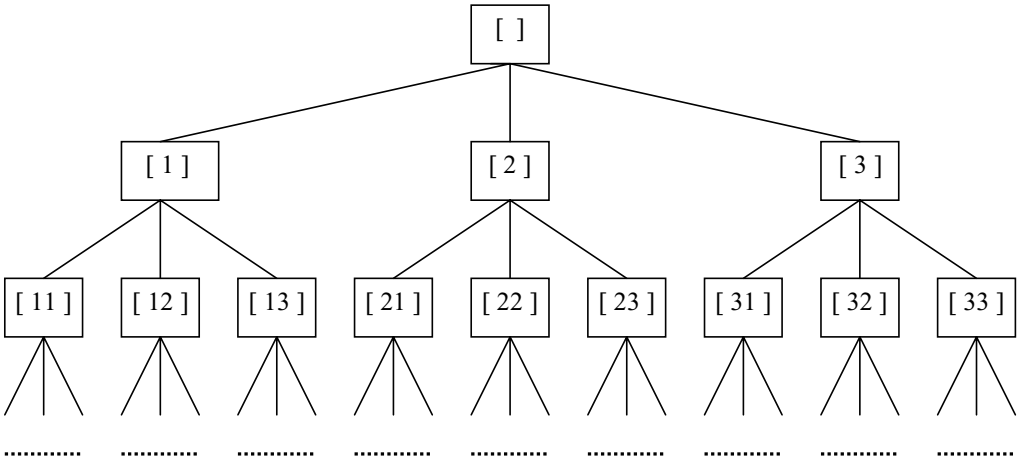

FIGURE 4. The state space for time series with three patterns.

iii) If the evaluation base is above its threshold whereas the discriminating power still not reaching the threshold, continue to expand this node with the hope of finding a satisfying sequence from its children.

The expansion of non-terminate nodes is proceeded in a level-by-level fashion. It is alike the breadth-first search in the sense that only when nodes at a current level have been expanded does the algorithm move on to the next level. On the other hand, in the context of this paper, we introduced certain special rules to deal with new nodes, which makes our search algorithm differ from the traditional breadth-first procedure in that (1) it does not attempt to expand every node encountered and a criterion is established to decide whether exploration needs to be proceeded at any given state; (2) it presumes multiple goals in the search space and thus the search procedure is not terminated when a single key sequence is found. Instead the search continues on other nodes until neither node from the current level needs to be further explored. A formal description of the proposed search algorithm is given as follows:

Algorithm for finding key sequences

1. Initialize the *Open* list with an empty sequence.
2. Remove the most left node t from the Open list.
3. Generate all child nodes of t.
4. For each child node, $C(t)$, of the parent node t
   (a) Evaluate $C(t)$ according to its discriminating power and evaluation base;
   (b) If the evaluation base and discriminating power are both above their respective thresholds, mark $C(t)$ as a key sequence and store it into the *Key_list;*
   (c) If the evaluation base of $C(t)$ is above its threshold but the discriminating power is not satisfying, put $C(t)$ on the right of the *Open* list.
5. If the *Open* list is not empty go to step 2, otherwise return the *Key_list* and terminate the search.

Finally it bears mentioning that the work of sequences discovery presented here differs from those (Agrawal and Srikant 1995; Srikant and Agrawal 1996; Garofalakis, Rastogi, and Shim 1999) in the literature of sequence mining. Usually the goal in sequence mining is merely to find all legal sequential patterns with their appearance frequencies above a user-specified threshold. However in our application context we have to consider the cause-outcome effect. Only the sequences exhibiting sufficient appearances and also strong discriminating power will be selected as the results of search.

3.3. Simulation Results on Sequence Discovery

To verify the feasibility of the mechanism addressed above we now present the simulation results on an artificially created case base. A case in this case base is depicted by a time series of 20 patterns and one diagnosis class as the outcome. A pattern in a time series belongs to $\{1, 2, 3, 4, 5\}$ and a diagnosis class is either A, B, or C. The four key sequences assumed are [1 4 3], [2 3 1], [4 5 2], and [5 1 5]. The first two sequences were supposed to have strong co-occurrences with class A and the third and fourth exhibit strong co-occurrence with classes B and C, respectively. Each case in the data set was created in such a way that both sequences [1 4 3] and [2 3 1] had a chance of 80% of being reproduced once in the time series of class A cases while sequences [4 5 2] and [5 1 5] were added into B and C cases, respectively, with a probability of 90%. After stochastic reproduction of key sequences the remaining patterns in the time series of all cases were generated randomly. The whole case

