Discovering Key Sequences in Time Series Data for Pattern Classification

Peter Funk and Ning Xiong

Department of Computer Science and Electronics
Mälardalen University
SE-72123 Västerås, Sweden
{peter.funk, ning.xiong@mdh.se}

Abstract. This paper addresses the issue of discovering key sequences from time series data for pattern classification. The aim is to find from a symbolic database all sequences that are both indicative and non-redundant. A sequence as such is called a key sequence in the paper. In order to solve this problem we first we establish criteria to evaluate sequences in terms of the measures of evaluation base and discriminating power. The main idea is to accept those sequences appearing frequently and possessing high co-occurrences with consequents as indicative ones. Then a sequence search algorithm is proposed to locate indicative sequences in the search space. Nodes encountered during the search procedure are handled appropriately to enable completeness of the search results while removing redundancy. We also show that the key sequences identified can later be utilized as strong evidences in probabilistic reasoning to determine to which class a new time series most probably belongs.

1 Introduction

Data mining attains growing importance to ease the knowledge acquisition bottleneck. It can be defined as efficiently discovering useful knowledge and information which are hidden somewhere in large databases. Extracting valuable knowledge from stored records/examples has been recognized as a non-trivial process of identifying novel, valid and potentially useful data patterns, and ideally also, to understand these data patterns for specific purpose [3].

Time series data bases present a relatively new research area for data mining. Unlike static databases where objects are described by attributes which are time independent, a time series database contains profiles of time-varying variables wherein pieces of data are associated with a timestamp and are meaningful only for a specific segment in a period. Analyses of time relevant data patterns are crucial for acquiring necessary knowledge to understand and predict the behavior of complex, dynamic processes.

Our paper studies the problem of key sequence discovery from symbolic time series data. The input is a collection of pairs of time series profiles and the associated classes. Each time series is a list of symbols corresponding to events that occurred in consecutive time segments. The task is to find all non-redundant sequences that are evaluated as frequent and indicative in discerning certain object classes. The non-redundancy of a sequence requires that it not contain any other sequence that has been identified to be indicative of the same class as it. A sequence that is both non-redundant and indicative is termed as a key sequence. The key sequences identified can later be utilized as strong evidences in probabilistic reasoning to determine to which class a new time series most probably belongs.

This study was primarily motivated by our AI project in stress medicine which aims at diagnosis of stresses based on sensor readings collected during patient respirations. Experimentally a patient is investigated through a series of 40-80 breathing cycles (including inhalation and exhalation). The classification of dysfunctional patterns for each breathing cycle has been implemented in the previous work using case based reasoning [12]. The next step is further to estimate the category of stress according to the series of breathing dysfunctions detected for successive respiration cycles. Related medical research has revealed that certain transitions of breathing patterns over time may possess high co-occurrence with stress categories of interest [16]. Finding such sequences from time series data is thus beneficial in offering valuable information to support clinical diagnoses.

Beside, there are many other application scenarios to which the work of this paper would be relevant. For instance, in health monitoring of engineering equipments, original sensor readings can be converted into discrete symbols [15], and some critical changes in time series of measurements like swell, sag, impulsive transients, might be signs indicating a present or potential anomaly. In telecommunications, useful information can be obtained from sequences of alarms produced by switches for analysis and prediction of network faults. In defense, sequences of deployments/actions of enemies would possibly betray their tactical intentions. Finally, in a medical scenario again, a data sequence of symptoms exhibited on a patient may help to forecast a disease that follows the emerging symptoms.

Some researches into time series data mining have been conducted recently. Three embedding methods were proposed by [5] to transform time series data into a vector space for classification purpose. Keogh and his colleagues addressed the issue of dimensionality reduction for indexing large time series databases [9] and also for fast search in these databases [10]. In [20] a family of three unsupervised methods was suggested to identify optimal and valid features given multivariate time series data. Similarity mining in time series was tackled by [8] and various methods for efficient retrieval of similar time sequences were discussed in [2, 6, 13, 19]. Algorithms for mining association rules were handled in [11, 14, 18] to model and predict time series behaviors in dynamic systems, and the application of association mining to disclose stock prices relations in time series was presented in [7].

This paper focuses on symbolic sequential data and proposes a novel approach to discovery of key sequences for time series classification. The remainder of the paper is organized as follows. In section 2 we briefly formulate our problem and show what kind of data sequences are targeted at. Sequences are evaluated in section 3 for distinguishing indicative ones. Section 4 details a sequence search algorithm with simulation results. Then, in section 5, we explain the utility of the discovered sequences in probabilistic diagnosis and classification. Finally the paper is concluded with summary remarks in section 6.

