Appl Intell (2008) 28: 247–260
DOI 10.1007/s10489-007-0059-x

Concise case indexing of time series in health care by means of key sequence discovery

Ning Xiong · Peter Funk

Published online: 16 June 2007

Abstract Coping with time series cases is becoming an important issue in applications of case based reasoning in medical cares. This paper develops a knowledge discovery approach to discovering significant sequences for depicting symbolic time series cases. The input is a case library containing time series cases consisting of consecutive discrete patterns. The proposed approach is able to find from the given case library all qualified sequences that are non-redundant and indicative. A sequence as such is termed as a key sequence. It is shown that the key sequences discovered are highly valuable in case characterization to capture important properties while ignoring random trivialities. The main idea is to transform an original (lengthy) time series into a more concise representation in terms of the detected occurrences of key sequences. Four alternative ways to develop case indexes based on key sequences are suggested and discussed in detail. These indexes are simply vectors of numbers that are easily usable when matching two time series cases for case retrieval. Preliminary experiment results have revealed that such case indexes utilizing key sequence information result in substantial performance improvement for the underlying case-based reasoning system.

Keywords Time series · Key sequence discovery · Case indexing · Case-based reasoning

N. Xiong (✉) · P. Funk
Department of Computer Science and Electronics, Mälardalen University, 72123 Vasteras, Sweden
e-mail: ning.xiong@mdh.se

P. Funk
e-mail: peter.funk@mdh.se

1 Introduction

Coping with time series in case based reasoning has become increasingly important in the medical domain but also in industrial applications. Many medical domains also exhibit dynamic properties. Unlike static cases where objects are described by attributes which are time independent, a time series case contains profiles of time-varying variables wherein pieces of data are associated with a timestamp and are meaningful only for a specific segment in the case duration. Temporal aspect of time series cases has to be taken into account in the tasks of case indexing and case retrieval. Abstraction and representation of temporal knowledge in CBR systems were discussed in [5, 8, 18].

Signal analysis techniques have been applied to extract relevant features from time series signals such as sensor readings. The most common methods used in applications are Discrete Fourier Transform and Wavelet Analysis, see [6, 14, 15, 22]. Both have the merit of capturing significant characteristics of the original signal with a compact representation, and the features extracted are directly usable in building similarity measures for case matching and retrieval. However the available signal processing techniques are inherently restricted to dealing with numerical values, they are not applicable to time series consisting of non-ordered discrete symbols.

This paper aims to extract useful sequences for depicting symbolic time series cases. As behaviors in dynamic processes are usually reflected from transitional patterns over time, occurrences of certain sequences are believed to be significant evidences to identify properties existing in historical sequential records. Deciding which sequences as characteristic while others as trivial in characterization of time series cases is largely domain dependent. Knowledge acquisition and discovery thus becomes imperative in circumstances when no prior knowledge is available.

The study presented is relevant to many medical health care applications where physicians have to investigate sequences of symptoms of patients before making clinical diagnoses, and where frequently changing conditions with patients are more important than their static states within single time segments. In particular this work is motivated by our ongoing project in diagnosis and treatment of stress where stress levels have to be estimated based on series of dysfunctional breathing patterns. Related medical research has revealed that certain transitions of breathing patterns over time may have high co-occurrence with stress levels of interest [19]. An outline of this application scenario and the problem to be addressed will be formulated in the next section.

We developed a knowledge discovery approach to sequence extraction employing a case base as the information source. The utilized case base is assumed to be symbolic and contains a collection of time series cases consisting of consecutive discrete patterns. The proposed approach is able to find from the given case library all those sequences that are non-redundant and indicative in having strong occurrences with a certain class. A sequence as such is termed as a key sequence. We show that the knowledge about key sequences is highly valuable in case characterization to capture important properties while ignoring randomly occurred trivialities in a dynamic process. Four alternative ways to index time series cases according to the set of discovered key sequences are suggested and discussed in detail. These indexes are simply vectors of numbers that are easily usable when matching two time series cases for case retrieval. Our preliminary experiments have shown the merit of such case indexes for enhancing CBR performance.

It is worth noting that the knowledge discovery treated here distinguishes itself from traditional learning included in a CBR cycle. The retain step in CBR typically stores a new case in the library or modifies some existing cases and may contain a number of sub-steps [1]. Learning therein is therefore case specific with knowledge stemming directly from newly solved cases. Contrarily, in our approach, learning is treated as a background task separated from the retain step and the whole case library is the input to the knowledge discovery process. Some relevant works combining knowledge discovery and CBR systems include: genetic-based knowledge acquisition for case indexing and matching [9], incremental learning to organize a case base [16], exploitation of background knowledge in text classification [23], and analysis of pros and cons for explanations in CBR systems [11].

The remainder of the paper is organized as follows. Section 2 briefly outlines a medical scenario motivating our research and also formulates the problem to be addressed. In Sect. 3 criteria to evaluate sequences are established, followed by presentation of the key sequence search algorithm in Sect. 4. How to index time series cases using discovered key sequences is addressed in Sect. 5. Then, in Sect. 6, we illustrate some results of experiments for discussion. Section 7 is devoted to related works and finally Sect. 8 ends the paper with concluding remarks.

2 A medical scenario and problem statements

In this section we first briefly outline a typical medical scenario in which patients' stress levels are to be determined based on a series of respiratory sinus arrhythmia (RSA) breathing patterns. After this some definitions are given and the formulation of the problem.

2.1 Classification via respiratory sinus arrhythmia

In stress medicine, RSA signals obtained from patients are typically employed to classify their stress levels. A patient is usually tested through a series of 40–80 breathing cycles (including inhalation and exhalation). Every respiration cycle lasts on average 5–15 seconds and corresponds to either a normal breathing pattern or one of the dysfunctional patterns. The patterns of breathing (also called RSA patterns) are identified from RSA measurements in the respective respiration periods. Further patterns from consecutive breathing cycles constitute a symbolic time series, which is to be investigated to find information reflecting stress levels of patients.

An overview of the stress medicine project is depicted in Fig. 1. First the RSA signal measured during the whole test period is decomposed into a collection of sub-signals. By sub-signal in Fig. 1 we denote the portion of the signal recorded for the ith cycle. Each sub-signal i is delivered to the block "signal classifier" to decide upon pattern i corresponding to it. The identified patterns are then composed into a symbolic series in terms of their appearance order in time. So far the part of signal classifier has been implemented in the previous work using wavelet analysis and case based reasoning [14]. The next step of the project is to further estimate the level of stress given a time series of respiration patterns. For applying CBR again in the second step we feel it necessary to acquire knowledge about key sequences to characterize and index time series cases.

