On the discovery of significant temporal rules

Julien Blanchard, Fabrice Guillet, Régis Gras

On the discovery of significant temporal rules

Julien Blanchard, Fabrice Guillet, Régis Gras

***Abstract*— The assessment of the interestingness of sequential rules (generally temporal rules) is a crucial problem in sequence analysis. Due to their unsupervised nature, frequent pattern mining algorithms commonly generate a huge number of rules. However, while association rule interestingness has been widely studied in the literature, there are few measures dedicated to sequential rules. In this article, we propose an original statistical measure for assessing sequential rule interestingness. This measure named Sequential Implication Intensity (SII) evaluates the statistical significance of the rules in comparison with a probabilistic model. Numerical simulations show that SII has unique features for a sequential rule interestingness measure.**

I. Introduction

Frequent pattern discovery in sequences of events[1] (generally temporal sequences) is a major task in data mining. Research work in this domain consists of two approaches:

- discovery of frequent *episodes* in a long sequence of events (approach initiated by Mannila, Toivonen, and Verkamo [14] [13]),
- discovery of frequent *sequential patterns* in a set of sequences of events (approach initiated by Agrawal and Srikant [1] [18]).

The similarity between *episodes* and *sequential patterns* is that they are sequential structures, i.e., a structure defined with an order (partial or total). Such a structure can be, for example:

breakfast then *lunch* then *dinner*

The structure is described by its frequency (or support) and generally by constraints on the event position, like a maximal time window "less than 12 hours stand between *breakfast* and *dinner*" [18] [15] [6] [11] [19].

The difference between *episodes* and *sequential patterns* lies in the measure of their frequency: frequency of *episodes* is an intra-sequence notion [15] [6] [20] [11] [19] [21], while frequency of *sequential patterns* is an inter-sequence notion [1] [18] [17] [22] [9] (see [12] for a synthesis on the different ways of assessing frequency). Thus, the frequent *episode* mining algorithms search for structures which often recur inside a single sequence. On the other hand, the frequent *sequential pattern* mining algorithms search for structures which recur in numerous sequences (independently of the repetitions in each sequence). These last algorithms are actually an extension to sequential data of the frequent itemset mining algorithms, used among other things to generate association rules [2] [10].

Just as the discovery of frequent itemsets leads to the generation of association rules, the discovery of *episodes*/*sequential patterns* is often followed by a sequential rule generation stage which enables predictions to be made within the limits of a time window [18] [15] [6] [17] [22] [20] [11] [19]. Such rules have been used to predict, for example, stock market prices [6] or events in a telecommunication network [15] [19]. A sequential rule can be for instance:

$$breakfast \xrightarrow{6h} lunch$$

This rule means "if one observe *breakfast* then one will certainly observe *lunch* less than 6 hours later".

In this article, we study the assessment of the interestingness of sequential rules. This is a crucial problem in sequence analysis since the frequent pattern mining algorithms are unsupervised and can produce a huge number of rules. While association rule interestingness has been widely studied in the literature (see [3] and [4] for a survey), there are few measures dedicated to sequential rules. In addition to frequency, one mainly finds an index of confidence (or precision) that can be interpreted as an estimation of the conditional probability of the conclusion given the condition [18] [15] [6] [17] [22] [20] [11] [19]. A measure of recall is sometimes used too; it can be interpreted as an estimation of the conditional probability of the condition given the conclusion [20] [19]. In [6] and [11], the authors have proposed an adaptation to sequential rules of the J-measure of Smyth and Goodman, an index coming from mutual information[2]. Finally, an entropic measure is presented in [21] to quantify the information brought by an episode in a sequence, but this approach only deals with episodes and not with prediction rules.

These measures have several limits. First of all, the J-measure is not very intelligible since it gives the same value to a rule $a \xrightarrow{\omega} b$ and to its opposite $a \xrightarrow{\omega} \overline{b}$, whereas these two rules make conflicting predictions. Confidence and recall, vary linearly, which makes them rather sensitive to noise. Above all, these measures increase with the size of the time window chosen. This behavior is absolutely counter-intuitive since a rule with a too large time window does not contribute to making good quality predictions. Indeed, the larger the time window, the greater the probability of

Knowledge & Decision (KOD) research team
LINA – FRE CNRS 2729
Polytechnic School of Nantes University, France
julien.blanchard@univ-nantes.fr

[1]Here we speak about sequences of qualitative variables. Such sequences are generally not called time series.

[2]The J-measure is the part of the average mutual information relative to the truth of the condition.

observing the conclusion which follows the condition in data, and the less significant the rule. Another major problem, which concerns confidence, recall, and J-measure, is that these indexes are all frequency-based: the phenomena studied in data are considered only in a relative way (by means of frequencies) and not in an absolute way (by means of cardinalities). Thus, if a sequence is made longer by repeating it x times one after the other, the indexes do not vary[3]. Statistically, the rules are all the more reliable since they are assessed on long sequences yet. In the end, a good interestingness measure for sequential rules should therefore decrease when the size of the time window is too large, and increase with sequence enlargement. These essential properties have never been highlighted in the literature.

