Condensed Representation of Sequential Patterns According to Frequency-Based Measures

Marc Plantevit and Bruno Crémilleux

GREYC-CNRS-UMR 6072
Université de Caen Basse-Normandie
Campus Côte de Nacre,
14032 Caen Cedex, France

Abstract. Condensed representations of patterns are at the core of many data mining works and there are a lot of contributions handling data described by items. In this paper, we tackle sequential data and we define an exact condensed representation for sequential patterns according to the frequency-based measures. These measures are often used, typically in order to evaluate classification rules. Furthermore, we show how to infer the best patterns according to these measures, i.e., the patterns which maximize them. These patterns are immediately obtained from the condensed representation so that this approach is easily usable in practice. Experiments conducted on various datasets demonstrate the feasibility and the interest of our approach.

1 Introduction

It is well-known that the "pattern flooding which follows data flooding" is unfortunate consequence in exploratory Knowledge Discovery in Databases (KDD) processes. There is a large range of methods to discover the patterns of a potential user's interest but the most significant patterns are lost among too much trivial, noisy and redundant information. Many works propose methods to reduce the collection of patterns, such as the constraint-based paradigm [15], the pattern set discovery approach [4,11], the so-called condensed representations [3,27] as well as the compression of the dataset by exploiting the Minimum Description Length Principle [19]. In practice, these methods often tackle data described by items (i.e., itemsets) and/or specific contexts, such as the largely studied frequent patterns extraction issue (a pattern X is said *frequent* if the number of examples in the database supporting X exceeds a given threshold). Many applications (e.g., security network, bioinformatics) require sequence mining. Oddly enough, even more than in the item domain, sequence mining suffers from the massive output of the KDD processes. However, little works focused on this aspect mainly because the difficult formalization required for sequential patterns. For instance, although there are many condensed representations of frequent itemsets, only closed sequential patterns have been proposed as a exact condensed representation for all the frequent sequential patterns [27]. Moreover,

N. Adams et al. (Eds.): IDA 2009, LNCS 5772, pp. 155–166, 2009.

some concise representations of itemset patterns cannot be used in order to condense frequent sequential patterns [17]. This illustrates the intrinsic difficulty to extend such works from itemsets to sequential patterns.

This paper addresses the issue of condensed representations of sequential patterns. The idea is to compute a representation $\mathcal{R}$ of the extracted patterns which is lossless: the whole collection of patterns can be efficiently derived from $\mathcal{R}$. This approach has been mainly developed in the context of frequency [3,27] and there are very few works addressing other measures [8,21,22]. In this paper, we investigate exact condensed representations of sequential patterns based on many interestingness measures, the so-called *frequency-based measures* (see Section 3). These measures (e.g., frequency, confidence, lift, growth rate, information gain) are precious in real-world applications to evaluate the interestingness of patterns and the quality of classification rules [20]. For instance, the emerging measure is very useful to characterize classes and classify them. Initially introduced in [5], emerging patterns (EPs) are patterns whose frequency strongly varies between two datasets (i.e., two classes). An EP can be seen as a classification rule and EPs are at the origin of various works such as powerful classifiers [13]. From an applicative point of view, we can quote many works on the characterization of biochemical properties or medical data [14]. A condensed representation of itemsets according to frequency-based measures has already been proposed [22], but it is only limited to the item domain.

The contribution of this paper is twofold. First, we define an exact condensed representation of sequential patterns according to the frequency-based measures. *Exact* means that we are able to infer not only the patterns, but also the measure values associated to the patterns without accessing the data. This is useful because the user is mainly interested in these values. For that purpose, the key idea is to show that the value of a frequency-based measure of any sequential pattern can be deduced from one of its closed sequential patterns. This idea has already been used in the item domain [22], but not in sequential data. Contrary to itemsets, a sequential pattern may have several closed sequential patterns, our method overcomes this difficulty. As this condensed representation is based on the closed sequential patterns and there are efficient algorithms to extract these patterns, these algorithms are also efficient to mine such a condensed representation. Second, we define the notion of strong sequential patterns (SPs) according to frequency-based measures. Given a frequency-based measure, these patterns maximize it. This is interesting because it highlights the best patterns with respect to the measure and moreover it reduces the output. On the other hand, the SPs are immediately obtained from the condensed representation. Finally, experiments conducted on various datatsets demonstrate the feasibility of our approach and quantify the interests of SPs.

This paper is organized as follows. Section 2 provides the preliminaries which are needed for the rest of the paper. In Section 3, we propose a condensed representation of sequential patterns according to the frequency-based measures. Section 4 defines the strong frequency-based measures and the SPs. Section 5 provides in depth experimental results and we review related work in Section 6.