TABLE 1. Sequences Discovered with the Threshold of the Evaluation Base Set at 50

Sequence Discovered	Discriminating Power (%)	Evaluation Base	Main Consequent
[1 4 3]	76.70	103	Class *A*
[2 3 1]	78.22	101	Class *A*
[4 5 2]	73.39	124	Class *B*
[5 1 5]	83.18	107	Class *C*

TABLE 2. Sequences Discovered with the Threshold of the Evaluation Base Set at 30

Sequence Discovered	Discriminating Power (%)	Evaluation Base	Main Consequent
[1 4 3]	76.70	103	Class *A*
[2 3 1]	78.22	101	Class *A*
[4 5 2]	73.39	124	Class *B*
[5 1 5]	83.18	107	Class *C*
[3 1 4 3]	82.86	35	Class *A*
[3 4 5 2]	84.85	33	Class *B*

base consists of 100 instances for each class. The a priori probabilities of different classes were estimated based upon their appearance frequencies in this data set.

The search algorithm was applied to this case base to find key sequences and potential co-occurrences hidden in the data. The threshold for discriminating power was set at 70% to ensure an adequate strength of the relationship discovered. We also specified the values for the threshold of the evaluation base for reliable assessment of probabilities. Tables 1 and 2 illustrate the results from the tests where the threshold for the evaluation base was specified as 50 and 30, respectively.

As seen from Table 1 we detected all the four key sequences previously assumed. They were recognized to potentially cause the respective consequents with probabilities ranging from 73.39% to 83.18%. This relationship with a degree of uncertainty is due to the many randomly generated patterns in the case base such that any sequence of patterns is more or less probable to appear in any class cases. But such non-deterministic property is prevalent in many real-world domains, particularly in medical diagnosis situations.

By reducing the evaluation base threshold to 30 we obtained the results in Table 2, which consists of two more sequences: [3 1 4 3] and [3 4 5 2]. Yet this is not outside our expectation because the sequence [3 1 4 3] includes the key sequence [1 4 3] and [3 4 5 2] includes the key sequence [4 5 2]. Undoubtedly, a sequence containing a known key sequence is still discriminative in diagnosis. On the other hand, it bears noting that these two new sequences are actually redundant in conveying no more information. Redundant sequences can be easily identified by checking possible inclusion between sequences returned by the search procedure. Redundant sequences may also be avoided by increasing the evaluation base threshold for the search algorithm (as shown in Table 1) because a redundant sequence appears in fewer cases than the sequence it includes.

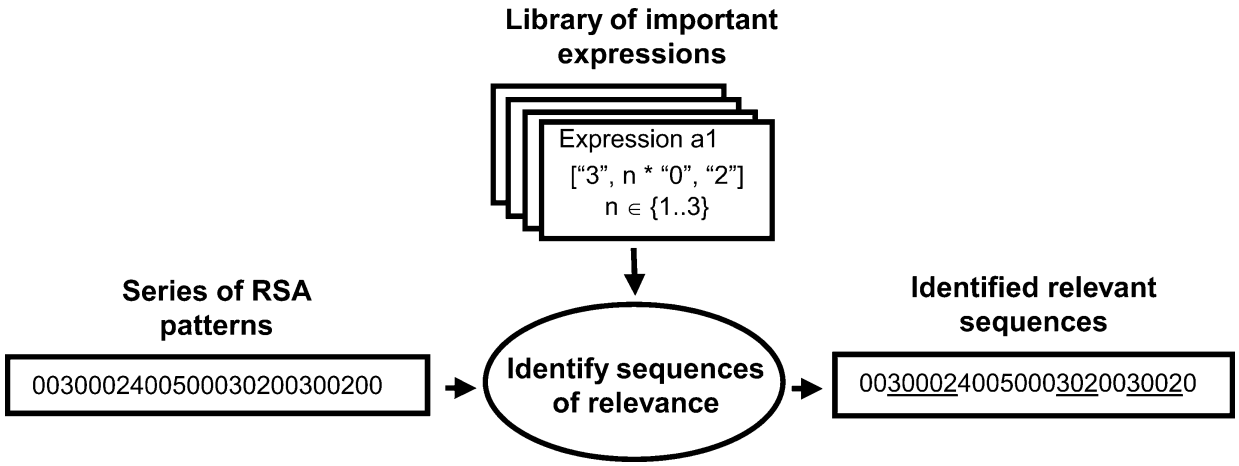

FIGURE 5. Identification of relevant sequences in a pattern series.