2.2 Problem statements

To clearly present our work fitting into the scenario, we now give descriptions of the various terms and concepts that are related. We begin with the definitions about time series, sequences, and time series case bases, and then we precisely formulate the problem to tackle.

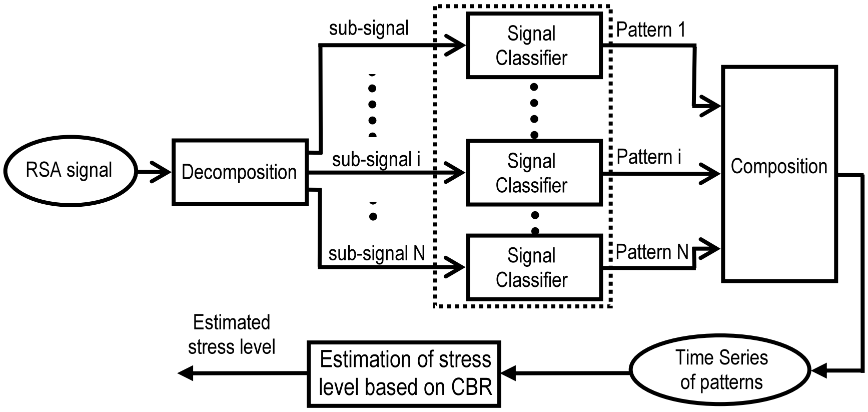

Fig. 1 An overview of the RSA based stress diagnosis system

Definition 1 A time series is a series of elements occurred sequentially over time, $X = \langle x(1), x(2), \ldots, x(i), \ldots, x(n)\rangle$, where i indexes the time segment corresponding to a recorded element and n can be very large.

The elements x in time series can be numerical or symbolic values. But in the discussions of this paper we restrict our attention to symbolic time series consisting of discrete patterns.

Moreover, every time series has an inherent class. The previous time series data are supposed to have been classified and they are stored in a case base together with their associated classes. A formal definition of time series case bases for purpose of classification is given as follows:

Definition 2 A time series case base is a set of pairs $\{(X_i, Z_i)\}_{i=1}^{K}$, where X_i denotes a time series and Z_i the class assigned to X_i and K is the number of time series cases in the case base.

With a time series case base at hand, the knowledge discovery process involves analyzing sequences that are included in the case base. A sequence in a time series is formally described in Definition 3.

Definition 3 A sequence S in a time series $X = \langle x(1), x(2), \ldots, x(n)\rangle$ is a list consisting of elements taken from contiguous positions of X, i.e., $S = \langle x(k), x(k+1), \ldots, x(k+m-1)\rangle$ with $m \le n$ and $1 \le k \le n - m + 1$.

Usually there is a very large amount of sequences included in the time series case base. But only a quite small part of them that carry useful information for estimating consequences are in line with our interest. Such sequences are referred to as indicative sequences and defined in the following:

Definition 4 A sequence is regarded as indicative given a time series case base provided that

(1) it appears in sufficient amount of time series cases of the case base;
(2) the discriminating power of it, assessed upon the case base, is above a specified threshold.

A measure for discriminating power together with the arguments that lie behind this definition will be elaborated in the next section. The intuitive explanation is that an indicative sequence is such a one that, on one hand, appears frequently in the case base, and on the other hand, exhibits high co-occurrence with a certain class.

Obviously, should a sequence be indicative, another sequence that contains it as subsequence may also be indicative for predicting the outcome. However, if these both are indicative of the same class, the second sequence is considered as redundant with respect to the first one because it conveys no more information. Redundant sequences can be easily recognized by checking possible inclusion between sequences encountered. The goal here is to find sequences that are not only indicative but also non-redundant and independent of each other.

Having given necessary notions and clarifications we can now formally define our problem to be addressed as follows:

Problem *Given a time series case base consisting of time series instances and associated classes, find a set of indicative sequences* $\{S_1, S_2, \ldots, S_p\}$ *that satisfy the following two criteria*:

(1) *For any* $i, j \in \{1, 2, \ldots, p\}$ *neither* $S_i \subseteq S_j$ *nor* $S_j \subseteq S_i$ *if* S_i *and* S_j *are indicative of a same class*;
(2) *For any sequence* S *that is indicative,* $S \in \{S_1, S_2, \ldots, S_p\}$ *if* S *is not redundant with respect to* S_j *for any* $j \in \{1, 2, \ldots, p\}$.

The first criterion above requests compactness of the set of sequences $\{S_1, S_2, \ldots, S_p\}$ in the sense that no sequence in it is redundant by having a subsequence indicative of the same class as it. A sequence that is both indicative and non-redundant is called a key sequence. The second criterion further requires that no single key sequence shall be lost, which signifies a demand for completeness of the set of key sequences to be discovered.

3 Evaluation of single sequences

This section aims to evaluate individual sequences to decide whether one sequence can be regarded as indicative. The main thread is to assess the discriminating power of sequences in terms of their co-occurrence relationship with possible time series classes. In addition we also illustrate the importance of sequence appearing frequencies in the case base for ensuring reliable assessments of the discriminating power.

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, $p(C_i|S)$, of C_i conditioned upon S. Sequence S is considered as discriminative in predicting outcomes as long as it has a strong co-occurrence with either of the possible outcomes. The discriminating power of S is defined as the maximum of the strengths of its relations with probable classes. Formally this definition of discriminating power PD is expressed as:

$$PD(S) = \max_{i=1,\ldots,k} P(C_i|S). \tag{1}$$

In addition we say that the class yielding the maximum strength of the co-occurrences, i.e.,

$$C = \underset{i=1,\ldots,k}{\arg} \max P(C_i|S),$$

is the class that sequence S is indicative of.

The conditional probabilities in (1) can be derived according to the Bayes theorem as:

$$P(C_i|S) = \frac{P(S|C_i)P(C_i)}{P(S)}. \tag{2}$$

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

$$P(S) = P(S|C_i)P(C_i) + P(S|\overline{C}_i)P(\overline{C}_i) \tag{3}$$

(2) for conditional probability assessment can be rewritten as

$$P(C_i|S) = \frac{P(S|C_i)P(C_i)}{P(S|C_i)P(C_i) + P(S|\overline{C}_i)P(\overline{C}_i)}. \tag{4}$$

Our aim here is to yield the conditional probability $P(C_i|S)$ in terms of (4). As $P(C_i)$ is a priori probability of occurrence of C_i which can be acquired from domain knowledge or approximated by experiences with randomly selected samples, the only things that remain to be resolved are the probabilities of S in (time series) 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 given case base. For instance we use the appearance frequency of sequence S in class C_i cases as an approximation of $P(S|C_i)$, thus we have:

$$P(S|C_i) \approx \frac{N(C_i, S)}{N(C_i)} \tag{5}$$

where $N(C_i)$ denotes the number of cases having class C_i in the case base and $N(C_i, S)$ is the number of cases having both class C_i and sequence S. Likewise the probability $P(S|\overline{C}_i)$ is approximated by

$$P(S|\overline{C}_i) \approx \frac{N(\overline{C}_i, S)}{N(\overline{C}_i)} \tag{6}$$

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

The denominator in (4) has to stay enough above zero to enable reliable probability assessment using the estimates in (5) and (6). Hence it is crucial to acquire an adequate amount of time series cases containing S in the case base. The more such cases available the more reliably the probability assessment could be derived. For this reason we refer the quantity $N(S) = N(C_i, S) + N(\overline{C}_i, S)$ as evaluation base of sequence S in this paper.