2 Preliminary Concepts and Definitions

Let $\mathcal{I} = \{i_1, i_2, \ldots, i_n\}$ be a finite set of items. An *itemset* I is a subset of $\mathcal{I}$. A sequence $s = \langle I_1, I_2, \ldots, I_n\rangle$ is an ordered list of itemsets. A sequence $s_\alpha = \langle A_1, A_2, \ldots, A_n\rangle$ is said to be contained in another sequence s_β=$\langle B_1, B_2, \ldots, B_m\rangle$ if there exist integer $1 \leq i_1 < i_2 < \ldots < i_n \leq m$ such that $A_1 \subseteq B_{i_1}, A_2 \subseteq B_{i_2}, \ldots, A_n \subseteq B_{i_n}$ (denoted by $s_\alpha \sqsubseteq s_\beta$). If the sequence s_α is contained in the sequence s_β, s_α is called a *subsequence* of s_β and s_β a *supersequence* of s_α.

Input data in sequential pattern mining consists in a collection of sequences. As previously highlighted in the introduction section, frequency based-measures are mainly used to assess the quality of classification rules and a class identifier is associated to each data sequence. Therefore, the input database $\mathcal{D}$ consists in a collection of tuples (sid, s, c) where sid is a sequence identifier, s is sequence and c is a class identifier (see the example given in Tab. 1). $\mathcal{D}$ corresponds to a partition of i subsets $\mathcal{D}_i$ where each $\mathcal{D}_i$ contains all tuples (sid, s, c_i) in $\mathcal{D}$. Each sequence belongs to a single subset $\mathcal{D}_i$. A tuple (sid, s, c) is said to *contain* a sequence s_α if $s_\alpha \sqsubseteq s$. The *intersection* of a set of sequences $S = \{s_1, s_2, \ldots, s_n\}$, denoted $\bigcap s_i \in S$, is the set of all maximal subsequences contained into all the s_i. For example, the intersection of $s = \langle c, b, c, a\rangle$ and $s' = \langle c, b, a, c, c, c\rangle$ is $\{\langle c, b, a, \rangle, \langle c, b, c\rangle\}$.

Table 1. Toy database $\mathcal{D}$ with class values

Seq_id	Sequence	Class
s_1	$\langle c, b, c, a\rangle$	c_1
s_2	$\langle c, b, a, c, c, c\rangle$	c_1
s_3	$\langle a, a, a, c, c, a, a\rangle$	c_2
s_4	$\langle a, a, b, a, c, c\rangle$	c_2

Frequency-based measures are linked to the notions of *support*. The *absolute support* of a sequence s_α in $\mathcal{D}$ is the number of tuples in $\mathcal{D}$ that contain s_α, denoted by $support(s_\alpha, \mathcal{D})$. The *relative support* of s_α is the percentage of tuples in $\mathcal{D}$ that contain s_α. For instance, $support(\langle c, a\rangle, \mathcal{D}) = 3$. Unless otherwise stated, we use the absolute support all along this paper.

Let *minsup* be a minimum support threshold. A sequence s_α is a frequent sequence on $\mathcal{D}$ if $support(s_\alpha, \mathcal{D}) \geq minsup$. A frequent sequence s_α is a closed frequent sequence if there does not exist a sequence s_β such that $support(s_\alpha, \mathcal{D}) = support(s_\beta, \mathcal{D})$ and $s_\alpha \sqsubset s_\beta$. Then, given $\mathcal{D}$ and *minsup*, the problem of mining frequent closed sequential patterns is to find the complete set of frequent closed sequences. Function $Closed(x, \mathcal{D})$ from Definition 1 return the set of closed sequential patterns in sequence database $\mathcal{D}$ which contains a sequence s.

Definition 1 (*$Closed(x, \mathcal{D})$*)**.** *Let x be a sequential pattern and $\mathcal{D}$ be a sequence database.*

$$Closed(x, \mathcal{D}) = \bigcap \{s \in \mathcal{D} | x \sqsubseteq s\}$$

Following our example in Table 1, we get: $Closed(\langle c, b\rangle, \mathcal{D}) = \{\langle c, b, a\rangle, \langle c, b, c\rangle\}$. These two sequences are closed in $\mathcal{D}$. Finally, we recall the notion of classification sequential rule.

Definition 2 (Classification sequential rules). *Let $\mathcal{C} = \{c_1, c_2, \ldots, c_m\}$ be a set of class values, a classification sequential rule is a rule $R = s \rightarrow c_i$ where s is a sequential pattern and $c_i \in \mathcal{C}$.*

3 Exact Condensed Representation of Sequential Pattern According to Frequency Based Measures

Various measures [7] are used to evaluate the quality of classification rules. Many measures are based on the frequency of the sequential patterns s and the concatenation of s and c_i, i.e. $\langle s, \{c\}\rangle$. These measures are called frequency-based measures and are defined as follows:

Definition 3 (Frequency-Based Measure). *Let $\mathcal{D}$ be a sequence database partitioned into k subsets denoted $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k$, a frequency-based measure M_i to characterize $\mathcal{D}_i$ is a function F of supports: $support(s, \mathcal{D}_1)$, $support(s, \mathcal{D}_2), \ldots, support(s, \mathcal{D}_k)$,* i.e. *$M_i(s) = F(support(s, \mathcal{D}_1), support(s, \mathcal{D}_2), \ldots, support(s, \mathcal{D}_k))$.*