Following the implication intensity for association rules [7] [8] [5], we propose in this article an original statistical measure for assessing sequential rule interestingness. More precisely, this measure evaluates the statistical significance of the rules in comparison with a probabilistic model. The next section is dedicated to the formalization of the notions of *sequential rule*, *example of a rule*, and *counter-example of a rule*, and to the presentation of the new measure, named *Sequential Implication Intensity (SII)*. In section 3, we study SII in several numerical simulations and compare it to other measures.

II. Measuring the statistical significance of sequential rules

A. Context

Our measure, SII, evaluates sequential rules extracted from **one unique sequence**. This approach can be easily generalized to several sequences, for example by computing an average or minimal SII on the set of sequences. Rules are of the form $a \xrightarrow{\omega} b$, where a and b are episodes (these ones can even be structured by intra-episode time constraints). However, in this article, we restrict our study to sequential rules where the episodes a and b are two single events.

The studied sequence is a continuous sequence of instantaneous events (adaptation to discrete sequences is trivial). It is possible that two different events occur at the same time. This amounts to using the same framework as the one introduced by Mannila, Toivonen, and Verkamo [15]. To extract the appropriate cardinalities from the sequence and compute SII, one only needs to apply their episode mining algorithm named Winepi [14] [15] (or one of its variants). In the following, we stand at the post-processing stage by considering that Winepi has already been applied on the sequence, and we directly work on the episode cardinalities that have been discovered. Here again, our approach could be generalized to other kinds of sequences, for which other episode mining algorithms have been proposed. For example, Höppner has studied sequences with time-interval events that have a non-zero duration and can overlap [11].

[3]We consider here that the size of the time window is negligible compared to the size of the sequence, and we leave aside the possible side effects which could make new patterns appear overlapping the end of a sequence and the beginning of the following repeated sequence.

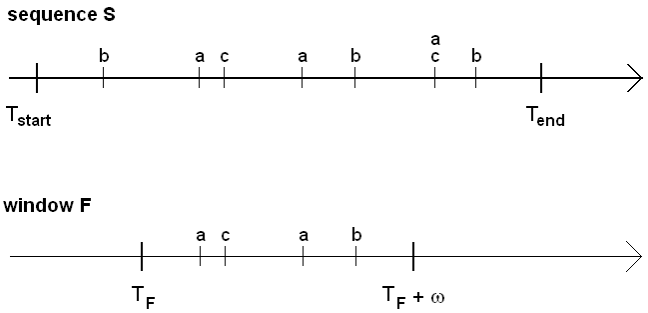


Fig. 1. A sequence S of events from $E = \{a, b, c\}$ and its window F of size ω beginning at T_F.

B. Notations

Let $E = \{a, b, c...\}$ be a finite set of *event types*. An *event* is a couple (e, t) where $e \in E$ is the type of the event and $t \in \mathbb{R}_+$ is the time the event occurred. It must be noted that the term event is often used to refer the event type without reducing intelligibility.

An *event sequence* S observed between the instants T_{start} and T_{end} is a finite series of events

$$S = \Big((e_1, t_1), (e_2, t_2), (e_3, t_3), ...(e_n, t_n)\Big)$$

such that:

$$\forall i \in \{1..n\}, (e_i \in E \wedge t_i \in [T_{start}, T_{end}])$$
$$\forall i \in \{1..n-1\}, t_i \leq t_{i+1}$$
$$\forall (i, j) \in \{1..n\}^2, t_i = t_j \Rightarrow e_i \neq e_j$$

The size of the sequence is $L = T_{end} - T_{start}$.

A *window* on a sequence S is a subsequence of S. For instance, a window F of size $\omega \leq L$ beginning at the instant $t_F \in [T_{start}, T_{end} - \omega]$ contains all the events (e_i, t_i) from S such as $t_F \leq t_i \leq t_F + \omega$.

In the following, we consider a sequence S of events from E.

C. Sequential rules

We establish a formal framework for sequence analysis by defining the notions of *sequential rule*, *example of a rule*, and *counter-example of a rule*. The examples and counter-examples of a sequential rule have never been defined in the literature about sequences.

*Definition 1: A **sequential rule** is a triple (a, b, ω) noted $a \xrightarrow{\omega} b$ where a and b are events of different types and ω is a strictly positive real number. It means: "if an event a appears in the sequence then an event b certainly appears within the next ω time units".*

*Definition 2: The **examples** of a sequential rule $a \xrightarrow{\omega} b$ are the events a which are followed by at least one event*

b within the next ω time units. Therefore the number of examples of the rule is the cardinality noted $n_{ab}(\omega)$:

$$n_{ab}(\omega) = \left|(a,t) \in S \mid \exists (b,t') \in S, 0 \leq t' - t \leq \omega\right|$$

Definition 3: The ***counter-examples*** *of a sequential rule $a \xrightarrow{\omega} b$ are the events a which are not followed by any event b during the next ω time units. Therefore the number of counter-examples of the rule is the cardinality noted $n_{a\bar{b}}(\omega)$:*

$$n_{a\bar{b}}(\omega) = \left|(a,t) \in S \mid \forall (b,t') \in S, (t' < t \vee t' > t + \omega)\right|$$

Contrary to association rules, n_{ab} and $n_{a\bar{b}}$ are not data constants but depend on the parameter ω.