At this point we realize that two requirements have to be satisfied for believing a sequence to be indicative of a certain class. Firstly the sequence has to possess an adequate evaluation base by appearing in a sufficient amount of time series cases. Obviously a sequence that occurred randomly in few occasions is not convincing and can hardly be deemed significant. Secondly, the conditional probability of that class under the sequence must be dominatingly high, signifying a strong discriminating power. These explain why indicative sequence is defined by the demands on its appearance frequency and discriminating power in Definition 4.

In real applications two minimum thresholds need to be specified for the evaluation base and discriminating power respectively, to judge sequences as indicative or not. The values of these thresholds are domain dependent and are to be decided by human experts in the related area. The threshold for discriminating power may reflect the minimum probability value that suffices to predict a potential outcome in a specific scenario. The threshold for the evaluation base indicates the minimum amount of samples required to fairly approximate the conditional probabilities of interest. This threshold value can be estimated in terms of the distribution of cases of classes in the case library as well as their prior probabilities. It is shown in the following.

Let $\delta > 0$ be the smallest distance for the denominator in (4) to remain sufficiently away from zero, we demand

$$\frac{N(C_i, S)}{N(C_i)} P(C_i) + \frac{N(\overline{C}_i, S)}{N(\overline{C}_i)} P(\overline{C}_i) \geq \delta. \tag{7}$$

Further the above relation has to hold for every class C_i to ensure reliable assessments of conditional probabilities for all the classes given sequence S. Next the lower bound for the left side of inequality (7) is yielded by

$$\begin{aligned} &\frac{N(C_i, S)}{N(C_i)} P(C_i) + \frac{N(\overline{C}_i, S)}{N(\overline{C}_i)} P(\overline{C}_i) \\ &\geq \frac{N(C_i, S) P(C_i) + N(\overline{C}_i, S) P(\overline{C}_i)}{\max[N(C_i), N(\overline{C}_i)]} \\ &\geq \frac{[N(C_i, S) + N(\overline{C}_i, S)] \cdot \min[P(C_i), P(\overline{C}_i)]}{\max[N(C_i), N(\overline{C}_i)]} \\ &= \frac{\min[P(C_i), P(\overline{C}_i)]}{\max[N(C_i), N(\overline{C}_i)]} N(S). \end{aligned} \tag{8}$$

Since this lower bound not being less than δ is a sufficient condition for satisfaction of inequality (7), we simply impose constraints on the information base $N(S)$ as given by

$$N(S) \geq \frac{\max[N(C_i), N(\overline{C}_i)]}{\min[P(C_i), P(\overline{C}_i)]} \cdot \delta \quad \forall i. \tag{9}$$

Therefore it can be clearly seen that the threshold value for the information base can be defined as the minimum number of $N(S)$ that satisfies all the constraints in (9) for every class C_i. Finally only those sequences that pass thresholds for both discriminating power and information base are evaluated as indicative ones.

4 Discovering a complete set of key sequences

With the evaluation of sequences being established, we now turn to exploration of qualified sequences in the problem space. The goal is to locate all key sequences that are non-redundant and indicative. A sequence search algorithm for this purpose is detailed here. Later we will demonstrate simulation results on a synthetic case base with the proposed algorithm in Sect. 6.1.

Discovery of key sequences can be considered as a search problem in 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 sequences. The state space for a scenario with three patterns a, b, c is illustrated in Fig. 2, where an arc connects two states if one can be created by extending the sequence of the other with a following pattern.

A systematic exploration in the state space is entailed for finding a complete set of key sequences. We start from a null sequence and generate new sequences by adding a single pattern to parent nodes for expansion. The child sequences are evaluated according to evaluation bases and discriminating powers. 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 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, do the redundancy checking for the sequence against the list of key sequences already identified. The sequence is redundant if at least one known key sequence constitutes its subsequence while both remaining indicative of the same class. Otherwise the sequence is considered non-redundant and hence is stored into the list of key sequences together with the class it indicates. After that this node is further expanded with the hope of finding, among its children, qualified sequences that might be indicative of other classes.

(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 qualified sequences among its children.

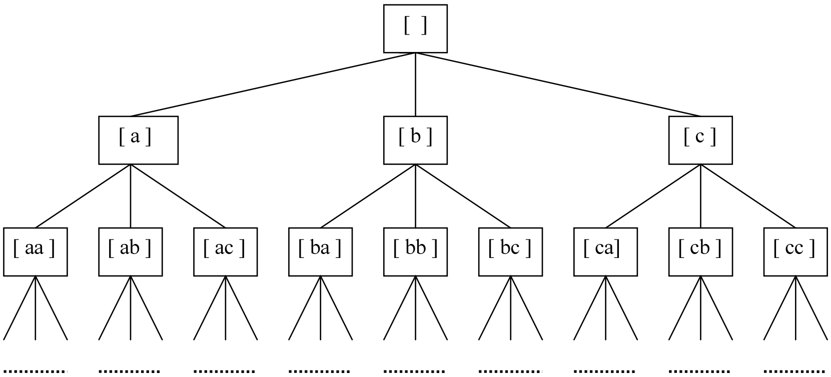

Fig. 2 The state space for sequences with three patterns

The expansion of non-terminate nodes is proceeded in a level-by-level fashion. A level in the search space consists of nodes for sequences of the same length and only when all nodes at a current level have been visited does the algorithm move on to the next level of sequences having one more pattern. This order of treating nodes is very beneficial for redundancy checking because a redundant sequence will always be encountered later than its subsequences including the key one(s) during the search procedure.