With the notation M_i, the subscript i denotes the dataset $\mathcal{D}_i$ which is characterized according to the measure M. A frequency-based measure consists of a finite combination of supports of a pattern s on several sequence data sets $\mathcal{D}_i$. More precisely, a frequency-based measure cannot contain other parameters. Table 2 lists some well-known frequency-based measures that are commonly used in the literature. These measures are given here by using the absolute support whereas the literature often writes them in term of conditional probabilities [7] ($P(X|c_i)$) corresponds to $\frac{support(X,\mathcal{D}_i)}{|\mathcal{D}_i|}$ where X is a (sequential) pattern. Note that some frequency-based measures (e.g., J-Measure, confidence, lift, growth rate) are expressed with supports that are not restricted to sets $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k$. However, these measures respect Definition 3 because these supports can be computed from $support(s, \mathcal{D}_1), support(s, \mathcal{D}_2), \ldots support(s, , \mathcal{D}_k)$. For instance, $support(s, \mathcal{D}) = \sum_{j=1}^{k} support(s, \mathcal{D}_j)$.

To compute the value of a frequency-based measure for a rule $s \rightarrow c_i$, computing the support of s in datasets $\mathcal{D}$ and $\mathcal{D}_i$ ($support(s, \mathcal{D})$ and $support(s, \mathcal{D}_i)$) is enough. An important result is that these frequencies can be obtained thanks to the set of closed sequential patterns in $\mathcal{D}$ and $\mathcal{D}_i$. Indeed, we have:

- $\forall e \in Closed(s, \mathcal{D}) \quad support(s, \mathcal{D}) = support(e, \mathcal{D})$
- $\forall e \in Closed(s, \mathcal{D}_i)\ support(s, \mathcal{D}_i) = support(e, \mathcal{D}_i)$

Consequently, the computation of $Closed(s, \mathcal{D})$ and $Closed(s, \mathcal{D}_i)$ are enough to compute $support(s, \mathcal{D})$ and $support(s, \mathcal{D}_i)$. Furthermore, the following property indicates that the computation of the function $Closed$ can be made only once:

Table 2. Examples of frequency-based measures characterizing $\mathcal{D}_i$

Frequency-based measure	Formula	Strong
J-Measure	$\frac{support(s,\mathcal{D}_i)}{\lvert\mathcal{D}\rvert} \times \log \frac{support(s,\mathcal{D}_i)\times\lvert\mathcal{D}\rvert}{\lvert\mathcal{D}_i\rvert\times support(s,\mathcal{D})}$	no
Relative support	$\frac{support(s,\mathcal{D}_i)}{\lvert\mathcal{D}\rvert}$	yes
Confidence	$\frac{support(s,\mathcal{D}_i)}{support(s,\mathcal{D})}$	yes
Sensitivity	$\frac{support(s,\mathcal{D}_i)}{\lvert\mathcal{D}_i\rvert}$	yes
Success rate	$\frac{support(s,\mathcal{D}_i)}{\lvert\mathcal{D}\rvert} + \frac{\lvert\mathcal{D}\setminus\mathcal{D}_i\rvert - support(s,\mathcal{D}\setminus\mathcal{D}_i)}{\lvert\mathcal{D}\rvert}$	yes
Specificity	$\frac{\lvert\mathcal{D}\setminus\mathcal{D}_i\rvert - support(s,\mathcal{D}\setminus\mathcal{D}_i)}{\lvert\mathcal{D}\rvert}$	yes
Piatetsky-Shapiro's (PS)	$\frac{support(s,\mathcal{D}_i)}{\lvert\mathcal{D}\rvert} - \frac{support(s,\mathcal{D})}{\lvert\mathcal{D}\rvert} \times \frac{\lvert\mathcal{D}_i\rvert}{\lvert\mathcal{D}\rvert}$	yes
Lift	$\frac{support(s,\mathcal{D}_i)\times\lvert\mathcal{D}\rvert}{support(s,\mathcal{D})\times\lvert\mathcal{D}_i\rvert}$	yes
Odd ratio (α)	$\frac{support(s,\mathcal{D}_i)\times(\lvert\mathcal{D}\setminus\mathcal{D}_i\rvert - support(s,\mathcal{D}\setminus\mathcal{D}_i))}{(support(s,\mathcal{D}) - support(s,\mathcal{D}_i))\times(\lvert\mathcal{D}_i\rvert - support(s,\mathcal{D}_i)}$	yes
Growth rate (GR)	$\frac{\lvert\mathcal{D}\rvert - \lvert\mathcal{D}_i\rvert}{\lvert\mathcal{D}\rvert} \times \frac{support(s,\mathcal{D}_i)}{support(s,\mathcal{D}) - support(s,\mathcal{D}_i)}$	yes
Information Gain	$\log \frac{support(s,\mathcal{D}_i)\times\lvert\mathcal{D}\rvert}{support(s,\mathcal{D})\times\lvert\mathcal{D}_i\rvert}$	yes