The originality of our approach is that it treats condition and conclusion in very different ways: the events a are used as references for searching the events b, i.e. only the windows which begin by an event a are taken into account. On the contrary, in the literature about sequences, the algorithms like Winepi move a window forward (with a fixed step) over the whole sequence [15]. This method amounts to considering as examples of the sequential rule any window that has an event a followed by b, even if it does not start by en event a. In comparison, our approach is algorithmically less complex.

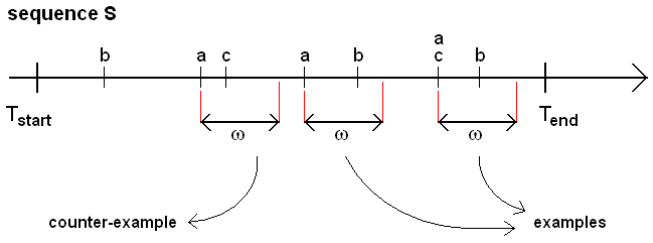


Fig. 2. Among the 3 windows of size ω beginning on events a, one can find 2 examples and 1 counter-example of the rule $a \xrightarrow{\omega} b$.

Let us note n_a the number of events a in the sequence. We have the usual equality $n_a = n_{ab} + n_{a\bar{b}}$. A sequential rule $a \xrightarrow{\omega} b$ is completely described by the quintuple $(n_{ab}(\omega), n_a, n_b, \omega, L)$. The examples of a sequential rule now being defined, we can specify our measure for the frequency of the rules:

Definition 4: The ***frequency*** *of a sequential rule $a \xrightarrow{\omega} b$ is the proportion of examples compared to the size of the sequence:*

$$frequency(a \xrightarrow{\omega} b) = \frac{n_{ab}(\omega)}{L}$$

With these notations, the confidence, recall, and J-measure are given by the following formula:

$$confidence(a \xrightarrow{\omega} b) = \frac{n_{ab}(\omega)}{n_a}$$

$$recall(a \xrightarrow{\omega} b) = \frac{n_{ab}(\omega)}{n_b}$$

$$J\text{-}measure(a \xrightarrow{\omega} b) = \frac{n_{ab}(\omega)}{L} \log_2 \frac{n_{ab}(\omega) L}{n_a n_b} + \frac{n_{a\bar{b}}(\omega)}{L} \log_2 \frac{n_{a\bar{b}}(\omega) L}{n_a (L - n_b)}$$

D. Random model

Following the implication intensity for association rules [8] [5], the sequential implication intensity SII measures the statistical significance of the rules $a \xrightarrow{\omega} b$. To do so, it quantifies the unlikelihood of the smallness of the number of counter-examples $n_{a\bar{b}}(\omega)$ with respect to the independence hypothesis between the types of events a and b. Therefore, in a search for a random model, we suppose that the types of events a and b are independent. Our goal is to determine the distribution of the random variable $\mathcal{N}_{a\bar{b}}$ (number of counter-examples of the rule) given the size L of the sequence, the numbers n_a and n_b of events of types a and b, and the size ω of the time window which is used.

We suppose that the arrival process of the events of type b satisfies the following hypotheses:

- the times between two successive occurrences of b are independent random variables,
- the probability that a b appears during $[t, t + \omega]$ only depends on ω.

Moreover, two events of the same type cannot occur simultaneously in the sequence S (see section II-B). In these conditions, the arrival process of the events of type b is a Poisson process of intensity $\lambda = \frac{n_b}{L}$. So, the number of b appearing in a window of size ω follows Poisson's Law with parameter $\frac{\omega . n_b}{L}$. In particular, the probability that no event of type b appears during ω time units is:

$$p = \mathrm{P}(Poisson(\frac{\omega . n_b}{L}) = 0) = e^{-\frac{\omega}{L} n_b}$$

Therefore, wherever it appears in the sequence, an event a has the fixed probability p of being a counter-example, and $1 - p$ of being an example. Let us repeat n_a times this random experiment to determine the theoretical number of counter-examples $\mathcal{N}_{a\bar{b}}$. If ω is negligible compared to L, then two randomly chosen windows of size ω are not likely to overlap, and we can consider that the n_a repetitions of the experiment are independent. In these conditions, the random variable $\mathcal{N}_{a\bar{b}}$ is binomial with parameters n_a and p:

$$\mathcal{N}_{a\bar{b}} = Binomial(n_a, e^{-\frac{\omega}{L} n_b})$$

When permitted, this binomial distribution can be approximated by another Poisson distribution (even in the case of "weakly dependent" repetitions –see [16]) .