2 Problem Statements

To clearly present the proposed work, 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 databases, and then we precisely formulate the problem this paper aims to tackle.

Definition 1. A time series (profile) is a series of elements occurred sequentially over time, $X = \langle x(1), x(2), \cdots x(i), . \cdots, x(n) \rangle$, 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. Numeric values in time series may depict the evolution of a continuous variable as time elapsed, while symbolic values correspond to discrete events that happened or agent actions that were taken in successive time segments. In the following discussions we restrict our attention to symbolic time series consisting of discrete symbols.

Moreover, every time series profile has an inherent class. The previous time series data are assumed to have been classified and they are stored in a database together with their associated classes to facilitate data mining. A formal definition of time series database in the context of classification is given as follows:

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

With a time series database at hand, the data mining process involves analyzing sequences that are included in the database. A sequence of a time series profile is formally described in definition 3.

Definition 3. A sequence S of a time series profile $X = \langle x(1), x(2), \cdots, x(n) \rangle$ is a list consisting of elements taken from contiguous positions of X, i.e., $S = \langle x(k), x(k+1), \cdots, x(k+m-1) \rangle$ 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 database. But only a 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 database provided that

1) it appears in sufficient amount of time series profiles of the database;

2) the discriminating power of it, assessed upon the database, 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 database, 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 consequent, 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:

Given a time series database consisting of time series profiles and associated classes, find a set of indicative sequences {S_1, S_2, …, S_p} that satisfy the following two criteria:

1) For any i, j∈{1, 2, …p} neither $S_i \subseteq S_j$ nor $S_j \subseteq S_i$ if S_i and S_j are indicative of a same consequent;

2) For any sequence S that is indicative, S∈{S_1, S_2, …, S_p} if S is not redundant with respect to S_j for any j∈{1, 2, …p}.

The first criterion above requests compactness of the set of sequences *{S_1, S_2, …, S_p}* in the sense that no sequence in it is redundant by having a subsequence indicative of the same consequent 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 database 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, …, C_k}. The strength of the co-occurrence between sequence *S* and class C_i (i=1…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 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 consequents. Formally this definition of discriminating power *PD* is expressed as:

$$PD(S) = \max_{i=1\cdots k} P(C_i \mid S) \qquad \textbf{(1)}$$

In addition we say that the class yielding the maximum strength of the co-occurrences, i.e., $C = \arg_{i=1\cdots k} \max P(C_i \big| S)$, is the consequent that sequence *S* is indicative of.

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

$$P(C_i \mid S) = \frac{P(S \mid C_i)P(C_i)}{P(S)} \quad \textbf{(2)}$$

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

$$P(S) = P(S \mid C_i)P(C_i) + P(S \mid \overline{C}_i)P(\overline{C}_i) \quad \textbf{(3)}$$

equation (2) for conditional probability assessment can be rewritten as

$$P(C_i \mid S) = \frac{P(S \mid C_i)P(C_i)}{P(S \mid C_i)P(C_i) + P(S \mid \overline{C}_i)P(\overline{C}_i)} \quad \textbf{(4)}$$

Our aim here is to yield the conditional probability $P(C_i \mid S)$ in terms of equation (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 database. For instance we use the appearance frequency of sequence S in class C_i cases as an approximation of $P(S \mid C_i)$, thus we have:

$$P(S \mid C_i) \approx \tfrac{N(C_i, S)}{N(C_i)} \quad \textbf{(5)}$$

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

$$P(S \mid \overline{C}_i) \approx \tfrac{N(\overline{C}_i, S)}{N(\overline{C}_i)} \quad \textbf{(6)}$$

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 database. 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. Finally only those sequences that pass both thresholds 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 indicate. We first detail a sequence search algorithm for this purpose in subsection 4.1 and then we demonstrate simulation results on a synthetic database with the proposed algorithm in subsection 4.2.

4.1 A Sequence Search Algorithm

Discovery of key sequences can be considered as a search problem in a state space in which each state represents a sequence of symbols. Connection between two states signifies an operator between them for transition, i.e. addition or removal of a single symbol in time sequences. The state space for a scenario with three symbols *a*, *b*, *c* is illustrated in Fig. 1, where an arc connects two states if one can be created by extending the sequence of the other with a following symbol.