3.4. Applying Discovered Sequences to New Cases

Once a set of key sequences has been discovered upon the case library, they can be compared with each other for possible merging. Our solution is to generalize those sequences relating to the same class and with similar discriminating power into an expression enabling certain variations. For instance, if the sequences in (2) exhibit similar co-occurrences with the same RSA dysfunction, they are generalized into the expression ["3", n * "0", "2"] with $*$ denoting n time repetition of the following label and $n \in \{1, \ldots, 3\}$. This generalized expression means that there is first an RSA pattern 3 followed by one, two, or three normal breathing patterns and finally an RSA pattern of class 2, and that the variation of the number of normal breathing cycles in between plays no significant role. The aim of such generalization is to recognize similar sequences and enable identical interpretation of certain varying RSA sequences as done by clinicians. When a clinician is investigating a measurement session, a search for similar, but not exactly matching sequences may be relevant and hence a similarity-based matching is preferred. This may indicate a variation of an RSA dysfunction or even a new type of RSA dysfunction not previously encountered.

In Figure 5 a series of classified RSA patterns is given as input (from the left). This series is a result from a measurement signal classified by the HR3Modul, a tool for classification of RSA (Nilsson et al. 2006). The HR3Modul has classified each RSA pattern in the measurement signal. "0" is a normal breathing cycle with no indication of dysfunction. The library of expressions at the top can be seen as the output from Figure 3 and it contains expressions and/or sequences of importance for classifying dysfunctions. The expressions in this library may, as mentioned previously, stem from experienced experts, but may also contain formulations of sequences automatically generated as described in Subsection 3.2. The "identify sequences of relevance" process in the middle is the matching process, discovering sequences similar to those formalized in the expression library. In the resulting output series on the right, the identified sequences are underlined. Such identified relevant sequences can be considered as distinguishing features for depicting the case of the input pattern series.

The result will present the recognized sequences in the RSA pattern series to the clinician in a suitable way (sequences may overlap each other so how the sequences are visualized for the clinician must be chosen carefully). This will help less experienced clinicians in making

an overall diagnosis of the patient and also ease the work load on experienced clinicians. Moreover, as the discovered sequences have strong discriminating power, they will also be valuable in explaining diagnoses of clinicians in a similar way as that was done for single features in McSherry (1999, 2004).

4. UTILITY OF THE KNOWLEDGE ABOUT KEY SEQUENCES FOR INDEXING TIME-SERIES CASES

The discovered key sequences are treated as significant features in capturing dynamic system behaviors. Rather than enumerating what happened in every consecutive time segment, we can now more concisely represent a time-series case in terms of occurrences of key sequences in it. Let $\{S_1, S_2, \ldots, S_p\}$ be the set of key sequences. We have to search for every S_i $(i = 1, \ldots, P)$ in a time series X to detect all possible appearances. Then case index for X can be established according to the results of key sequence detection. In the following three alternate ways to index X based on key sequences are suggested.

4.1. Naïve Case Index

A naïve means of indexing a time-series case X is to depict it by a vector of binary numbers each of which corresponds to a key sequence. A number in the vector is unity if the corresponding sequence is detected in X and zero otherwise. This means that, by the naïve method, the index of X is given by

$$Id_1(X \mid S_1, \ldots, S_P) = [b_1, b_2, \ldots, b_P] \tag{10}$$

$$\text{where} \quad b_i = \begin{cases} 1 & \text{if } S_i \text{ is subsequence of } X \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

This index has the merit of imposing low demand in computation. It also enables the similarity between two cases to be calculated as the proportion of the positions where their indexing vectors have identical values. Suppose two time-series cases X_1 and X_2 which are indexed by binary vectors $[b_{11}, \ldots, b_{1P}]$ and $[b_{21}, \ldots, b_{2P}]$, respectively, the similarity between them is simply defined as

$$\text{Sim}_1(X_1, X_2) = 1 - \frac{1}{P} \sum_{j=1}^{P} |b_{ij} - b_{2j}|. \tag{12}$$

4.2. Case Index Using Sequence Appearance Numbers

With a binary structure the case index in Subsection 4.1 carries a little limited content and would be usable only in relatively simple circumstances. A main reason is that the index cannot reflect how many times a key sequence has appeared in a series of consideration. To incorporate that information, an alternate way is to directly employ the numbers of appearances of single key sequences in describing time-series cases. By doing this we acquire the second method of indexing time series X by an integer vector as

$$Id_2(X \mid S_1, \ldots, S_P) = [f_1, f_2, \ldots, f_P], \tag{13}$$

where f_i denotes the number of occurrences of sequence S_i in series X.