Definition 5: The ***sequential implication intensity (SII)*** *of a rule $a \xrightarrow{\omega} b$ is defined by:*

$$SII(a \xrightarrow{\omega} b) = \mathrm{P}(\mathcal{N}_{a\bar{b}} > n_{a\bar{b}}(\omega))$$

Numerically, we have:

$$SII(a \xrightarrow{\omega} b) = 1 - \sum_{k=0}^{n_{a\bar{b}}(\omega)} \mathrm{C}_{n_a}^{k} (e^{-\frac{\omega}{L} n_b})^k (1 - e^{-\frac{\omega}{L} n_b})^{n_a - k}$$

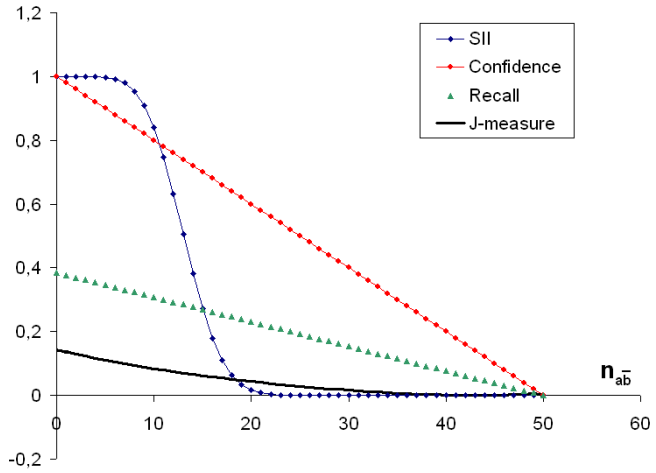


Fig. 3. SII, confidence, recall, and J-mesure w.r.t. the number of counter-examples. ($n_a = 50$, $n_b = 130$, $\omega = 10$, $L = 1000$)

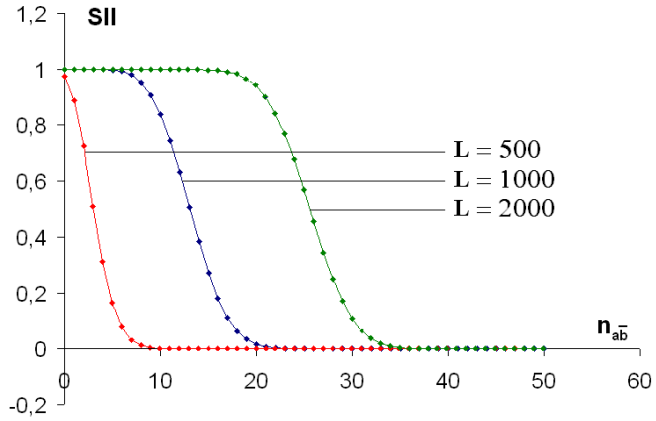


Fig. 4. SII with sequence enlargement. ($n_a = 50$, $n_b = 130$, $\omega = 10$)

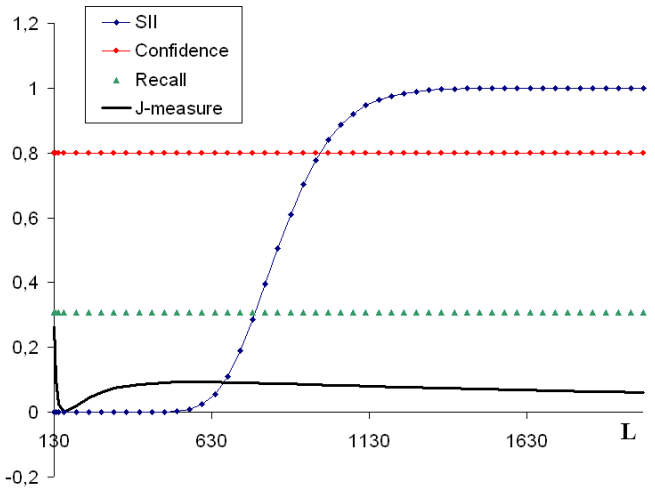


Fig. 5. SII, confidence, recall, and J-mesure with sequence enlargement. ($n_a = 50$, $n_b = 130$, $n_{a\bar{b}} = 10$, $\omega = 10$)

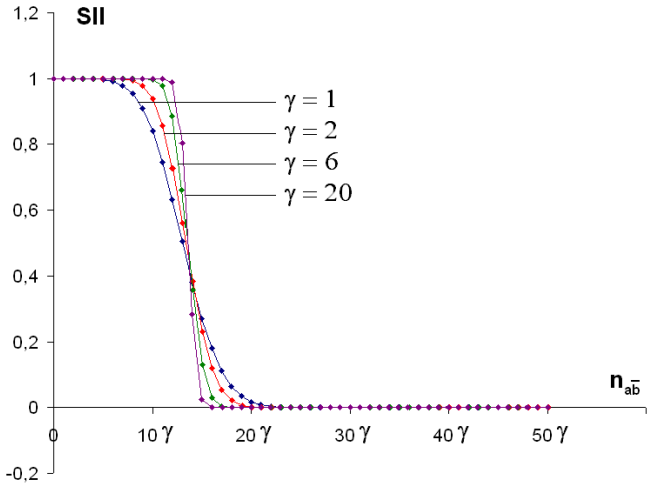


Fig. 6. SII with sequence repetition. ($n_a = 50 \times \gamma$, $n_b = 130 \times \gamma$, $\omega = 10$, $L = 1000 \times \gamma$)

III. Properties and comparisons

SII quantifies the unlikelihood of the smallness of the number of counter-examples $n_{a\bar{b}}(\omega)$ with respect to the independence hypothesis between the types of events a and b. In particular, if $SII(a \xrightarrow{\omega} b)$ is worth 1 or 0, then it is unlikely that the types of event a and b are independent (deviation from independence is significant and oriented in favor of the examples or of the counter-examples). This new index can be seen as the complement to 1 of the p-value of a hypothesis test. However, following the implication intensity [8] [5], the aim here is not testing a hypothesis but actually using it as a reference to evaluate and sort the rules.