From a general structure, the proposed sequence search algorithm is a little similar to the traditional breadth-first procedure. However, there are still substantial differences between both. The features distinguishing our search algorithm are: (1) it does not attempt to expand every node encountered and criteria are 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 prospective nodes until none of the nodes in the latest level needs to be expanded. A formal description of the proposed search algorithm is given as follows:

Algorithm for finding a complete set of key sequences

```
1. Initialize the Open list with an empty sequence.
2. Initialize the Key_List to be an empty list.
3. Remove the most left node $t$ from the Open list.
4. Generate all child nodes of $t$.
5. 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, do the redundancy
       checking for $C(t)$ against the sequences in the
       Key_list. Store $C(t)$ into the Key_list if it is judged
       as not redundant. Finally put $C(t)$ on the right of the
       Open 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.
6. If the Open list is not empty go to step 3, otherwise return
   the Key_list and terminate the search.
```

5 Case indexing based on key sequences

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 four alternative ways to index X based on key sequences are suggested.

5.1 Naive case index

A naive 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 naive method, the index of X is given by

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

where

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

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

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

5.2 Case index using sequence appearance numbers

With a binary structure the case index in Sect. 5.1 carries a little limited content and would be usable only in relatively simple circumstances. A main reason is that the index can not reflect how many times a key sequence has appeared in a series of consideration. To incorporate that information, an alternative 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

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

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 [17] 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.

5.3 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 class, 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:

$$Id3(X|S_1, \ldots, S_P) = [g_1, g_2, \ldots, g_P] \tag{15}$$

with

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

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. Assuming the appearances of S_i are independent of each other, this probability can be obtained by applying the Bayes theorem in a sequential procedure. Considering a two class problem without loss of generality, this procedure is depicted here by a series of equations as follows:

$$P(C|S_i) = \frac{P(S_i|C)P(C)}{P(S_i|C)P(C) + P(S_i|\overline{C})P(\overline{C})}, \tag{17}$$

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

$$\vdots$$

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

$$\vdots$$

$$DP(f_i * S_i) = P(C|f_i * S_i) = \frac{P(S_i|C)P(C|(f_i-1) * S_i)}{P(S_i|C)P(C|(f_i-1) * S_i) + P(S_i|\overline{C})P(\overline{C}|(f_i-1) * S_i)} \tag{20}$$

where the probabilities $P(S_i|C)$ and $P(S_i|\overline{C})$ can be estimated according to (5) and (6) respectively. The probability updated in (17) represents the probability for class C given one appearance of S_i, which is further updated in (18) by the second appearance of S_i producing a higher probability considering both occurrences. Generally, the probability $P(C|t * S_i)$ is yielded by updating the prior probability $P(C|(t-1) * S_i)$ with one more occurrence of S_i in (19). Finally we obtain the ultimate probability assessment incorporating all appearances, i.e. the required discriminating power, by (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 (C_1 and C_2) 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 C_1. The a priori probability for class C_1 is 40% and the probabilities of sequences S_1, S_2 in situations of class C_1 and C_2 are shown below:

$$P(S_1|C_1) = 0.5, \qquad P(S_1|C_2) = 0.2,$$
$$P(S_2|C_1) = 0.8, \qquad P(S_2|C_2) = 0.3.$$

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_1 with the first appearance of S_1 by

$$P(C_1|S_1) = \frac{P(S_1|C_1)P(C_1)}{P(S_1|C_1)P(C_1) + P(S_1|C_2)P(C_2)}$$
$$= \frac{0.5 \cdot 0.4}{0.5 \cdot 0.4 + 0.2 \cdot 0.6} = 0.6250.$$

2. Refine the probability $P(C_1|S_1)$ with the second appearance of S_1, producing the discriminating power for the appearances of S_1

$$DP(2 * S_1) = P(C_1|2 * S_1)$$
$$= \frac{P(S_1|C_1)P(C_1|S_1)}{P(S_1|C_1)P(C_1|S_1) + P(S_1|C_2)P(C_2|S_1)}$$
$$= \frac{0.5 \cdot 0.625}{0.5 \cdot 0.625 + 0.2 \cdot 0.375} = 0.8065.$$

It is clearly seen here that the power of discrimination is increased from 0.6250 to 0.8065 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_2 upon S_2 as

$$DP(1 * S_2) = P(C_1|S_2)$$
$$= \frac{P(S_2|C_1)P(C_1)}{P(S_2|C_1)P(C_1) + P(S_2|C_2)P(C_2)}$$
$$= \frac{0.8 \cdot 0.4}{0.8 \cdot 0.4 + 0.3 \cdot 0.6} = 0.6400.$$

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:

$$Id_3(X|S_1, S_2, S_3) = [0.8065, 0.6400, 0].$$

With this case indexing scheme, we first calculate the dissimilarity between two time series X_1 and X_2 as the average of the differences in discriminating powers over all key sequences as follows:

$$Dis_3(X_1, X_2) = \frac{1}{P}\sum_{j=1}^{P} |g_{ij} - g_{2j}| \tag{21}$$

where g_{1j} and g_{2j} denote the jth elements in the case indexes (15) for X_1 and X_2 respectively. Since the dissimilarity measure in (21) is opposite to that of similarity, the degree of similarity between X_1 and X_2 is simply given by

$$Sim_3(X_1, X_2) = 1 - Dis_3(X_1, X_2). \tag{22}$$

5.4 Case indexing with key sequence union

In the preceding section cases are indexed according to the discriminating powers of occurrences of single key sequences. Such work could be extended by regarding the key sequences that are indicative of a common class as a collective union. This view motivates us to group occurrences of key sequences in time series X into a set of clusters. For every class C_i there is a cluster V_i corresponding to it. V_i is a collection of events for occurrences of those key sequences that are indicative of class C_i. The discriminating power of cluster V_i is defined as the probability of class C_i in light of the events included in the cluster. Hence we write

$$DP(V_i) = \begin{cases} P(C_i|\{e_j|e_j \in V_i\}) & \text{if } V_i \neq \emptyset, \\ 0 & \text{if } V_i = \emptyset. \end{cases} \tag{23}$$

Further, the discriminating powers of clusters of events representing key sequences occurrences are utilized to index a time series case. Hence the index for time series X is given by

$$Id_4(X|S_1, \ldots, S_P) = [DP(V_1), DP(V_2), \ldots, DP(V_K)] \tag{24}$$

where K denotes the number of classes of interest.

It is clear that the case index in the form of (24) is highly concise. It reduces the length of index vector to the number of classes. This is achieved by calculating the discriminating power for a union of key sequences that are consistent. Consequently every component in the vector of (24) contains rich information by fusion of occurrences from multiple key sequences. This proposed case index is valuable for further dimensionality reduction particularly under the circumstances when the number of key sequences discovered is still quite large.