Property 1. Let s be a sequential pattern and $\mathcal{D}_i$ a subset of $\mathcal{D}$, $\forall e \in Closed(s,\mathcal{D})$, $support(s,\mathcal{D}_i) = support(e,\mathcal{D}_i)$

Proof. According to Definition 1, sequence e from $Closed(s,\mathcal{D})$ is a supersequence of s having the same support in $\mathcal{D}$. Since s is a subsequence of e, all sequences from $\mathcal{D}$ that contain e also contain s. Moreover, sequences s and e have the same support in $\mathcal{D}$. Thus, they are contained by the same sequences of $\mathcal{D}$. Since $\mathcal{D}_i$ is a subset of $\mathcal{D}$, sequences e and s are contained in the same sequences of $\mathcal{D}_i$. Thus, $support(s,\mathcal{D}_i) = support(e,\mathcal{D}_i)$.

As said in Section 2, a sequential pattern s may have several closed patterns. Theorem 1 shows that all closed patterns of s have the same value for a measure. Consequently, the value of a frequency-based measure for s can be deduced from any of its closed sequential patterns:

Theorem 1. *Let s be a sequential pattern, we have:*

$$\forall e \in Closed(s,\mathcal{D}),\ M_i(s) = M_i(e)$$

Proof. Let s be a sequential pattern. Since $\forall e \in Closed(s,\mathcal{D})$, $support(s,\mathcal{D}_i) = support(e,\mathcal{D}_i)$ (property 1), we can express $M_i(s) = F(support(s,\mathcal{D}_1), support(s,\mathcal{D}_2), \ldots, support(s,\mathcal{D}_k))$ by $M_i(s) = F(support(e,\mathcal{D}_1), support(e,\mathcal{D}_2), \ldots, support(e,\mathcal{D}_k)) = M_i(e)$ where $e \in Closed(s,\mathcal{D})$. Thus $M_i(s) = M_i(e)$.

For example, $Closed(\langle c,b\rangle,\mathcal{D}) = \{\langle c,b,a\rangle, \langle c,b,c\rangle\}$, and $Confidence_{c_1}(\langle c,b\rangle) = Confidence_{c_1}(\langle c,b,a\rangle) = 1$. The closed sequential patterns with their values of the measure M_i are enough to synthesize the set of sequential patterns according to M_i. As a consequence, the closed sequential patterns with their values of the measure M_i are an exact condensed representation of the whole set of sequential

patterns according to M_i. In practice, the number of closed patterns is lower (and often much lower) than the complete set of sequential patterns. More generally, this condensed representation benefits from all the advantages of the condensed representation based on the closed sequential patterns [27,25].

4 Strong Sequential Patterns According to Frequency-Based Measures

In practice, the number of patterns satisfying a given threshold for a measure M_i can be very large and hampers their individual analysis. In this section, we show that our approach easily enables us to highlight the best patterns according to measures, that is to say the patterns which maximize such measures. To achieve this result, we have to consider a slightly different set of measures, the strong frequency-based measures:

Definition 4 (Strong Frequency-Based Measure). *A frequency-based measure M_i which decreases with $support(s, \mathcal{D})$ when $support(s, \mathcal{D}_i)$ remains unchanged, is a strong frequency-based measure.*

Most frequency-based measures are also strong frequency-based measures (in Table 2, only the J-measure is not a strong frequency-based measure). More generally, Definition 4 is less restrictive than the property P_3 of Piatetsky-Shapiro's framework [16] which defines three properties which have to be satisfied by an interestingness measure to be qualified as a "good" one.

Theorem 2 indicates that the closed sequential patterns satisfy an interesting property w.r.t. the strong frequency-based measures.

Theorem 2. *Let M_i be a strong frequency-based measure and s be a sequential pattern, we have $\forall e \in Closed(s, \mathcal{D}_i), M_i(s) \leq M_i(e)$. The elements from $Closed(s, \mathcal{D}_i)$ are called strong sequential patterns (SPs) in class i or dominant sequential patterns for M_i.*

Proof. Let M_i be a strong frequency-based measure and s be a sequential pattern. $\forall e \in Closed(s, \mathcal{D}_i)$, we have $support(s, \mathcal{D}_i) = support(e, \mathcal{D}_i)$ (see Definition 1). As $s \sqsubseteq e$, we obtain that $support(s, \mathcal{D}) \geq support(e, \mathcal{D})$ thanks to the anti-monotonicity of the support. By definition 4, we conclude that $M_i(s) \leq M_i(e)$.

The result given by Theorem 2 is important: it means that the closed sequential patterns in $\mathcal{D}_i$ maximize any strong frequency-based measure M_i. In other words, a sequential pattern that is not closed in $\mathcal{D}_i$ has a lower (or equal) value than one of its closed sequential patterns in $\mathcal{D}_i$ for any measure M_i.