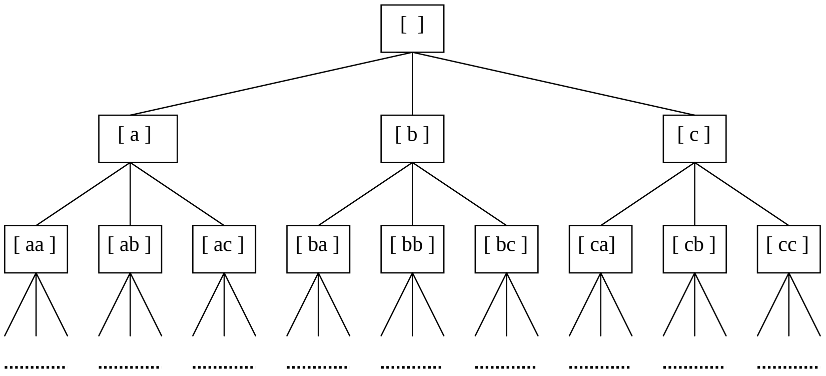

Fig. 1. The state space for sequences with three symbols

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 symbol 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 consequent. Otherwise the sequence is considered as non-redundant and hence is stored into the list of key sequences together with the consequent 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 consequents;

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.

The expansion of non-terminate nodes are 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 symbol. 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.
```

Finally it bears mentioning that finding key sequences in our context differs from those [1, 4, 17] in the literature of sequence mining. Usually the goal in sequence mining is merely to find all legal sequential patterns with their frequencies of appearances above a user-specified threshold. Here we have to consider the cause-outcome effect for classification purpose. Only those non-redundant sequences which are not only frequent but also possess strong discriminating power will be selected as the results of search.

4.2 Simulation Results

To verify the feasibility of the mechanism addressed above we now present the simulation results on a synthetic database. A case in this database is depicted by a time series of 20 symbols and one diagnosis class as the outcome. A symbol 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 d c], [b c a], [d e b], and [e 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 database was created in such a way that both sequences [a d c] and [b c a] had a chance of 80% of being reproduced once in the time series of class 1 while sequences [d e b] and [e a e] were added into class 2 and class 3 cases respectively with a probability of 90%. After stochastic reproduction of these key sequences, the remaining symbols in the time series of all cases were generated randomly. The whole database consists of 100 instances for each class. Presuming such time series cases to be randomly selected samples from a certain domain, a priori probability of each class is believed to be one third.

The sequence search algorithm was applied to this database to find key sequences and potential co-occurrences hidden in the data. The threshold for the discriminating power was set at 70% to ensure an adequate strength of the relationships discovered. We also specified 50 as the threshold of the evaluation base for reliable assessment of probabilities. The sequences found in our test are shown in table 1 below.

Table 1. Sequences discovered on a synthetic database

Sequence Discovered	Discriminating power	Evaluation base	Dominating Consequent
[a d c]	76.70%	103	Class 1
[b c a]	78.22%	101	Class 1
[d e b]	73.39%	124	Class 2
[e a e]	83.18%	107	Class 3

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%. These relationships with a degree of uncertainty are due to the many randomly generated symbols in the database such that any sequence of symbols is more or less probable to appear in time series of any class. But such nondeterministic property is prevalent in many real world domains.

5 Applying Key Sequences in Probabilistic Reasoning

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 characterize a dynamic time series in terms of what key sequences it includes as well as how many times each included key sequence has occurred. Further, as the key sequences have strong co-occurrences with a certain class, they can be used as discriminative evidences to update our beliefs concerning probabilities of classes an unknown time series may belong to.

Given a new time series X to be classified, the first task is characterization of the series according to the set of key sequences, say $\{S_1, S_2, \ldots, S_p\}$. Hence X has to be scanned thoroughly to detect all occurrences of key sequences in it. Every appearance of a key sequence S_j (j=1…P) in X is treated as an evidence for brief updating in classification. In view of this, the time series X can be characterized by a collection of evidences, i. e. $EV(X) = \{e_1, e_2, \cdots, e_T\}$ with $e_i \in \{S_1, \cdots, S_P\}$ for any i from 1 to T.

Important to note is that it is possible to have $e_i = e_j$ for $i \neq j$, implying that a key sequence appearing in X more than once is considered to cause multiple evidences.