Let $V_i = \{e_1, e_2, \ldots, e_T\}$ be a cluster of events of key sequences occurrences corresponding to class C_i. We now want to obtain the discriminating power of cluster V_i by calculating the conditional probability $P(C_i|e_1, e_2, \ldots, e_T)$. This probability is yielded by exploiting the events e_j as evidences for probability updating in separate steps. At every step we use a single event to revise prior probabilities according to the Bayes theorem and these updated probability estimates are then propagated as prior beliefs to the next step. The procedure of probability updating using events in cluster V_i is depicted by a series of equations as follows:

$$P(C_i|e_1) = \frac{P(e_1|C_i)P(C_i)}{P(e_1|C_i)P(C_i) + P(e_1|\overline{C}_i)P(\overline{C}_i)}, \tag{25}$$

$$P(C_i|e_1, e_2) = \frac{P(e_2|C_i)P(C_i|e_1)}{P(e_2|C_i)P(C_i|e_1) + P(e_2|\overline{C}_i)P(\overline{C}_i|e_1)}, \tag{26}$$

$$\vdots$$

$$P(C_i|e_1, \dots, e_i) = \frac{P(e_i|C_i)P(C_i|e_1, \dots, e_{i-1})}{P(e_i|C_i)P(C_i|e_1, \dots, e_{i-1}) + P(e_i|\overline{C}_i)P(\overline{C}_i|e_1, \dots, e_i)}, \tag{27}$$

$$\vdots$$

$$P(C_i|e_1, \dots, e_T) = \frac{P(e_T|C_i)P(C_i|e_1, \dots, e_{T-1})}{P(e_T|C_i)P(C_i|e_1, \dots, e_{T-1}) + P(e_T|\overline{C}_i)P(\overline{C}_i|e_1, \dots, e_{T-1})} \tag{28}$$

where the probabilities $P(e_i|C_i)$ and $P(e_i|\overline{C}_i)$ for $i \in \{1, \dots, T\}$ can be estimated according to (5) and (6) respectively, as e_i is considered as the occurrence of a sequence. The probability updated in (25) represents the probability for class C_i given event e_1, which is further updated in (26) by event e_2 producing a more refined belief considering both e_1 and e_2. Generally the probability $P(C|e_1, \dots, e_i)$ is yielded by updating the prior probability $P(C|e_1, \dots, e_{i-1})$ with a new event e_i in (27). Finally we obtain the ultimate probability assessment incorporating all available events by (28).

At this stage one may question the order in which single events from a cluster are used to refine probability assessments. This seems a fundamental issue and involves allocation of events to different steps of a sequential procedure. Fortunately our study has clarified that the order of events used in probability updating is completely indifferent. The final probability value remains constant as long as each piece of event is assigned to a distinct step. The claims as such are formally based on the following theorems.

Lemma *Let $\{e_1, \dots, e_T\}$ be a cluster of events representing appearances of certain key sequences in a time series X. The probability for class C given the cluster is not affected if two adjacent events exchange their positions in the order of events used for probability refinements. This means that the relation $P(C|e_1, \dots, e_i, e_{i+1}, \dots, e_T\} = P(C|e_1, \dots, e_{i+1}, e_i, \dots, e_T\}$ holds for $i \in \{1, \dots, T-1\}$.*

Proof For proof of the lemma with the statement that $P(C|e_1, \dots, e_{i-1}, e_i, e_{i+1}, \dots, e_T\} = P(C|e_1, \dots, e_{i-1}, e_{i+1}, e_i, \dots, e_T\}$, we only need to establish the relation for $P(C|e_1, \dots, e_{i-1}, e_i, e_{i+1}\} = P(C|e_1, \dots, e_{i-1}, e_{i+1}, e_i\}$, which is equivalent to the lemma.

We start to consider the probability $P(C|e_1, \dots, e_i, e_{i+1}\}$ which is acquired by updating the prior belief $P(C|e_1, \dots, e_i\}$ with a new evidence e_{i+1}, hence it can be written as

$$P(C|e_1, \dots, e_i, e_{i+1}) = \frac{P(e_(i+1)|C)P(C|e_1, \dots, e_i)}{P(e_{i+1}|C)P(C|e_1, \dots, e_i) + P(e_{i+1}|\overline{C})P(\overline{C}|e_1, \dots, e_i)}. \tag{29}$$

Further the probability $P(C|e_1, \dots, e_i)$ is formulated by taking $P(C|e_1, \dots, e_{i-1})$ as its prior estimate such that

$$P(C|e_1, \dots, e_i) = \frac{P(e_i|C)P(C|e_1, \dots, e_{i-1})}{P(e_i|e_1, \dots, e_{i-1}}. \tag{30}$$

Likewise we obtain

$$P(\overline{C}|e_1, \dots, e_i) = \frac{P(e_i|\overline{C})P(\overline{C}|e_1, \dots, e_{i-1})}{P(e_i|e_1, \dots, e_{i-1})}. \tag{31}$$

Combining (30) and (31) into (29) gives rise to a transformed formulation as

$$P(C|e_1,\dots,e_i,e_{i+1}) = \frac{P(e_{i+1}|C)P(e_i|C)P(C|e_1,\dots,e_{i-1})}{P(e_{i+1}|C)P(e_i|C)P(C|e_1,\dots,e_{i-1}) + P(e_{i+1}|\overline{C})P(e_i|\overline{C})P(\overline{C}|e_1,\dots,e_{i-1})}. \tag{32}$$

Next we express the conditional probabilities $P(e_{i+1}|C)$, $P(e_{i+1}|\overline{C})$, $P(e_i|C)$, $P(e_i|\overline{C})$ with their Bayes forms by

$$P(e_{i+1}|C) = \frac{P(C|e_{i+1})P(e_{i+1})}{P(C)}, \tag{33}$$

$$P(e_{i+1}|\overline{C}) = \frac{P(\overline{C}|e_{i+1})P(e_{i+1})}{P(\overline{C})}, \tag{34}$$

$$P(e_i|C) = \frac{P(C|e_i)P(e_i)}{P(C)}, \tag{35}$$

$$P(e_i|\overline{C}) = \frac{P(\overline{C}|e_i)P(e_i)}{P(\overline{C})} \tag{36}$$

where $P(C)$ and $P(\overline{C})$ denote the initial probability estimates for class C and its complementary without any events about key sequences appearances. Using the Bayes forms from (33) to (36), (32) is finally rewritten as

$$P(C|e_1,\dots,e_i,e_{i+1}) = \frac{P^2(\overline{C})P(C|e_{i+1})P(C|e_i)P(C|e_1,\dots,e_{i-1})}{P^2(\overline{C})P(C|e_{i+1})|P(C|e_i)P(C|e_1,\dots,e_{i-1}) + P^2(\overline{C})P(\overline{C}|e_{i+1})|P(\overline{C}|e_i)P(\overline{C}|e_1,\dots,e_{i-1})}. \tag{37}$$

Clearly we see from (37) that the order between e_i and e_{i+1} has no effect at all on the probability $P(C|e_1,\dots,e_i,e_{i+1})$ assessed. It follows that

$$\begin{aligned} &P(C|e_1,\dots,e_{i-1},e_i,e_{i+1}) \\ &\quad = P(C|e_1,\dots,e_{i-1},e_{i+1},e_i) \end{aligned} \tag{38}$$

and here from the lemma is proved. □

With the lemma justified by the proof above, we further contemplate the implication of it. This leads to a corollary presented below.