Further, considering the case index in (13) as a state vector, we use the cosine matching function (Salton 1968) as the similarity measure between two time-series cases X_1 and X_2. Thus we have

$$\mathrm{Sim}_2(X_1, X_2) = \frac{\sum_{j=1}^{P} f_{1j} f_{2j}}{\sqrt{\sum_{j=1}^{P} f_{1j}^2} \sqrt{\sum_{j=1}^{P} f_{2j}^2}}, \tag{14}$$

with f_{1j}, f_{2j} denoting the numbers of occurrences of key sequence S_j in X_1 and X_2, respectively.

4.3. Case Index in Terms of Discriminating Power

Although the case index in (13) can distinguish two cases having a same key sequence but with different numbers of appearances, it still might not be an optimal representation to capture the exact nature of the problem. Recall that the value of a key sequence is conveying a degree of confidence in the sense of discriminating power for predicting a potential consequent, a time series X would be more precisely characterized by the discriminating powers of the appearances of single key sequences. Intuitively two times of occurrences of a key sequence would give a stronger discriminating power than occurring just once, but not twice in the quantity of the strength. From view of this we suggest indexing X as a vector of real numbers, representing discriminating powers for the appearances of single key sequences, as follows:

$$Id_3(X \mid S_1, \ldots, S_P) = [g_1, g_2, \ldots, g_P], \tag{15}$$

$$\text{with} \quad g_i = \begin{cases} DP(f_i * S_i) & \text{if } f_i \geq 1 \\ 0 & \text{if } f_i = 0. \end{cases} \tag{16}$$

With $DP(f_i * S_i)$ we denote the discriminating power by sequence S_i appearing f_i times in X.

Let C be the class that the key sequence S_i is indicative of. We define the discriminating power $DP(f_i * S_i)$ as the probability for class C given f_i appearances of sequence S_i. This probability can be obtained by applying the Bayesian theorem in a sequential procedure. Assuming a two-class problem without loss of generality, this procedure is depicted here by a series of equations as follows:

$$P(C \mid S_i) = \frac{P(S_i \mid C)P(C)}{P(S_i \mid C)P(C) + P(S_i \mid \bar{C})P(\bar{C})}, \tag{17}$$

$$P(C \mid 2 * S_i) = \frac{P(S_i \mid C)P(C \mid S_i)}{P(S_i \mid C)P(C \mid S_i) + P(S_i \mid \bar{C})P(\bar{C} \mid S_i)}, \tag{18}$$

$$P(C \mid t * S_i) = \frac{P(S_i \mid C)P(C \mid (t-1) * S_i)}{P(S_i \mid C)P(C \mid (t-1) * S_i) + P(S_i \mid \bar{C})P(\bar{C} \mid (t-1) * S_i)}, \tag{19}$$

$$\begin{aligned} DP(f_i * S_i) &= P(C \mid f_i * S_i) \\ &= \frac{P(S_i \mid C)P(C \mid (f_i - 1) * S_i)}{P(S_i \mid C)P(C \mid (f_i - 1) * S_i) + P(S_i \mid \bar{C})P(\bar{C} \mid (f_i - 1) * S_i)}, \end{aligned} \tag{20}$$

where the probabilities $P(S_i \mid C)$ and $P(S_i \mid \bar{C})$ can be estimated according to equations (8) and (9), respectively. The probability updated in equation (17) represents the probability for class C given one appearance of S_i, which is further updated in equation (18) by the second appearance of S_i producing a higher probability considering both occurrences. Generally, the probability $P(C \mid t * S_i)$ is yielded by updating the prior probability $P(C \mid (t-1) * S_i)$ with one more occurrence of S_i in equation (19). Finally we obtain the ultimate probability assessment incorporating all appearances, i.e., the required discriminating power, by equation (20).

