Information Sciences 296 (2015) 175–200

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Statistical significance of episodes with general partial orders

Avinash Achar *, P.S. Sastry

Dept. of Electrical Engineering, Indian Institute of Science, Bangalore, India

ARTICLE INFO

Article history:
Received 15 April 2013
Received in revised form 16 September 2014
Accepted 29 September 2014
Available online 22 October 2014

Keywords:
Episode discovery
Partial order episodes
Non-overlapped frequency
Statistical significance

ABSTRACT

Frequent episode discovery is one of the methods used for temporal pattern discovery in sequential data. An episode is a partially ordered set of nodes with each node associated with an event type. For more than a decade, algorithms existed for episode discovery only when the associated partial order is total (serial episode) or trivial (parallel episode). Recently, the literature has seen algorithms for discovering episodes with general partial orders. In frequent pattern mining, the threshold beyond which a pattern is inferred to be interesting is typically user-defined and arbitrary. One way of addressing this issue in the pattern mining literature has been based on the framework of statistical hypothesis testing. This paper presents a method of assessing statistical significance of episode patterns with general partial orders. A method is proposed to calculate thresholds, on the non-overlapped frequency, beyond which an episode pattern would be inferred to be statistically significant. The method is first explained for the case of injective episodes with general partial orders. An injective episode is one where event-types are not allowed to repeat. Later it is pointed out how the method can be extended to the class of all episodes. The significance threshold calculations for general partial order episodes proposed here also generalize the existing significance results for serial episodes. Through simulations studies, the usefulness of these statistical thresholds in pruning uninteresting patterns is illustrated.

1. Introduction

Over the years, frequent pattern mining has been found useful in many applications of data mining. The field, which emerged with itemset (association rule) mining, has seen many other interesting patterns like sequential patterns, episodes, graphs, trees and so on [18]. In the frequent pattern mining paradigm, a pattern is considered interesting when it occurs often enough, that is, it is *frequent*. However, whether or not a pattern is frequent, depends on arbitrary (user-chosen) thresholds on the frequency (or support) of the pattern. Hence, the issue of choosing these thresholds and assessing the interestingness (or significance) of a pattern using sound statistical principles is important. Typically, significance of patterns is assessed based on ideas from statistical hypothesis testing. Thresholds beyond which patterns are assessed to be significant are typically chosen by rejection of a suitable null hypothesis at a given level of confidence. The null hypothesis that is most often used in data mining context is a hypothesis of independence. For example, in the itemset context an independence hypothesis is that the records or transactions in the database are independent realizations of a collection

* Corresponding author.
E-mail addresses: achar.avinash@gmail.com (A. Achar), sastry@ee.iisc.ernet.in (P.S. Sastry).

http://dx.doi.org/10.1016/j.ins.2014.09.063
0020-0255/© 2014 Elsevier Inc. All rights reserved.

of independent binary random variables, where each binary random variable codes for the presence or absence of an item. In this paper, this issue of statistical significance in the context of frequent episode mining is addressed [32].

1.1. Statistical methods in pattern discovery

Hypothesis testing problem is to find a decision rule (also called test) which can lead to a conclusion about the underlying true hypothesis (out of the two hypotheses – the null and the alternate) using an observed sample. The function of the sample or data that is used for decision making is called the test statistic. The set of values of the test statistic for which the null hypothesis is rejected is called the critical region. The error where one decides to reject the null hypothesis, but the null is actually true, is called the type-1 error. The probability of type-1 error is called the significance level of the test.

In pattern mining literature, significance of patterns is mostly assessed by rejection of a suitable null hypothesis, with no alternate hypothesis considered. Given that the patterns that occur often enough are considered interesting in frequent pattern mining, a natural choice for the test statistic is the frequency of the pattern. Hence the test will be of the form: if the frequency is above a (calculated) threshold then reject the null hypothesis and conclude that the pattern is statistically significant. In general, for any given null hypothesis, the distribution of the frequency of an episode would, in general, depend on the structure of the episode and hence different episodes would have different frequency thresholds for being significant. However, most data mining algorithms function with a single frequency threshold. Thus, often the thresholds calculated using statistical analysis are used for pruning the set of frequent episodes discovered (as a post-processing). Sometimes it may be possible to make this process more efficient. For example, if the thresholds for statistical significance are the same for all episodes of a given size (which could be the case, e.g., for serial episodes) then one can use these thresholds in an apriori-style data mining algorithm by having different support-thresholds for different passes of the algorithm.