Corollary *Let $\{e_1,\dots,e_T\}$ be a cluster of events representing appearances of certain key sequences in a time series X. The probability for X in class C given the cluster is independent of the order according to which single events $e_{1,}e_{2,\dots,}e_T$, are used in probability refinements.*

The proof of the above corollary is obvious. According to the lemma, an element in a given order of events can be moved to an arbitrary position by repeatedly exchanging its position with an adjacent one while not affecting the final probability assessments. As this can be done to every piece of event, we enable transitions to any orders of events without altering the estimated value of the probability.

This corollary is important in providing theoretic arguments allowing for an arbitrary order of sequences to be used in probability fusion based on the Bayes theorem. The connotation is that when a key sequence occurred in the time series does not matter for the case index. Instead only the numbers of appearances of key sequences affect the likelihoods of classes given respective occurrence clusters, which are included as components in the case index vector.

Now let us study an illustrative example to better understand how the above sequential procedure works in derivation of required probabilities using clusters of events as evidences. Consider a time series X with two probable classes. Suppose that four key sequences S_1, S_2, S_3, and S_4 are detected in X, and S_1, S_2 are indicative of class C while S_3 and S_4 are indicative of the complementary of C. The a priori probability of class C is 50%, and the probabilities of sequences S_1, S_2, S_3, and S_4 in situations of class C and its complementary are shown below:

$P(S_1|C) = 0.56,$	$P(S_1|\overline{C}) = 0.24,$
$P(S_2|C) = 0.80,$	$P(S_2|\overline{C}) = 0.40,$
$P(S_3|C) = 0.35,$	$P(S_3|\tilde{C}) = 0.62,$
$P(S_4|C) = 0.18,$	$P(S_4|\tilde{C}) = 0.30.$

Further we assume that sequence S_1 appears twice in X and S_2, S_3, S_4 appear once, hence the clusters of key sequence occurrences for X are notated as $V_1(X) = \{S_1, S_1, S_2\}$ and $V_2(X) = \{S_3, S_4\}$. With these three occurrences detected, the probability of class C yielded in the following three steps:

Step A1: Update the a priori probability $P(C)$ with the first appearance of S_1 by

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

Step A2: Refine the probability updated in step A1 with the second appearance of S_1, thus we have

$$P(C|S_1, S_1) = \frac{P(S_1|C)P(C|S_1)}{P(S_1|C)P(C|S_1) + P(S_1|\overline{C})P(\overline{C}|S_1)} = \frac{0.56 \cdot 0.70}{0.56 \cdot 0.70 + 0.24 \cdot 0.30} = 0.8448.$$

Step A3: Refine the probability updated in step 2 with the occurrence of S_2, and we acquire the final probability assessment taking into account all events by

$$P(C|S_1, S_1, S_2) = \frac{P(S_2|C)P(C|S_1, S_1)}{P(S_2|C)P(C|S_1, S_1) + P(S_2|\overline{C})P(\overline{C}|S_1, S_1)} = \frac{0.80 \cdot 0.8448}{0.80 \cdot 0.8448 + 0.40 \cdot 0.1552} = 0.9159.$$

Likewise we calculate the probability $P(\overline{C}|S_3, S_4)$ with two steps as follows:

Step B1: Update the prior probability $P(\overline{C})$ with occurrence of S_3

$$P(\tilde{C}|S_3) = \frac{P(S_3|\tilde{C})P(\tilde{C})}{P(S_3|C)P(C) + P(S_3|\tilde{C})P(\tilde{C})} = \frac{0.62 \cdot 0.5}{0.35 \cdot 0.5 + 0.62 \cdot 0.5} = 0.6392.$$

Step B2: Refine the probability updated in step B1 with appearance of S_4

$$P(\tilde{C}|S_3, S_4) = \frac{P(S_4|\tilde{C})P(\tilde{C}|S_3)}{P(S_4|C)P(C|S_3) + P(S_4|\tilde{C})P(\tilde{C}|S_3)} = \frac{0.30 \cdot 0.6392}{0.18 \cdot 0.3608 + 0.30 \cdot 0.6392} = 0.7470.$$

Finally, with the required probabilities at hand, we can establish the case index for the time series X as follows

$$\begin{aligned} Id_4(X|S_1, S_2, S_3, S_4) &= [DP(V_1), DP(V_2)] \\ &= [P(C|S_1, S_1, S_2), P(C|S_3, S_4)] \\ &= [0.9159, 0.7470]. \end{aligned}$$

For similarity assessment, we first calculate the dissimilarity between two time series X_1 and X_2 as the average of the differences in discriminating powers over all key sequences clusters

$$Dis_4(X_1, X_2) = \frac{1}{K} \sum_{j=1}^{K} |DP(V_{1j}) - DP(V_{2j})| \tag{39}$$

where V_{1j} and V_{2j} denote the jth clusters of key sequences corresponding to class C_i, for X_1 and X_2 respectively. Since the concept of dissimilarity is opposite to that of similarity, the degree of similarity between X_1 and X_2 is simply defined as unity subtracted by the dissimilarity value

$$Sim_4(X_1, X_2) = 1 - Dis_4(X_1, X_2). \tag{40}$$

6 Experiment results

This section presents some experimental results to demonstrate the feasibility and usefulness of the proposed approaches. We first verify the ability of our search mechanism to find key sequences from a symbolic time series data set. Subsequently we examine the performance of case-based classification using these discovered key sequences.