In the following, we study SII in several numerical simulations and compare it to confidence, recall, and J-measure. These simulations point out the intuitive properties of a good interestingness measure for sequential rules.

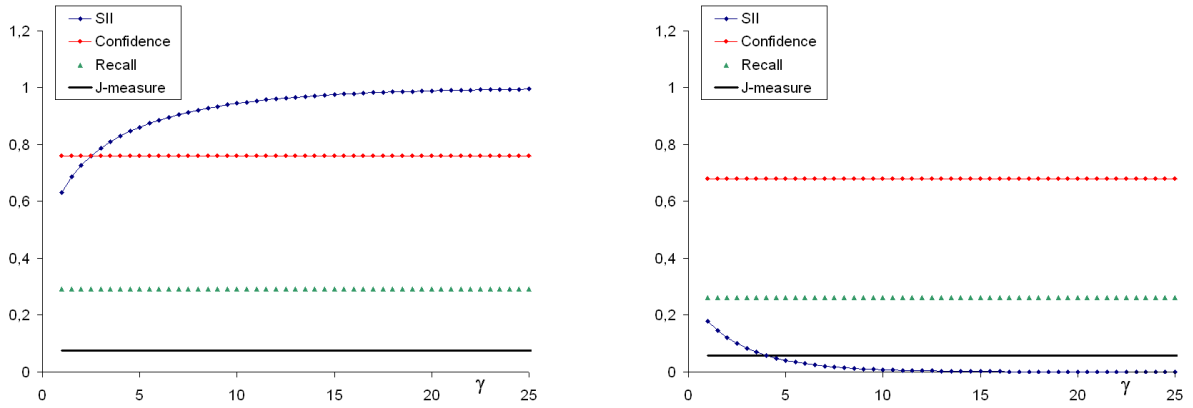
A. Counter-example increase

In this section, we study the measures when the number $n_{a\bar{b}}$ of counter-examples increases (with all other parameters constant). For a rule $a \xrightarrow{\omega} b$, this can be seen as making the events a and b more distant in the sequence while keeping the same numbers of a and b. This operation transforms events a from examples to counter-examples.

Fig. 3 shows that SII clearly distinguishes between acceptable numbers of counter-examples (assigned to values close to 1) and non-acceptable numbers of counter-examples (assigned to values close to 0) with respect to the other parameters n_a, n_b, ω, and L. On the contrary, confidence and recall vary linearly, while J-measure provides very little discriminative power. Due to its entropic nature, the J-measure could even increase when the number of counter-examples increases, which is disturbing for a rule interestingness measure.

B. Sequence enlargement

We call sequence enlargement the operation which makes the sequence longer by adding new events (of new types) at the beginning or at the end. For a rule $a \xrightarrow{\omega} b$, such an operation does not change the cardinalities $n_{ab}(\omega)$ and



(a) $n_{a\bar{b}} = 12 \times \gamma$ (b) $n_{a\bar{b}} = 16 \times \gamma$

Fig. 7. SII, confidence, recall, and J-mesure with sequence repetition. ($n_a = 50 \times \gamma$, $n_b = 130 \times \gamma$, $\omega = 10$, $L = 1000 \times \gamma$)

$n_{a\bar{b}}(\omega)$ since the layout of the events a and b remain the same. Only the size L of the sequence increase.

Fig. 4 shows that SII increases with sequence enlargement. Indeed, for a given number of counter-examples, a rule is more surprising in a long sequence rather than in a short one since the a and b are less likely to be close in a long sequence. On the contrary, measures like confidence and recall remain unchanged since they do not take L into account (see Fig. 5). The J-measure varies with L but only slightly. It can even decrease with L, which is counter-intuitive.

C. Sequence repetition

We call sequence repetition the operation which makes the sequence longer by repeating it γ times one after the other (we leave aside the possible side effects which could make new patterns appear by overlapping the end of a sequence and the beginning of the following repeated sequence). With this operation, the frequencies of the events a and b and the frequencies of the examples and counter-examples remain unchanged.

Fig. 6 shows that the values of SII are more extreme (close to 0 or 1) with sequence repetition. This is due to the statistical nature of the measure. Statistically, a rule is all the more significant when it is assessed on a long sequence with lots of events: the longer the sequence, the more one can trust the imbalance between examples and counter-examples observed in the sequence, and the more one can confirm the good or bad quality of the rule. On the contrary, the frequency-based measures like confidence, recall, and J-measure do not vary with sequence repetition (see Fig. 7).