However, Theorem 2 means that mining all closed sequential patterns in each $\mathcal{D}_i$ is needed, which indeed require a lot of computation. Lemma 1 links closed sequential patterns in $\mathcal{D}_i$ with closed sequential patterns in $\mathcal{D}$. For that, we first have to define the sequence concatenation. Given a sequence $s_\alpha = \langle A_1, A_2, \ldots, A_n \rangle$ and a class c, the concatenation of sequence s_α with $\langle c \rangle$, denoted $s_\alpha \bullet c$ is $\langle A_1, A_2, \ldots, A_n, \{c\} \rangle$. We then consider the sequence database $\mathcal{D}'$

from $\mathcal{D}$ where each data sequence contains a new item that represents their class value. For each tuple (sid, s, c) we add the tuple $(sid, s \bullet c, c)$ in $\mathcal{D}'$. Then, like $\mathcal{D}$, $\mathcal{D}'$ corresponds to a partition of i subsets $\mathcal{D}'_i$ where each $\mathcal{D}'_i$ contains all tuples $(sid, s \bullet c_i, c_i)$. Note that we have the relation $support(s, \mathcal{D}_i) = support(s \bullet c_i, \mathcal{D}'_i)$.

Lemma 1. *If the sequence $s \bullet c_i$ is a closed sequential pattern in $\mathcal{D}'_i$ then $s \bullet c_i$ is a closed sequential pattern in $\mathcal{D}'$.*

Proof. By construction of the subsets of $\mathcal{D}'$, a class value c_i is only contained in $\mathcal{D}'_i$ and not in the other datasets. So we have $support(s, \mathcal{D}'_i) = support(s \bullet c_i, \mathcal{D}')$.

Thanks to Lemma 1, we can give Property 2 which indicates that mining only the closed sequential patterns in $\mathcal{D}'$ is enough. In other words, only one extraction of closed patterns is needed.

Property 2 (SPs: computation of their frequency-based measure). If s is a strong sequential pattern in $\mathcal{D}_i$, then $M_i(s)$ can directly be computed with the supports of the condensed representation based on the closed sequential patterns of $\mathcal{D}'$.

Proof. Let s be a SP in $\mathcal{D}_i$. Thus, $s \bullet c_i$ is a closed sequential pattern in $\mathcal{D}'_i$. To compute $M_i(s)$, it is necessary to know $support(s, \mathcal{D}'_i)$ and $support(s, \mathcal{D}')$. By definition of $\mathcal{D}'_i$, $support(s, \mathcal{D}'_i) = support(s \bullet c_i, \mathcal{D}')$ and lemma 1 ensures that $s \bullet c_i$ is closed in $\mathcal{D}'$. As a consequence, its support is provided by the condensed representation of the closed sequential patterns of $\mathcal{D}'$. To compute $support(s, \mathcal{D}')$, two cases are possible: (i) if s is a closed sequential pattern in $\mathcal{D}'$, its support is directly available; (ii) if, s is not a closed sequential pattern in $\mathcal{D}'$, then $s \bullet c_i$ belongs to $Closed(s, \mathcal{D}')$ and $support(s, \mathcal{D}') = support(s \bullet c_i, \mathcal{D}')$.

We have defined a theoretical framework for SPs in database $\mathcal{D}$ and its subset $\mathcal{D}_i$. In practice, these patterns can be discovered in $\mathcal{D}'$ and their frequencies can also be computed in $\mathcal{D}'$ thanks to any closed sequential pattern mining algorithm. Indeed, if s is a strong sequential pattern in $\mathcal{D}_i$, then $s \bullet c_i$ is a closed sequential pattern in $\mathcal{D}'$, $support(s, \mathcal{D}_i) = support(s \bullet c_i, \mathcal{D}')$ and $support(s, \mathcal{D}) = support(s, \mathcal{D}')$.

Example 1. Following our example in Table 1, with $minsup = 2$, we have 11 closed frequent sequential patterns. In particular, $\langle c, b, a \rangle$ and $\langle c, b, c \rangle$ are SPs for class c_1, $\langle a, a, a, c, c \rangle$ is a SP for class c_2. Thus these sequences maximize any frequency-based measure M_i.

Let M_2 be the confidence measure, $\langle a, a, a, c, c \rangle \bullet c_2$ is a closed sequential pattern in $\mathcal{D}'$. To compute its confidence, we need to know $support(\langle a, a, a, c, c \rangle, \mathcal{D}')$. Since sequence $\langle a, a, a, c, c \rangle$ is not a closed sequential pattern in $\mathcal{D}'$, then $support(\langle a, a, a, c, c \rangle, \mathcal{D}') = support(\langle a, a, a, c, c \rangle \bullet c_2, \mathcal{D}') = 2$. Thus the confidence of SP $\langle a, a, a, c, c \rangle$ for class c_2 is 1.