The next task is to update the probabilities of different classes using detected evidences to reduce the uncertainty. Assuming conditional independence of key sequences occurrences under any class, the evidences available can be utilized separately for probability updating in individual steps. At every step we use a single evidence to revise prior probabilities according to the Bayes theorem and these updated probability estimates are then propagated as prior beliefs to the next step. Considering a two class problem without loss of generality, the procedure of probability updating using a set of evidences $\{e_1, e_2, \cdots, e_T\}$ is depicted by a series of equations as follows:

$$P(C \mid e_1) = \frac{P(e_1 \mid C)P(C)}{P(e_1 \mid C)P(C) + P(e_1 \mid \overline{C})P(\overline{C})} \qquad \textbf{(7)}$$

$$P(C \mid e_1, e_2) = \frac{P(e_2 \mid C)P(C \middle| e_1)}{P(e_2 \mid C)P(C \middle| e_1) + P(e_2 \mid \overline{C})P(\overline{C} \middle| e_1)} \qquad \textbf{(8)}$$

…………………………………………………

$$P(C \mid e_1, \cdots, e_i) = \frac{P(e_i \mid C)P(C \middle| e_1, \cdots, e_{i-1})}{P(e_i \mid C)P(C \middle| e_1, \cdots, e_{i-1}) + P(e_i \mid \overline{C})P(\overline{C} \middle| e_1, \cdots e_i)} \qquad \textbf{(9)}$$

……………………………………………………………

$$P(C \mid e_1, \cdots, e_T) = \frac{P(e_T \mid C)P(C \middle| e_1, \cdots, e_{T-1})}{P(e_T \mid C)P(C \middle| e_1, \cdots, e_{T-1}) + P(e_T \mid \overline{C})P(\overline{C} \middle| e_1, \cdots e_{T-1})} \qquad \textbf{(10)}$$

where the probabilities $P(e_i \middle| C)$ and $P(e_i \middle| \overline{C})$ for $i \in \{1, \ldots, T\}$ can be estimated according to equations (5) and (6) respectively, as e_i is a sequence. The probability updated in equation (7) represents the probability for class C given evidence e_1, which is further updated in equation (8) by evidence e_2 producing a more refined belief considering both e_1 and e_2. Generally the probability $P(C \mid e_1, \cdots, e_i)$ is yielded by updating the prior probability $P(C \mid e_1, \cdots, e_{i-1})$ with a new evidence e_i in equation (9). Finally we obtain the ultimate probability assessment incorporating all available evidences by equation (10).

We now give a concrete example to illustrate how the above sequential procedure works in probability refinements using key sequence appearances as evidences. Consider a problem of classifying a time series X into one of the two classes. Suppose that two key sequences S_1 and S_2 are detected in X and both are indicative of a certain class C. The a priori probability of class C is 50% and the probabilities of sequences S_1, S_2 in situations of class C and its complementary are shown below:

$$P(S_1 \middle| C) = 0.56 \qquad\qquad P(S_1 \middle| \overline{C}) = 0.24$$

$$P(S_2|C) = 0.80 \qquad\qquad P(S_2|\overline{C}) = 0.40$$

Further we assume that sequence S_1 appears twice in X and S_2 appears once, hence the collection of evidences for X is notated as $EV(X) = \{S_1, S_1, S_2\}$. With these three evidences detected, the probability of class C for time series X is refined gradually in the following three steps:

Step 1: 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 2: Refine the probability updated in step 1 with the second appearance of S_1, thus we have

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

It is clearly seen that the belief in class C is increased from 0.70 to 0.8448 due to the key sequence occurring for the second time.

Step 3: Refine the probability updated in step 2 with the occurrence of S_2, and we acquire the final probability assessment taking into account all evidences by

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

Here, the appearance of S_2 makes the probability be enhanced to an even higher value of 0.9159. As both sequences S_1 and S_2 are consistent in being indicative of the same consequence, each appearance of them contributes to increase the probability of C with a certain extent.

At last let us consider the order in which single evidences are used to refine probability assessments. This seems a fundamental issue and involves allocation of evidences to different steps of a sequential procedure. Fortunately our study has clarified that the order of evidences used in probability updating is completely indifferent. The final probability value remains constant as long as each piece of evidence is assigned to a distinct step. The claims as such are formally based on the following theorems.