In the context of frequent itemsets, an interesting statistical significance analysis is presented in [8]. Here the null hypothesis is one of independence and the statistical test is essentially to establish whether or not the corresponding binary random variables of the itemset are independent. Recently a method to assess significance of itemsets based on rejecting a composite null hypothesis that is more general than that of independence, was proposed in [37]. Statistical significance analysis of sequential patterns [5,10,21] (which are essentially sequences of itemsets) has also been addressed, e.g., in [24].

In this paper, the focus would be on statistical significance assessment of episode patterns based on their frequency. Frequent episode discovery [32] is a framework for discovering temporal patterns in symbolic time series data, with applications in several domains like manufacturing [27,41], telecommunication [32,19], WWW [30], biology [7,35], finance [34], intrusion detection [31,42], text mining [23] etc. The data in this framework is a single long time-ordered sequence of events and each temporal pattern (called an episode) is essentially a small, partially ordered collection of event-types. The partial order in the episode constrains the time-order in which events (of appropriate types) should appear in the data, in order for the events to constitute an occurrence of the episode. There have been many notions of frequency proposed in the episodes literature (see [1] for a discussion on different frequency measures used in the episodes framework). Every frequency measure captures some notion of how often an episode occurs in the data. Given a frequency measure, the computational task is to discover all episodes whose frequency in the data exceeds a user-defined threshold. For many years, the algorithms for discovering frequent episodes restricted themselves to only *serial* episodes (where the partial order is a total order) or only *parallel* episodes (where the partial order is trivial) [32,9,28,27,35,45]. Algorithms for discovering episodes with general partial orders have recently been proposed [39,2].

In the context of episodes, an often used null model is an i.i.d. (independent identically distributed) model. This null hypothesis states that the event sequence consists of i.i.d. realizations of the random variable that takes values from the set of all event-types, $\mathcal{E}$. The probability of each event-type, E, is denoted by p_E and $\sum_{E\in\mathcal{E}} p_E = 1$. If each $p_E = 1/|\mathcal{E}|$, then the model is referred to as a uniform i.i.d. model. For any other symbol probabilities, we refer to it as a non-uniform i.i.d. model. Apart from the i.i.d. null hypothesis, people have also considered a null hypothesis where the underlying event sequence, is assumed to form a Markov chain. The literature has seen statistical significance assessment of serial episode patterns under some of the frequency notions. Under the windows based frequency and i.i.d. null hypothesis, [16] show that the random variable representing frequency of a serial episode (after being suitably normalized) is asymptotically normal. Under such a Gaussian approximation, they calculate significance thresholds for a given probability of type-1 error. They extend this to general partial order episodes (including parallel episodes) in [6]. The extension of these results for a Markov null hypothesis is considered in [17]. Statistical significance results under the non-overlapped frequency and uniform i.i.d. null model for serial episodes are reported in [28]. A method to assess the significance of general partial order episodes based on minimal occurrences based frequency has been proposed in [38]. This method uses span information of the minimal windows present in the data as the test statistic rather than the frequency of episodes. There is also some recent work on assessing significance of serial episodes with a fixed gap constraint under a composite null hypothesis based on some conditional probability bounds [36].

As explained above, there are approaches for assessing significance of general partial order episodes under the windows based and minimal windows based frequency. In this paper, statistical significance of episodes with general partial orders under the non-overlapped frequency is considered.

The approach we follow for analyzing statistical significance of discovered episodes under non-overlapped frequency can be summarized as follows. Most of the algorithms for frequent episode discovery employ (implicitly) some finite state automata (FSA) for recognizing the occurrence of an episode [1]. These automata make state transitions based on the input