D. Window enlargement

Window enlargement consists of increasing the size ω of the time window. As the function $n_{a\bar{b}}(\omega)$ is unknown ($n_{a\bar{b}}$ is given by a data mining algorithm, it depends on the data),

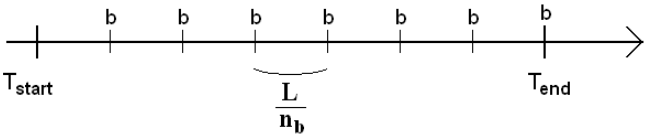


Fig. 8. A sequence where the events b are regularly spread.

we model it in the following way:

$$n_{a\bar{b}}(\omega) = n_a - \frac{n_a n_b}{L}\omega \ , \quad \text{if } \omega \leq \frac{L}{n_b}$$

$$n_{a\bar{b}}(\omega) = 0 \ , \quad \text{otherwise.}$$

This is a simple model, considering that the number of examples observed in the sequence is proportional to ω: $n_{ab}(\omega) = \frac{n_a n_b}{L}\omega$. The formula is based on the following postulates:

- According to definitions 2 and 3, n_{ab} must increase with ω and $n_{a\bar{b}}$ must decrease with ω.
- If $\omega = 0$ then there is no time window, and the data mining algorithm cannot find any example[4]. So we have $n_{ab} = 0$ and $n_{a\bar{b}} = n_a$.
- Let us consider that the events b are regularly spread over the sequence (Fig. 8). If $\omega \geq \frac{L}{n_b}$, then any event a can capture at least one event b within the next ω time units. So we are sure that all the events a are examples, i.e. $n_a = n_{ab}$ and $n_{a\bar{b}} = 0$.

In practice, since the events b are not regularly spread over the sequence, the maximal gap between two consecutive events b can be greater than $\frac{L}{n_b}$. So the threshold $\omega \geq \frac{L}{n_b}$ is not enough to be sure that $n_a = n_{ab}$. This is the reason why we introduce a coefficient k into the function $n_{a\bar{b}}(\omega)$:

$$n_{a\bar{b}}(\omega) = n_a - \frac{n_a n_b}{L}\frac{\omega}{k} \ , \quad \text{if } \omega \leq \frac{kL}{n_b}$$

$$n_{a\bar{b}}(\omega) = 0 \ , \quad \text{otherwise.}$$

[4]We consider that two events a and b occurring at the same time do not make an example.

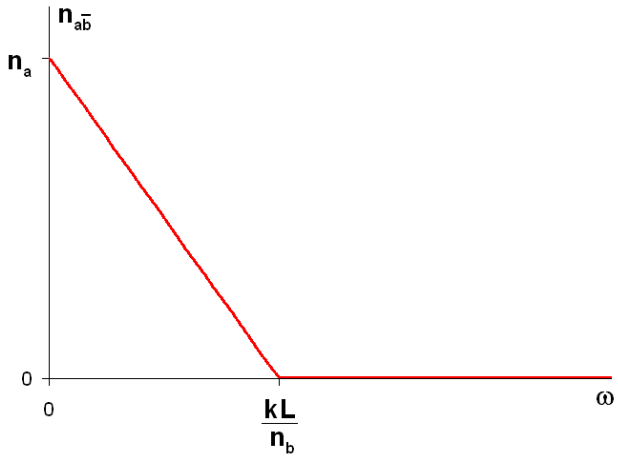


Fig. 9. Model for $n_{a\bar{b}}(\omega)$.

The coefficient k can be seen as a non-uniformity index for the events b in the sequence. We have $k = 1$ only if the events b are regularly spread over the sequence (Fig. 8).

With this model for $n_{a\bar{b}}(\omega)$, we can now study the interestingness measures with regard to ω and k. Several interesting behaviors can be pointed out for SII (see illustration in Fig. 10):

- There exists a range of values for ω which allows SII to be maximized. This is intuitively satisfying[5]. The higher the coefficient k, the smaller the range of values.
- If ω is too large, then $SII = 0$. Indeed, the larger the time window, the greater the probability of observing a given series of events in the sequence, and the less significant the rule.
- As for the small values of ω (before the range of values which maximizes SII):
  - If $k \approx 1$, then n_{ab} increases fast enough with ω to have SII increase (Fig. 10 at the top).
  - If k is larger, then n_{ab} does not increase fast enough with ω. SII decreases until n_{ab} becomes more adequate (Fig. 10 at the bottom).

On the other hand, confidence (idem for recall) increases linearly with ω (see Fig. 11 with a logarithmic scale). Above all, the three measures confidence, recall, and J-measure do not tend to 0 when ω is large[6]. Indeed, these measures depend on ω only through $n_{a\bar{b}}$, i.e. the parameter ω does not explicitly appear in the formulas of the measures. If ω is large enough to capture all the examples, then $n_{a\bar{b}} = 0$ is fixed and the three measures become constant functions (with a good value since there is no counter-example). This behavior is absolutely counter-intuitive. Only SII takes ω explicitly into account and allows rules with too large time window to be discarded.

[5]When using a sequence mining algorithm to discover a specific phenomenon in data, lots of time is spent to find the "right" value for the time window ω.

[6]This does **not** depend on any model chosen for $n_{a\bar{b}}(\omega)$.

IV. Conclusion

In this article, we have studied the assessment of the interestingness of sequential rules. First, we have formalized the notions of *sequential rule*, *example of a rule*, and *counter-example of a rule*. We have then presented the *Sequential Implication Intensity (SII)*, an original statistical measure for assessing sequential rule interestingness. SII evaluates the statistical significance of the rules in comparison with a probabilistic model. Numerical simulations show that SII has interesting features. In particular, SII is the only measure that takes sequence enlargement, sequence repetition, and window enlargement into account in an appropriate way.