5 Experiments

Experiments have been carried out on real datasets by considering the emerging measure (Growth Rate) [5]. The emerging measure is very useful to characterize

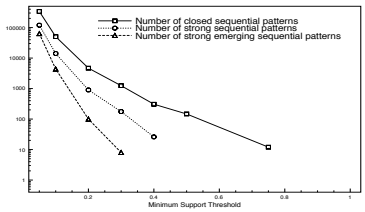
classes and classify them. Emerging patterns (EPs) are patterns whose frequency strongly varies between two datasets (i.e., two classes). Note that any frequency-based measure can be used. However, due to the space limitation, we only report experiments on strong emerging frequent sequential patterns (SESPs). To mine closed sequential patterns, we have implemented Bide algorithm [25] in Java language (JVM 1.5). Furthermore, we do not report results about the runtime of the discovery of SEPSs. However, it is important to note that the computation of SP growth rates is negligible compared to the step of frequent closed sequential pattern mining. We can conclude that the scalability issue of SPs' discovery is Bide-dependent and Bide is known as being a scalable and robust algorithm. Consequently, the discovery of SESPs is then scalable.

In these experiments, we consider the following real datasets:

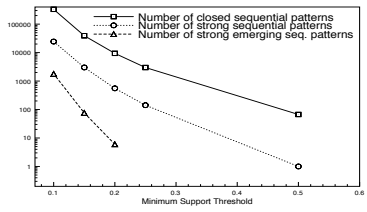
- *E.Coli Promoters dataset:* The *E. Coli* Promoters data set [23] is available on the UCI machine learning repository [1]. The data set is divided into two classes: 53 *E. Coli* promoter instances and 53 non-promoter instances. We consider pairs of monomers (*e.g.,* aa, ac, etc.) as items.
- *PSORTdb v.2.0 cytoplasmic dataset:* The cytoplasmic data set was obtained from PSORTdb v.2.0 [6]. The data set contains 278 cytoplasmic Gram-negative sequences and 194 Gram-positive sequences. We consider items in the same way as in the previous dataset.
- *Greenberg's Unix dataset:* We transform the original Unix dataset [9] into a new data set that contains 18681 data sequences where a data sequence contains a session of a Unix command shell user. These sequences are divided into 4 classes: 7751 sequences about navigation of computer scientists, 3859 sequences for *experienced-programmers*, 1906 sequences for *non-programmers* and 5165 sequences about *novice-programmers*.
- *Entree Chicago Recommendation Dataset:* We use the data set underlying the Entree system [2]. This data set is also available on the UCI machine learning repository [1]. For each restaurant, a sequence of features is associated. We consider 8 classes (Atlanta, Boston, Chicago, Los Angeles, New Orleans, New York, San Francisco and Washington DC) that respectively contain $267, 438, 676, 447, 327, 1200, 414$ and 391 sequences.

These experiments aim at studying several quantitative results of the discovery of strong sequential patterns satisfying both a growth rate threshold and a support threshold.

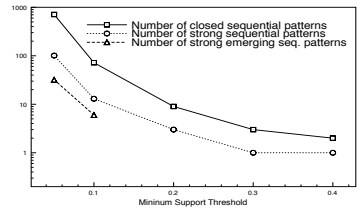
Figures from Fig. 1 report the number of frequent closed, strong and emerging sequential patterns according to the minimum support threshold. Obviously, the number of patterns decreases when the support threshold increases. We note that the number of SESPs is much lower than the number of SPs which is itself significantly lower than the number of sequential patterns (the figure uses a logarithmic scale). It indicates a high condensation of patterns reducing the output and highlighting the most valuable patterns according to the measures. Note that there is no SESP (and no SP in E.coli and Entree datasets) when *minsup* is high because no pattern can satisfy the growth rate measure.



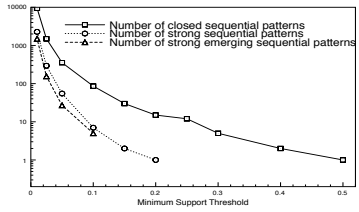
(a) E.coli dataset ($mingr = 3$)



(b) PSORTdb v.2.0 cytoplasmic dataset ($mingr = 1.5$)

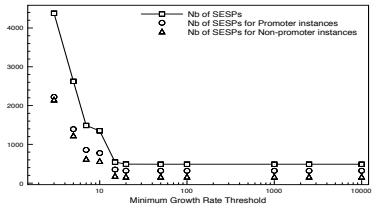


(c) Unix log dataset ($mingr = 3$)

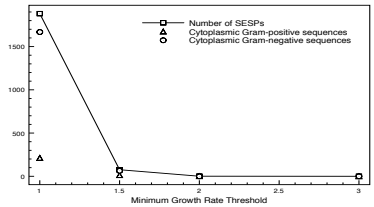


(d) Entree dataset ($mingr = 3$)