6.1 Finding key sequences from time series data

A synthetic time series data set was created to test the feasibility of our key sequence search mechanism. A case in this data set is depicted by a time series of 60 patterns and one diagnosis class as the outcome. A pattern in a time series belongs to $\{a, b, c, d, e\}$ and a diagnosis class is either 1, 2, or 3. The four key sequences assumed are $[a\ c\ e\ b]$, $[d\ b\ a\ c]$, $[b\ c\ b\ e]$, and $[d\ d\ a\ e]$. The first two sequences were supposed to have strong co-occurrences with class 1 and the third and fourth exhibit strong co-occurrences with classes 2 and 3 respectively. Each time series in the data set was created in such a way as follows. The sequence $[a\ c\ e\ b]$ was reproduced once with the probability of 75% for cases of class 1, while the sequences $[d\ b\ a\ c]$ and $[b\ c\ b\ e]$ were created twice with the probability of 60% for both class 1 and class 2 cases. Moreover, with a chance of 50%, the sequence $[d\ d\ a\ e]$ was inserted three times into cases of class 3. After stochastic reproduction of these key sequences, the remaining patterns in the time series of all cases were generated randomly. The whole data set consists of 100 instances for each class. Presuming such time series cases to be randomly selected samples from a practical domain, a priori probability of each class is believed to be one third.

The sequence search algorithm was applied to this data set to find key sequences and potential co-occurrences hidden in the data. The threshold for the discriminating power was set at 70% to ensure adequate strengths of relationships discovered. For reliable assessment of probabilities, we also defined the threshold of the evaluation base according to (9) with δ being specified as 0.1. The sequences found in our test are shown in Table 1.

As seen from Table 1 we detected all the four key sequences previously assumed. They were recognized as potentially related to the respective classes with probabilities ranging from 83.51% to 87.50%. These relationships with a degree of uncertainty are due to the many randomly generated patterns in the data set such that any sequence of patterns is more or less probable to appear in time series of any class. But this would reflect non-deterministic property prevalent in many real world situations.

6.2 Case-based classification using key sequences

The next step is to utilize the information of key sequences to transform original symbolic time series into numerical feature vectors. Each of the case indexing schemes suggested in Sect. 5 can be used here for this purpose. The point of departure is that the data generated above are strongly characterized by some crucial transitions of patterns rather than single pattern values. As a consequence, it makes no sense to compare two time series cases in terms of the distance between them over the whole time span. Such judgment has also been verified by conducted tests in which the kNN method was applied on the original time series cases using the similarity metric as:

$$Similarity(X_1, X_2) = \frac{1}{K}\sum_{j=1}^{K}\begin{cases}1, & \text{if } X_1(j) = X_2(j),\\ 0, & \text{if } X_1(j) \neq X_2(j)\end{cases} \qquad (41)$$

where K is the length of the original symbolic time series and by $X_i(j)$ we denote the jth sequential pattern in time series X_i.

Table 1 Key sequences discovered on a synthetic data set

Key sequences discovered	Discriminating power	Evaluation base	Dominating class
[*a c e b*]	84.71%	85	Class 1
[*d b a c*]	83.51%	97	Class 1
[*b c b e*]	86.54%	104	Class 2
[*d d a e*]	87.50%	104	Class 3

The results of these tests using (41) as similarity assessment are shown in Table 2, which includes the leave-one-out accuracy of the kNN classifications with $k = 1, 3, 5, 7, 9, 11$. It is seen from the table that no improvement was achieved by *kNN* in classification accuracy over the prior probabilities of classes. The reason lies in the simple distance measurement applied in similarity matching, which appears knowledge poor and ignores all the information about occurrences of key sequences in time series cases.

To better characterize problems for CBR tasks, we converted the symbolic time series data according to occurrences of key sequences in our further experiments. The kNN method was then applied on the newly converted compact cases where a numerical vector was adopted as case index to convey descriptions of problems. All the four suggested case indexes (naive, sequence appearance number, discriminating power, sequence union) were investigated, leading to the employment of the similarity metrics in (12), (14), (22), and (40) respectively for case matching and retrieval. Table 3 illustrates the leave-one-out performances of the kNN classifications in association with different case indexes. Observing the results in Table 3 enables us to draw the statements as follows:

(1) The use of case indexes based on key sequences results in substantial improvement of classification accuracy in all cases against the situations with simple distances as similarity criterion.

(2) The case index in terms of discriminating powers seems to achieve the better classification performance than the naive index and the index using sequence appearance numbers, regardless the value of k specified for the kNN method. This probably can be explained by the more precise information carried by the discriminating powers compared with the binary or appearance number descriptions connected with key sequences.

(3) The case index upon key sequence unions causes almost the same performance as the index based on discrimi-

Table 2 The leave-one-out accuracy of kNN using the similarity metric in (41)

	$k = 1$	$k = 3$	$k = 5$	$k = 7$	$k = 9$	$k = 11$
Accuracy	34.00%	33.33%	31.00%	35.33%	31.33%	29.33%

Table 3 The leave-one-out accuracy of kNN with the case indexes based on key sequences

	$k = 1$	$k = 3$	$k = 5$	$k = 7$	$k = 9$	$k = 11$
Naive index	57.00%	85.67%	89.33%	90.00%	88.33%	86.33%
Index using appearance numbers	58.67%	89.67%	88.67%	89.33%	90.00%	86.33%
Index upon discriminating powers	73.33%	92.33%	91.33%	90.00%	91.67%	92.00%
Index upon key sequence unions	73.67%	92.33%	91.33%	89.00%	92.00%	91.00%

nating powers. The former index can be understood as compressing discriminating powers of key sequences of identical unions into single values. The merit of doing so is further reduction of the dimension of the case index, particularly when the number of key sequences is still large.

7 Related works

Representation and retrieval of time dependent situations has received increasing research efforts during the recent years. The two most common methods are Fourier and Wavelet transforms which aim to convert time-evolving profiles into somehow simplified and shorter vectors that still preserve core properties. The usages of Fourier and Wavelet transforms for retrieving similar cases to support medical and industrial diagnoses have been shown in [12, 14] and [15] respectively. Besides, the theory of temporary interval is also shown a suitable tool for representing temporal relations insides cases [8]. The idea is to maintain temporary information in a temporary network where nodes represent individual intervals and all possible relationships between nodes are processed by means of a predefined transitivity table. This temporal approach was applied in a CBR system Creek for prediction of unwanted events in oil well drilling [8].

A general framework for tackling cases in time dependent domain was proposed by [13], in which temporal knowledge embedded in cases are represented at two levels: case level and history level. The case level is tasked to depict single cases with features varying within case durations, while consecution of cases occurrences have to be captured in the history level to reflect the evolution of the system as a whole. It was also recommended by the authors that, at both of the two levels, the methodology of temporal abstraction [4, 20] could be exploited to derive series of qualitative states or behaviors, which facilitate easy interpretation as well as pattern matching for case retrieval.