To continue this research work, we are developing a rule mining platform for sequence analysis. Experimental studies of SII on real data (Yahoo Finance Stock Exchange data) will be available soon.

References

[1] R. Agrawal and R. Srikant, *Mining sequential patterns*, Proceedings of the international conference on data engineering (ICDE), IEEE Computer Society, 1995, pp. 3–14.

[2] Rakesh Agrawal and Ramakrishnan Srikant, *Fast algorithms for mining association rules*, Proceedings of the twentieth international conference on very large data bases (VLDB 1994) (Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, eds.), Morgan Kaufmann, 1994, pp. 487–499.

[3] Julien Blanchard, Fabrice Guillet, Henri Briand, and Regis Gras, *Assessing rule interestingness with a probabilistic measure of deviation from equilibrium*, Proceedings of the eleventh international symposium on Applied Stochastic Models and Data Analysis ASMDA-2005, ENST, 2005, pp. 191–200.

[4] Julien Blanchard, Fabrice Guillet, Regis Gras, and Henri Briand, *Using information-theoretic measures to assess association rule interestingness*, Proceedings of the fifth IEEE international conference on data mining ICDM'05, IEEE Computer Society, 2005, pp. 66–73.

[5] Julien Blanchard, Pascale Kuntz, Fabrice Guillet, and Régis Gras, *Implication intensity: from the basic statistical definition to the entropic version*, Statistical Data Mining and Knowledge Discovery (Hamparsum Bozdogan, ed.), Chapman and Hall/CRC Press, 2003, chapter 28, pp. 473–485.

[6] Gautam Das, King-Ip Lin, Heikki Mannila, Gopal Renganathan, and Padhraic Smyth, *Rule discovery from time series*, Proceedings of the fourth ACM SIGKDD international conference on knowledge discovery and data mining (Rakesh Agrawal, Paul E. Stolorz, and Gregory Piatetsky-Shapiro, eds.), AAAI Press, 1998, pp. 16–22.

[7] Régis Gras, *L'implication statistique : nouvelle méthode exploratoire de données*, La Pensée Sauvage Editions, 1996, in French.

[8] Sylvie Guillaume, Fabrice Guillet, and Jacques Philippe, *Improving the discovery of association rules with intensity of implication*, Proceedings of the second European conference on principles of data mining and knowledge discovery (PKDD'98) (J.M. Zytkow and M. Quafafou, eds.), Lecture Notes in Artificial Intelligence, vol. 1510, Springer-Verlag, 1998, pp. 318–327.

[9] J. Han, J. Pei, and X. Yan, *Sequential pattern mining by pattern-growth: Principles and extensions*, Recent Advances in Data Mining and Granular Computing (Mathematical Aspects of Knowledge Discovery) (W. W. Chu and T. Y. Lin, eds.), Springer-Verlag, 2005, pp. 183–220.

[10] Jochen Hipp, Ulrich Güntzer, and Gholamreza Nakhaeizadeh, *Algorithms for association rule mining a general survey and comparison*, SIGKDD Explorations **2** (2000), no. 1, 58–64.

[11] F. Höppner, *Learning dependencies in multivariate time series*, Proceedings of the ECAI'02 workshop on knowledge discovery in spatio-temporal data, 2002, pp. 25–31.

[12] Mahesh Joshi, George Karypis, and Vipin Kumar, *A universal formulation of sequential patterns*, Tech. report, University of Minnesota, 1999, TR 99-021.