We now give a concrete example to illustrate how a case index can be built in terms of occurrences of key sequences. Suppose a two-class situation in which three key sequences S_1, S_2, and S_3 are discovered. Sequence S_1 appears twice in time series X and S_2 appears once while S_3 is not detected. S_1 and S_2 are both indicative of a certain class C. The a priori probability for class C is 50% and the probabilities of sequences S_1, S_2 in situations of class C and its complementary are shown below

$$P(S_1 \mid C) = 0.56 \quad P(S_1 \mid \bar{C}) = 0.24$$

$$P(S_2 \mid C) = 0.80 \quad P(S_2 \mid \bar{C}) = 0.40.$$

With all the information assumed above, the discriminating powers for the appearances of S_1 and S_2 are calculated in the following:

1. Calculate the probability for C with the first appearance of S_1 by

$$P(C \mid S_1) = \frac{P(S_1 \mid C)P(C)}{P(S_1 \mid C)P(C) + P(S_1 \mid \bar{C})P(\bar{C})} = \frac{0.56 \cdot 0.5}{0.56 \cdot 0.5 + 0.24 \cdot 0.5} = 0.70.$$

2. Refine the probability $P(C \mid S_1)$ with the second appearance of S_1, producing the discriminating power for the appearances of S_1

$$DP(2 * S_1) = P(C \mid 2 * S_1) = \frac{P(S_1 \mid C)P(C \mid S_1)}{P(S_1 \mid C)P(C \mid S_1) + P(S_1 \mid \bar{C})P(\bar{C} \mid S_1)}$$
$$= \frac{0.56 \cdot 0.70}{0.56 \cdot 0.70 + 0.24 \cdot 0.30} = 0.8448.$$

It is clearly seen here that the power of discrimination is increased from 0.70 to 0.8448 due to the key sequence occurring for the second time.

3. Derive the discriminating power for the occurrence of S_2 by calculating the conditional probability for C upon S_2 as

$$DP(1 * S_2) = P(C \mid S_2) = \frac{P(S_2 \mid C)P(C)}{P(S_2 \mid C)P(C) + P(S_2 \mid \bar{C})P(\bar{C})}$$
$$= \frac{0.80 \cdot 0.50}{0.80 \cdot 0.50 + 0.40 \cdot 0.50} = 0.6667.$$

Moreover, because S_3 is not detected in X, there is no discriminating power for it. Hence we construct the index for this time series case as

$$Id_3(X \mid S_1, S_2, S_3) = [0.8448, 0.6667, 0].$$

Finally, with this case-indexing scheme, we use the cosine function again as the similarity measure for case retrieval. So the similarity between two time series X_1 and X_2 is given by

$$\mathrm{Sim}_3(X_1, X_2) = \frac{\sum_{j=1}^{P} g_{1j} g_{2j}}{\sqrt{\sum_{j=1}^{P} g_{1j}^2} \sqrt{\sum_{j=1}^{P} g_{2j}^2}}, \tag{21}$$

where g_{1j} and g_{2j} denote the jth elements in the case indexes (15) for X_1 and X_2, respectively.

5. SUMMARY AND CONCLUSIONS

We have in this paper discussed the value of combining KD and case-based reasoning in medical applications in which time series and patterns of events in these time series are relevant. We propose to treat KD as a separate process, outside the traditional CBR cycle. In contrast to knowledge retaining which is directly related to case-specific experience, the purpose of KD is to discover new knowledge that is more general and, by adding this new knowledge to improve the overall performance of the CBR system.

The approach is exemplified in a medical domain (diagnosis of stress) in which the diagnosis is based on time series of classified breathing patterns. KD is used to discover key sequences in previously classified time series of breathing cycles. Single classified breathing patterns are not always sufficiently reliable for classification. New sequences that may have a causal correlation to specific diagnoses are generated and thereafter evaluated against all classified time series. If there is a correlation between the sequence and a particular diagnosis, the sequence is saved and used for improved classification of new unclassified series of breathing patterns. If the case library is representative of the real world, the key sequences can be used to improve the case-based reasoning systems ability to classify new problems. The identified key sequences may also have a value for experts who wish to explore and discover new causal relations not known previously. Hence the proposed approach also makes the case libraries valuable assets for clinical research. The suggested combination of KD and case-based reasoning is generic and will work for similar domain where time series contain information that improves classification and diagnosis.