This paper would be a valuable supplementary to the framework by Montani and Portinale in the sense that our key sequence discovery approach can be beneficially applied to the series of symbols abstracted from original time series. The point of departure is that, in many practical circumstances, significant transitional patterns in history are more worthy of attentions than the states or behaviors themselves associated with single episodes. It follows that the key sequences discovered will offer us useful knowledge to focus on what are really important in case characterization. Moreover, as the number of key sequences is usually is smaller than the number of elements in the series, indexing cases in terms of key sequences exhibits a further dimensionality reduction from series obtained via temporal abstraction.

Finding sequential patterns was widely addressed in the literature of sequence mining [2, 7, 21], where the goal was merely to find all legal sequential patterns with their frequencies of appearances above a user-specified threshold. High occurrence frequency was regarded in [3] as strong evidence to identify coherent sequences in a case-based system for recommendation of songs playlist. Identifying key sequences in our context differs from those in sequence mining in that we have to consider the cause-outcome effect for classification purpose. Only those non-redundant sequences that are not only frequent but also possess strong discriminating power will be selected.

Finally, but not the least, Martin and Plaza [10] investigated temporarily related cases prevalent in many real world domains. They defined sequential case as an assemblage of a few sub-cases among which a temporal order is established. Based on that a new model termed ceaseless CBR was proposed, which consists of the steps of ceaseless retrieval and ceaseless reuse. Ceaseless retrieval aims to continuously compare the sequence of alerts at hand with sequential cases in the case base to update the set of hypotheses on the occurrences of similar past cases, while ceaseless reuse is tasked to search for the combinations of such hypotheses to best explain the sequence of observational data so far received.

8 Conclusion

This paper aims to identify significant sequences to interpret and deal with dynamic properties of time series cases consisting of discrete, symbolic patterns. A knowledge discovery approach is proposed for this purpose. This approach uses the whole case library as available resources and is able to find from the problem space all qualified sequences that are non-redundant and indicative. An indicative sequence exhibits a high co-occurrence with a certain class and is hence valuable in offering discriminative strength for prediction. A sequence that is both indicative and non-redundant is termed as a key sequence.

It is shown that the key sequences discovered are highly usable to characterize time series cases in case based reasoning. The idea is to transform an original (lengthy) time series into a more concise representation in terms of the occurrences of key sequences detected. Four alternative ways to develop case indexes based on key sequences are suggested. Preliminary results of experiments have shown that these case indexes can lead to much better performance of the CBR system compared with using the whole symbolic series as problem descriptions. Further comparative studies regarding performance and applicability of these four case indexes will be done in conjunction with a number of medical application scenarios in future.

References

1. Aamodt A, Plaza E (1994) Case-based reasoning: foundational issues, methodological variations and systems approaches. AI Commun 7:39–59
2. Agrawal R, Srikant R (1995) Mining sequential patterns. In: Proceedings of the 11th international conference on data engineering, pp 3–14
3. Baccigalupo C, Plaza E (2006) Case-based sequential ordering of songs for playlist recommendation. In: Roth-Berghofer TR et al (eds) Proceedings of the European conference on case-based reasoning. Springer, Berlin, pp 286–300
4. Bellazzi R, Larizza C, Riva A (1998) Temporal abstractions for interpreting diabetic patients monitoring data. Intell Data Anal 2:97–122
5. Bichindaritz I, Conlon E (1996) Temporal knowledge representation and organization for case-based reasoning. In: Proceedings TIME-96. IEEE Computer Society, Washington, pp 152–159
6. Chan KP, Fu AW (1999) Efficient time series matching by wavelets. In: Proceedings of the international conference on data engineering, pp 126–133
7. Garofalakis MN, Rajeev R, Shim K (1999) SPIRIT: sequential pattern mining with regular expressing constraints. In: Proceedings of the 25th international conference on very large data bases, pp 223–234
8. Jaere MD, Aamodt A, Skalle P (2002) Representing temporal knowledge for case-based prediction. In: Craw S, Preece A (eds) Proceedings of the European conference on case-based reasoning, pp 174–188
9. Jarmulak J, Craw S, Rowe R (2000) Genetic algorithms to optimise CBR retrieval. In: Blanzieri E, Portinale L (eds) Proceedings of the European conference on case-based reasoning. Springer, Berlin, pp 136–147
10. Martin FJ, Plaza E (2004) Ceaseless case-based reasoning. In: Funk P, Gonzales Calero PA (eds) Proceedings of the European conference on case-based reasoning. Springer, Berlin, pp 287–301
11. McSherry D (2004) Explaining the Pros and Cons of conclusions in CBR. In: Proceedings of the European conference on case-based reasoning, pp 317–330
12. Montani S et al (2006) Case-based retrieval to support the treatment of end stage renal failure patients. Artif Intell Med 37:31–42
13. Montani S, Portinale L (2005) Case based representation and retrieval with time dependent features. In: Proceedings of the international conference on case-based reasoning. Springer, Berlin, pp 353–367
14. Nilsson M, Funk P (2004) A Case-based classification of respiratory sinus arrhythmia. In: Proceedings of the 7th European conference on case-based reasoning, Madrid. Springer, Berlin, pp 673–685
15. Olsson E, Funk P, Xiong N (2004) Fault diagnosis in industry using sensor readings and case-based reasoning. J Intell Fuzzy Syst 15:41–46
16. Perner P (2003) Incremental learning of retrieval knowledge in a case-based reasoning system. In: Ashley KD, Bridge DG (eds) Proceedings of the international conference on case-based reasoning. Springer, Berlin, pp 422–436
17. Salton G (1968) Automatic information organization and retrieval. McGraw–Hill, New York
18. Schmidt R, Heindl B, Pollwein B, Gierl L (1996) Abstraction of data and time for multiparametric time course prognoses. In: Advances of case-based reasoning. Lecture notes in artificial intelligence, vol 1168. Springer, Berlin, pp 377–391
19. von Schéele B (1999) Classification Systems for RSA, ETCO2 and other physiological parameters. PBM Stressmedicine. Technical Report, Heden 110, 82131 Bollnäs, Sweden
20. Shahar Y (1997) A framework for knowledge-based temporal abstractions. Artif Intell 90:79–133
21. Srikant R, Agrawal R (1996) Mining sequential patterns: generalizations and performance improvements. In: Proceedings of the 5th international conference on extending database technology, pp 3–17
22. Wu Y, Agrawal D, El Abbadi A (2000) A comparison of DFT and DWT based similarity search in time series databases. In: Proceedings of the 9th ACM CIKM conference on information and knowledge management, McLean, VA, pp 488–495
23. Zelikovitz S, Hirsh H (2002) Integrating background knowledge into nearest-neighbor text classification. In: Craw S, Preece A (eds) Proceedings of the European conference on case-based reasoning. Springer, Berlin, pp 1–5