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

A proof of the lemma is given in the appendix. Contemplating the implication of this lemma led us to a corollary presented below.

Corollary: Let $\{e_1, \cdots e_T\}$ be a set of evidences representing appearances of the key sequences in a time series X. The final probability for X in class C is independent of the order according to which single evidences $e_{1,}$ $e_{2, \ldots,}$ 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 evidences 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 evidence, we enable transitions to any orders of evidences without altering the classification result.

This corollary is important in providing theoretic arguments allowing for an arbitrary order of sequences to be used in probability refinement based on the Bayes theorem. The connotation is that when a key sequence occurred in the time series does not matter for the final result of classification. Instead only the numbers of appearances of key sequences effect our beliefs concerning the likelihoods of probable outcomes.

6 Conclusion

This paper tackles discovery of key sequences for time series data classification. The input is a symbolic database which consists of pairs of time series profiles and their associated classes. The problem is to find from the database a complete set of sequences that are both indicative and non-redundant. A sequence as such is called a key sequence in the paper.

Novel solutions are suggested here to deal with this problem. First we establish criteria to evaluate sequences in terms of the measures of evaluation base and discriminating power. The main idea is to accept those sequences appearing frequently and possessing high co-occurrences with consequents as indicative ones. Secondly a sequence search algorithm is proposed for exploration of indicative sequences in the problem space. One property of the search algorithm is that it always visits nodes of longer sequences after nodes of shorter ones such that redundant sequences can be detected easily for exclusion. The other property is that it terminates expansion only at the nodes where there is no prospect to find qualified sequences from their off-springs, guaranteeing the completeness of the search results.

The discovered key sequences are considered as important features in characterizing time series cases. We show that appearances of key sequences in time series can be used as evidences in probabilistic reasoning. A sequential procedure is presented to update beliefs for classification using found evidences. We also demonstrate that the order in which single evidences are used for brief refinements is indifferent to the final results.

Appendix: Proof of the Lemma

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

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

$$P(C|e_1,\cdots,e_i,e_{i+1}) = \frac{P(e_{i+1}|C)P(C|e_1,\cdots,e_i)}{P(e_{i+1}|C)P(C|e_1,\cdots,e_i) + P(e_{i+1}|\overline{C})P(\overline{C}|e_1,\cdots,e_i)} \qquad \textbf{(11)}$$

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

$$P(C|e_1,\cdots,e_i) = \frac{P(e_i|C)P(C|e_1,\cdots,e_{i-1})}{P(e_i|e_1,\cdots,e_{i-1})} \qquad \textbf{(12)}$$

Likewise we obtain

$$P(\overline{C}|e_1,\cdots,e_i) = \frac{P(e_i|\overline{C})P(\overline{C}|e_1,\cdots,e_{i-1})}{P(e_i|e_1,\cdots,e_{i-1})} \qquad \textbf{(13)}$$

Combining (12) and (13) into equation (11) gives rise to a transformed formulation as

$$P(C|e_1,\cdots,e_i,e_{i+1}) = \frac{P(e_{i+1}|C)P(e_i|C)P(C|e_1,\cdots,e_{i-1})}{P(e_{i+1}|C)P(e_i|C)P(C|e_1,\cdots,e_{i-1}) + P(e_{i+1}|\overline{C})P(e_i|\overline{C})P(\overline{C}|e_1,\cdots,e_{i-1})} \qquad \textbf{(14)}$$

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

$$P(e_{i+1}|C) = \frac{P(C|e_{i+1})P(e_{i+1})}{P(C)} \qquad \textbf{(15)}$$

$$P(e_{i+1}|\overline{C}) = \frac{P(\overline{C}|e_{i+1})P(e_{i+1})}{P(\overline{C})} \qquad \textbf{(16)}$$

$$P(e_i|C) = \frac{P(C|e_i)P(e_i)}{P(C)} \qquad \textbf{(17)}$$

$$P(e_i|\overline{C}) = \frac{P(\overline{C}|e_i)P(e_i)}{P(\overline{C})} \qquad \textbf{(18)}$$

where $P(C)$ and $P(\overline{C})$ denote the initial probability estimates for class C and its complementary without any evidences. Using the Bayes forms from (15) to (18), equation (14) is finally rewritten as