REFERENCES

AAMODT, A., and E. PLAZA. 1994. Case-based reasoning: Foundational issues, methodological variations and system approaches. Artificial Intelligence Com, **7**:39–59.

AGRAWAL, R., and R. SRIKANT. 1995. Mining sequential patterns. *In* Proceedings of the 11th International Conference on Data Engineering, Taipei, Taiwan, pp. 3–14.

BICHINDARITZ, I., E. KANSU, and K. M. SULLIVAN. 1998. Case-based reasoning in care-partner: Gathering evidence for evidence-based medical practice. *In* Proceedings of the 4th European Workshop on Case-Based Reasoning, Dublin, Ireland, pp. 334–345.

BICHINDARITZ, I. 1996. MNAOMIA: Improving case-based reasoning for an application in psychiatry. *In* Artificial Intelligence in Medicine: Applications of Current Technologies, Stanford, CA, pp. 14–20.

BICHINDARITZ, I. 1995. Case-based reasoning adaptive to several cognitive tasks. *In* Proceedings of the 1st International Conference on Case-Based Reasoning, Sesimbra, Portugal, pp. 391–400.

FAYYAD, U., G. PIATETSKY-SHAPIRO, and P. SMYTH. 1996. From data mining to knowledge discovery. *In* Advances in Knowledge Discovery and Data Mining, MIT Press, Cambridge, MA, pp. 1–36.

GAROFALAKIS, M. N., R. RASTOGI, and K. SHIM. 1999. SPIRIT: Sequential pattern mining with regular expression constraints. *In* Proceedings of the 25th International Conference on Very Large Databases, Edinburgh, Scotland, pp. 223–234.

MARLING, C., and P. WHITEHOUSE. 2001. Case-based reasoning in the care of Alzheimer's disease patients. *In* Proceedings of the 4th International Conference on Case-Based Reasoning, Vancouver, Canada, pp. 702–715.

MCSHERRY, D. 1999. Dynamic and static approaches to clinical data mining. Artificial Intelligence in Medicine, **16**(1):97–115.

MCSHERRY, D. 2004. Explaining the pros and cons of conclusions in CBR. *In* Proceedings of the 7th European Conference on Case-Based Reasoning, Madrid, Spain, pp. 317–330.

MONTANI, S., and R. BELLAZZI. 2001. Intelligent knowledge retrieval for decision support in medical applications. *In* Proceedings of MEDINFO, IOS Press, London, pp. 498–502.

NILSSON, M., P. FUNK, E. OLSSON, B. VON SCHÉELE, and N. XIONG, 2006. Clinical decision support for diagnosing stress related disorders by applying psychophysiological medical knowledge to an instance based learning system. Artificial Intelligence in Medicine, **36**:159–176.

NILSSON, M., and M. SOLLENBORN. 2004. Advancements and trends in medical case-based reasoning: An overview of systems and system development. *In* Proceedings of the 17th International FLAIRS Conference, Miami Beach, FL, pp. 178–183.

OLSSON, E., P. FUNK, and N. XIONG. 2004. Fault diagnosis in industry using sensor readings and case-based reasoning. Journal of Intelligent & Fuzzy Systems, **15**:41–46.

PERNER, P., and A. BÜHRING. 2004. Case-based object recognition. *In* Proceedings of the 7th European Conference on Case-Based Reasoning, Madrid, Spain, pp. 375–388.

RICHTER, M. 1995. The knowledge contained in similarity measures. *In* Proceedings of the 1st International Conference on Case-Based Reasoning, Sesimbra, Portugal.

SALTON, G. 1968. Automatic Information Organization and Retrieval. McGraw-Hill, New York.

SCHMIDT, R., and L. GIERL. 2001. Temporal abstractions and case-based reasoning for medical course data: Two prognostic applications. *In* Machine Learning and Data Mining in Pattern Recognition; Lecture Notes in Computer Science, Vol. 2123. Spring-Verlag, New York, pp. 23– 34.

SRIKANT, R., and R. AGRAWAL. 1996. Mining sequential patterns: Generalizations and performance improvements. *In* Proceedings of the 5th International Conference on Extending Database Technology, Avignon, France, pp. 3–17.

VON SCHÉELE, B. 1999. Classification systems for RSA, ETCO2 and other physiological parameters. PBM Stressmedicine, Technical report, www.pbmstressmedicine.se.