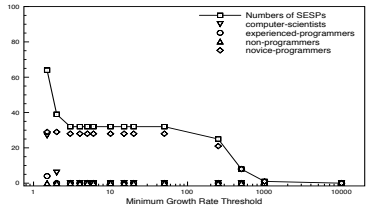
Fig. 1. Numbers of closed, strong and emerging sequential patterns according to the minimum support threshold



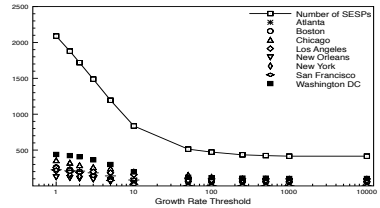
(a) E.coli dataset ($minsup = 0.1$)



(b) PSORTdb v.2.0 cytoplasmic dataset ($minsup = 0.15$)



(c) Unix log dataset (minsup=0.05)



(d) Entree dataset (minsup=0.01)

Fig. 2. Number of strong emerging sequential patterns and their repartition according to the growth rate threshold

Figures from Fig. 2 report the number of SESPs and their distribution among the different class values according to the growth rate threshold. The number of SESPs decreases when the growth rate threshold increases. However, this

number does not always tend to zero. Indeed, some SESPs with an infinite growth rate can appear (see Fig. 2(a) and Fig. 2(d)). These particular SESPs are called jumping SESPs (JSESPs). They are sequential patterns that appear for only one class value, and never appear for other class values. It should be noticed that the repartition of SESPs among the class values is not necessary uniform. For instance, class value *novice-programmers* for Fig. 2(c) and class value *Washington DC* for Fig. 2(d) contain a significantly greater number of SESPs than the others.

Discovered SESPs: Sequential pattern $\langle$(aa)(at)(ta)(gc)$\rangle$ is a JSESP for promoter sequences in *E. coli* dataset while $\langle$(tg)(cg)(ac)(tg)$\rangle$ is a JSESP for non-promoter sequences. According to *Entree* dataset, the sequence $\langle$(Week-end Dining)(Parking-Valet)$\rangle$ is a SESP for class *Washington DC* with a growth rate $gr = 10.02$. Sequential pattern $\langle$pix, umacs, pix, umacs$\rangle$ is a SESP for class *novice-programmers* with a growth rate $gr = 450$. Let us recall that to the best of our knowledge, our method is the unique method to discover such patterns.

6 Related Work

Main works on condensed representations have been outlined in the introduction. A condensed representation of frequency-based measures has already been proposed in [22], but it is limited to the item domain and our work can be seen as an extension of [22] to the sequence framework. To the best of our knowledge, there is no work in the literature that addresses condensed representations of sequential patterns w.r.t. any frequency-based measure.

In the literature, classification on sequence data has been extensively studied. In [26], the authors introduce the problem of mining sequence classifiers for early prediction. Criteria for feature selection are proposed in [12]. The authors use the confidence to quantify the features. Our work can lead to a generalization of this work by allowing the use of any frequency-based measure. In [18] frequent subsequences are used for classification but the interestingness of a pattern is valued according to the only confidence measure.

An approach to detect sequential pattern changes between two periods is proposed in [24]. First, two sequential pattern sets are discovered in the two-period databases. Then, the dissimilarities between all pairs of sequential patterns are considered. Finally, a sequential pattern is classified as one of the following three change-types: an emerging sequential pattern, an unexpected sequence change, and an added/perish sequence. These latter correspond to jumping emerging sequences. Note that the notion of EPs differs from [5]. This work does not consider condensed representations. Moreover, two databases have to be mined and then similarities between each pair of sequences have to be computed whereas our framework needs only one database mining and no computation of sequence similarities.

7 Conclusion

In this paper, we have investigated condensed representations of sequential patterns according to many interestingness measures and we have proposed an exact condensed representation of sequential patterns according to the frequency-based measures. Then, we have defined the strong sequential patterns which are the best patterns according to the measures. These patterns are straightforwardly obtained from the condensed representation so that this approach can be easily used in practice. Experiments show the feasibility and the interest of the approach.

We think that condensed representations of patterns have a lot of applications and their use is not limited to obtain more efficiently patterns associated to their interestingness measures. As they can be used as cache mechanisms, they make interactive KDD processes more easily and are a key concept of inductive databases. Moreover, their properties are useful for high-level KDD tasks such as classification or clustering. Finally, the behavior of interestingness measures has been studied in [10] and the next step is to determine lower bounds for weighted combinations of frequency-based measures in order to ensure a global quality according to a set of measures.

Acknowledgments. The authors would like to thank Arnaud Soulet (Université François Rabelais de Tours, Fr) for very fruitful comments and invaluable discussions. This work is partly supported by the ANR (French National Research Agency) funded project Bingo2 ANR-07-MDCO-014.