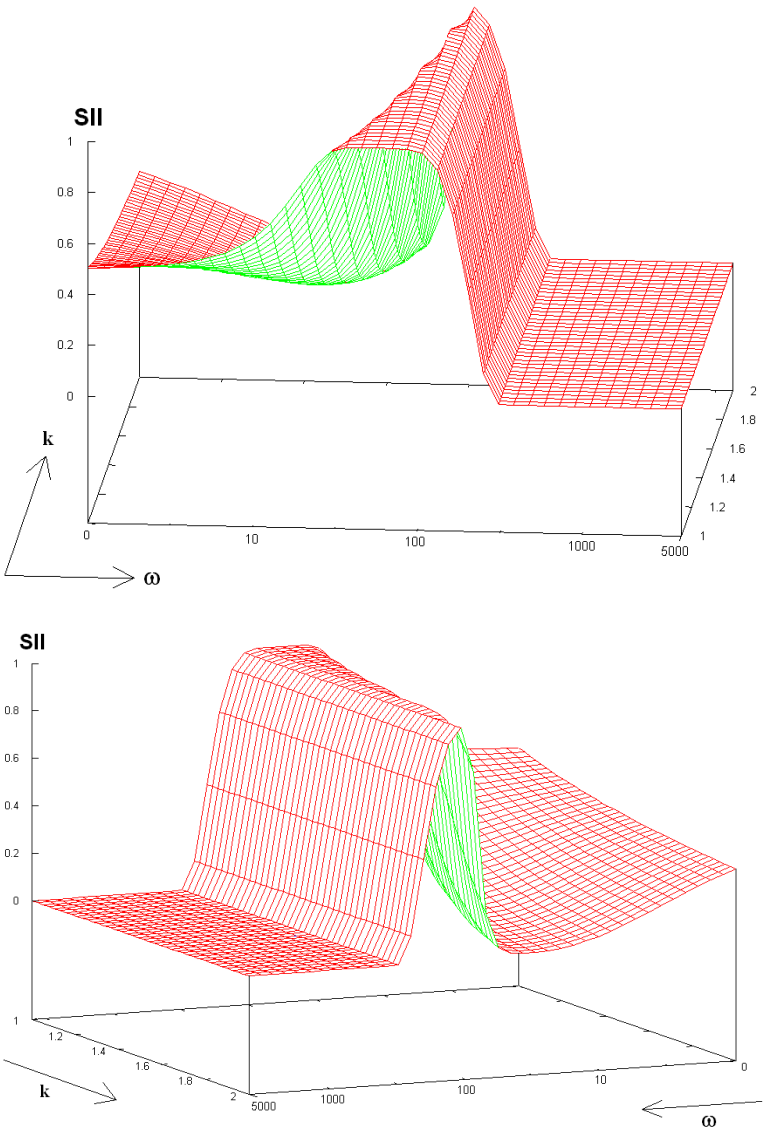


Fig. 10. SII with window enlargement. ($n_a = 50$, $n_b = 100$, $L = 5000$)

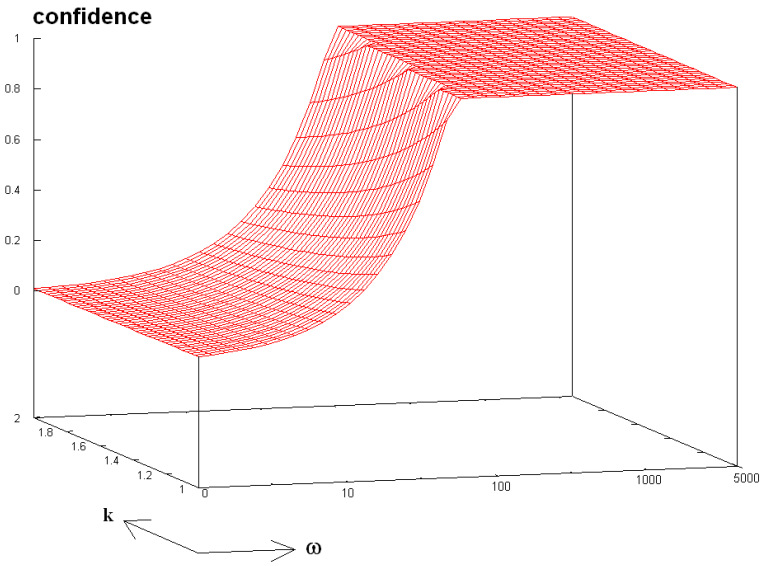


Fig. 11. Confidence with window enlargement. ($n_a = 50$, $n_b = 100$, $L = 5000$)

[13] H. Mannila and H. Toivonen, *Discovering generalized episodes using minimal occurrences*, Proceedings of the second ACM SIGKDD international conference on knowledge discovery and data mining, AAAI Press, 1996, pp. 146–151.

[14] H. Mannila, H. Toivonen, and A. I. Verkamo, *Discovering frequent episodes in sequences*, Proceedings of the first ACM SIGKDD international conference on knowledge discovery and data mining, AAAI Press, 1995, pp. 210–215.

[15] Heikki Mannila, Hannu Toivonen, and A Inkeri Verkamo, *Discovery of frequent episodes in event sequences*, Data Mining and Knowledge Discovery **1** (1997), no. 3, 259–289.

[16] Sheldon M. Ross, *Introduction to probability models*, 2006, 9th edition.

[17] Myra Spiliopoulou, *Managing interesting rules in sequence mining*, PKDD'99: Proceedings of the third European conference on principles of data mining and knowledge discovery, Springer-Verlag, 1999, pp. 554–560.

[18] Ramakrishnan Srikant and Rakesh Agrawal, *Mining sequential patterns: generalizations and performance improvements*, EDBT'96: Proceedings of the fifth International Conference on Extending Database Technology, Springer-Verlag, 1996, pp. 3–17.

[19] Xingzhi Sun, Maria E. Orlowska, and Xiaofang Zhou, *Finding event-oriented patterns in long temporal sequences*, Proceedings of the seventh Pacific-Asia conference on knowledge discovery and data mining (PAKDD2003) (Kyu-Young Whang, Jongwoo Jeon, Kyuseok Shim, and Jaideep Srivastava, eds.), Lecture Notes in Computer Science, vol. 2637, Springer-Verlag, 2003, pp. 15–26.

[20] Gary M. Weiss, *Predicting telecommunication equipment failures from sequences of network alarms*, Handbook of knowledge discovery and data mining, Oxford University Press, Inc., 2002, pp. 891–896.

[21] Jiong Yang, Wei Wang, and Philip S. Yu, *Stamp: On discovery of statistically important pattern repeats in long sequential data*, Proceedings of the third SIAM international conference on data mining (Daniel Barbará and Chandrika Kamath, eds.), SIAM, 2003.

[22] Mohammed J. Zaki, *SPADE: an efficient algorithm for mining frequent sequences*, Machine Learning **42** (2001), no. 1-2, 31–60.