$$P(C|e_1,\cdots,e_i,e_{i+1}) = \frac{P^2(\overline{C})P(C|e_{i+1})|P(C|e_i)P(C|e_1,\cdots,e_{i-1})}{P^2(\overline{C})P(C|e_{i+1})|P(C|e_i)P(C|e_1,\cdots,e_{i-1}) + P^2(C)P(\overline{C}|e_{i+1})|P(\overline{C}|e_i)P(\overline{C}|e_1,\cdots,e_{i-1})} \qquad \textbf{(19)}$$

Clearly we see from equation (19) that the order between e_i and e_{i+1} has no effect at all on the probability $P(C|e_1,\cdots,e_i,e_{i+1})$ assessed. It follows that

$$P(C|e_1,\cdots,e_{i-1},e_i,e_{i+1}) = P(C|e_1,\cdots,e_{i-1},e_{i+1},e_i) \qquad \textbf{(20)}$$

and here from the lemma is proved.

References

1. Agrawal, R., Srikant, R.: Mining sequential patterns. In: Proceedings of the 11th International Conference on Data Engineering. (1995) 3-14
2. Chan, K. P., Fu, A. W.: Efficient time series matching by wavelets. In: Proceedings of the International Conference on Data Engineering. (1999) 126-133
3. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery. In: Advances in Knowledge Discovery and Data Mining. MIT Press (1996) 1-36
4. Garofalakis, M. N., Rajeev, R., Shim, K.: SPIRIT: Sequential sequential pattern mining with regular expression constraints. In: Proceedings of the 25th International Conference on Very Large Databases. (1999) 223-234
5. Hayashi, A., Mizuhara, Y., Suematsu, N.: Embedding time series data for classification. In: Perner, P., Imiya, A. (eds.): Proceedings of the IAPR International Conference on Machine Learning and Data Mining in Pattern Recognition. Leipzig (2005) 356-365
6. Hetland, M. L.: A survey of recent methods for efficient retrieval of similar time sequences. In: Last, M., Kandel, A., Bunke, H. (eds.): Data Mining in Time Series Databases. World Scientific (2004)
7. Huang, C. F., Chen, Y. C., Chen, A. P.: An association mining method for time series and its application in the stock prices of TFT-LCD industry. In: Perner, P. (ed.): Proceedings of the 4th Industrial Conference on Data Mining. Leipzig (2004)
8. Huhtala, Y., Kärkkäinen, J., Toivonen, H.: Mining for similarities in aligned time series using wavelets. In: Data Mining and Knowledge Discovery: Theory, Tools, and Technology. SPIE Proceedings Series, Vol. 3695. Orlando, FL (1999) 150-160
9. Keogh, E., Chakrabarti, K., Pazzani, M., Mehrotra, S.: Locally adaptive dimensionality reduction for indexing large time series databases. In: Proceedings of ACM SIGMOD Conference on Management of Data. Santa Barbara, CA (2001) 151-162
10. Keogh, E., Chakrabarti, K., Pazzani, M., Mehrotra, S.: Dimensionality reduction for fast similarity search in large time series databases. Journal of Knowledge and Information Systems (2001)
11. Last, M., Klein, Y., Kandel, A.: Knowledge discovery in time series databases. IEEE Trans. Systems, Man, and Cybernetics --- Part B: Cybernetics 31 (2001) 160-169
12. Nilsson, M., Funk, P.: A Case-Based Classification of Respiratory Sinus Arrhythmia. In: Proceedings of the 7th European Conference on Case-Based Reasoning. Madrid (2004) 673-685
13. Park, S., Chu, W. W., Yoon, J., Hsu, C.: Efficient search for similar subsequences of different lengths in sequence databases. In: Proceedings of the International Conference on Data Engineering. (2000) 23-32

14. Pray, K. A., Ruiz, C.: Mining expressive temporal associations from complex data. In: Perner, P., Imiya, A. (eds.): Proceedings of the IAPR International Conference on Machine Learning and Data Mining in Pattern Recognition. Leipzig (2005) 384-394
15. Ray, A.: Symbolic dynamic analysis of complex systems for anomaly detection. Signal Processing 84 (2004) 1115-1130
16. von Schéele, B.: Classification Systems for RSA, ETCO2 and other physiological parameters. PBM Stressmedicine, Technical report, www.pbmstressmedicine.se, (1999)
17. Srikant, R., Agrawal, R.: Mining sequential patterns: Generalizations and performance improvements. In: Proceedings of the 5th International Conference on Extending Database Technology. (1996) 3-17
18. Tung, A. K. H., Lu, H., Han, J., Feng, L.: Breaking the barrier of transactions: Mining inter-transaction association rules. In: Proceedings of ACM Conference on Knowledge Discovery and Data Mining. (1999) 297-301
19. Wu, Y., Agrawal, D., Abbadi, A. EI: 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 (2000) 488-495
20. Yoon, H., Yang, K., Shahabi, C.: Feature subset selection and feature ranking for multivariate time series. IEEE Trans. Knowledge and Data Engineering 17 (2005) 1186-1198