References

1. Asuncion, A., Newman, D.: UCI machine learning repository (2007), `http://www.ics.uci.edu/~mlearn/MLRepository.html`
2. Burke, R.D.: The wasabi personal shopper: A case-based recommender system. In: AAAI/IAAI, pp. 844–849 (1999)
3. Calders, T., Rigotti, C., Boulicaut, J.-F.: A survey on condensed representations for frequent sets. In: Constraint-Based Mining and Inductive Databases, pp. 64–80 (2004)
4. De Raedt, L., Zimmermann, A.: Constraint-based pattern set mining. In: SDM (2007)
5. Dong, G., Li, J.: Efficient mining of emerging patterns: discovering trends and differences. In: ACM SIGKDD 1999, San Diego, CA, pp. 43–52. ACM Press, New York (1999)
6. Gardy, J.L., Spencer, C., Wang, K., Ester, M., Tusnady, G.E., Simon, I., Hua, S.: PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria. Nucl. Acids Res. 31(13), 3613–3617 (2003)
7. Geng, L., Hamilton, H.J.: Interestingness measures for data mining: A survey. ACM Comput. Surv. 38(3) (2006)
8. Giacometti, A., Laurent, D., Diop, C.T.: Condensed representations for sets of mining queries. In: Knowledge Discovery in Inductive Databases, 1st International Workshop, KDID 2002 (2002)

9. Greenberg, S.: Using Unix: Collected traces of 168 users. Research Report, 88/333/45, Department of Computer Science, University of Calgary, Calgary, Canada (1988), http://grouplab.cpsc.ucalgary.ca/papers/
10. Hébert, C., Crémilleux, B.: A unified view of objective interestingness measures. In: Perner, P. (ed.) MLDM 2007. LNCS (LNAI), vol. 4571, pp. 533–547. Springer, Heidelberg (2007)
11. Knobbe, A.J., Ho, E.K.Y.: Pattern teams. In: Fürnkranz, J., Scheffer, T., Spiliopoulou, M. (eds.) PKDD 2006. LNCS (LNAI), vol. 4213, pp. 577–584. Springer, Heidelberg (2006)
12. Lesh, N., Zaki, M.J., Ogihara, M.: Mining features for sequence classification. In: KDD, pp. 342–346 (1999)
13. Li, J., Dong, G., Ramamohanarao, K.: Making use of the most expressive jumping emerging patterns for classification. Knowledge and Information Systems 3(2), 131–145 (2001)
14. Li, J., Wong, L.: Emerging patterns and gene expression data. Genome Informatics 12, 3–13 (2001)
15. Ng, R.T., Lakshmanan, L.V.S., Han, J., Pang, A.: Exploratory mining and pruning optimizations of constrained associations rules. In: ACM SIGMOD 1998, pp. 13–24. ACM Press, New York (1998)
16. Piatetsky-Shapiro, G.: Discovery, analysis, and presentation of strong rules. In: Knowledge Discovery in Databases, pp. 229–248. AAAI/MIT Press (1991)
17. Raïssi, C., Calders, T., Poncelet, P.: Mining conjunctive sequential patterns. Data Min. Knowl. Discov. 17(1), 77–93 (2008)
18. She, R., Chen, F., Wang, K., Ester, M., Gardy, J.L., Brinkman, F.S.L.: Frequent-subsequence-based prediction of outer membrane proteins. In: Getoor, L., Senator, T.E., Domingos, P., Faloutsos, C. (eds.) KDD, pp. 436–445. ACM, New York (2003)
19. Siebes, A., Vreeken, J., van Leeuwen, M.: Item sets that compress. In: Proceedings of the Sixth SIAM International Conference on Data Mining, Bethesda, MD, USA. SIAM, Philadelphia (2006)
20. Smyth, P., Goodman, R.M.: Rule induction using information theory. In: Knowledge Discovery in Databases, pp. 159–176. AAAI Press, Menlo Park (1991)
21. Soulet, A., Crémilleux, B.: Adequate condensed representations of patterns. Data Min. Knowl. Discov. 17(1), 94–110 (2008)
22. Soulet, A., Crémilleux, B., Rioult, F.: Condensed representation of eps and patterns quantified by frequency-based measures. In: KDID 2004, Revised Selected and Invited Paperss, pp. 173–190 (2004)
23. Towell, G.G., Shavlik, J.W., Noordewier, M.O.: Refinement ofapproximate domain theories by knowledge-based neural networks. In: AAAI, pp. 861–866 (1990)
24. Tsai, C.-Y., Shieh, Y.-C.: A change detection method for sequential patterns. Decis. Support Syst. 46(2), 501–511 (2009)
25. Wang, J., Han, J., Li, C.: Frequent closed sequence mining without candidate maintenance. IEEE Trans. Knowl. Data Eng. 19(8), 1042–1056 (2007)
26. Xing, Z., Pei, J., Dong, G., Yu, P.S.: Mining sequence classifiers for early prediction. In: SDM, pp. 644–655 (2008)
27. Yan, X., Han, J., Afshar, R.: Clospan: Mining closed sequential patterns in large databases. In: SDM (2003)