A Framework for Mining Sequential Patterns from Spatio-Temporal Event Datasets

Paper by Yan Huang, Liqin Zhang, and Pusheng Zhang
Narrative by Anuj Karpatne, Vijay Borra

Problem Statement

The paper addresses the problem of mining sequentially occurring patterns in a spatio-temporal domain by looking at a large record of spatio-temporal events. A spatio-temporal event has several attributes associated with it such as event ID, time, location and 'event type'. Event types are the class labels associated with the events, that characterize the behaviour of the event. In this context, the problem of mining spatio-temporal sequential patterns refers to finding sequences of 'event types' that happen serially in time, with one leading to another, depicting a 'chain reaction'. Such sequential patterns occur frequently in real-world scenarios where events follow other events in space and time. Some of its application domains include earth science, epidemiology, ecology and climatology.

The first challenge in addressing the above problem is in developing a scoring mechanism to assess the significance of a given pattern in a given dataset $D$, and then exploit this significance scoring scheme to come up with an algorithmic design for mining significant patterns. The interest measure may or may not preserve downward closure property, thus requiring the development of newer algorithms to account for this issue. Also, while dealing with large databases of events, memory requirement of the algorithm is a constraint which need to be addressed by partitioning spatio-temporal datasets.

Major Contributions

There are 2 major contributions of this paper:

- Deriving a framework for assessing the significance of a sequential spatio-temporal pattern using 2 interest measures – density ratio (for sequences of size utmost 2), and sequence index (an extension of density ratio for sequences with size greater than 2). Density ratio takes into account the statistical occurrence of a given event type in the neighbourhood of another event type as compared to the distribution of the given event type in the entire space.
- Using the sequence index as a scoring measure, the paper proposes a novel algorithm, STS-Miner, for mining significant sequential patterns. Since the sequence index does not guarantee the downward closure property, the algorithm utilizes the weak anti-monotone property of sequence index to develop a depth-first expand based algorithmic model for mining significant event type patterns. Further, they extend the algorithm to develop Slicing-STS-Miner, which utilizes temporal slicing to partition the data set into overlapping slices according to time when the number of events is too large to be processed in memory. The algorithmic challenges associated with creating partitions using temporal slicing is addressed in Slicing-STS-Miner by keeping an account of the crossing events in each overlapping interval. Computational complexity of both STS-Miner and Slicing-STS-Miner has been stated.
Novelty of the work

Previous approaches in the area of spatio-temporal sequential pattern mining can be broadly categorized into the following two categories, each with their own inherent limitations:

- **Sequential pattern mining in the market-basket data analysis:** In this approach, they extend association analysis based methods used in market basket data for identifying sequential patterns in scenarios where the events are discrete and can be considered as a transaction in time. Examples of such dataset types include Web log click streams, DNA sequences and medical treatments. However the notion of a 'transaction', which is crucial in the association analysis framework, does not apply in spatio-temporal domain where events cannot be represented as transactions as space and time is continuous. Partitioning the spatio-temporal space for 'transactionization' of the data in such cases would lead to the loss of spatial and temporal relationships in the events around the partitions.

- **Mining trajectory patterns in spatio-temporal data:** The trajectory of a moving object is typically a collection of consecutive spatial signatures at different time stamps. Mining frequent patterns in the trajectory data of different moving objects might reveal insights into the underlying travelling patterns of the objects. Since trajectory data is a collection of the locations visited by the same object, trajectory analysis can only be applied if the trajectories have been provided a priori, which is not relevant in the context of spatio-temporal data.

The proposed model

An event $e_0$ is said to follow another event $e$, if and only if $e_0$ is in the spatio-temporal neighbourhood of $e$. A simple spatio-temporal neighbourhood constructs an enclosing boundary in space and forward in time using constraints in both space and time. Neighbourhood of an event discretely measures the occurrence of another event in the vicinity of the given event in both space and time. Time-varying neighbourhoods have been proposed as alternate neighbourhood definitions in the paper, to account for more complex scenarios.

Using this neighbourhood definition, the density ratio of an event type $E'$ to event type $E$ is defined as the average density of events in event type $E'$ in the neighbourhood of $E$ as compared to the overall density of events in event type $E'$ in the universal space. Event type $E'$ is said to follow $E$ if the density ratio is greater than 1, $E$ repels $E'$ if density ratio is less than 1, and event types $E'$ and $E$ are independent if density ratio is equal to 1.

For a given event sequence $S$ of size $k$, the tail event set of sequence $S$ is the set of events of type $S(k)$ that participate in at least one of the event sequences of $S$. Let $S'$ be the subsequence of $S$ of size $k-1$, including event types $S(1)$ to $S(k-1)$. The sequence index for $S$ is then defined as:

- Density Ratio of $S(1)$ to $S(2)$, if size of the sequence is 2
- Minimum of (Sequence Index of $S'$, Density Ratio of tail event set of $S'$ and $S(k)$), otherwise

Such a recursive definition of sequence index measures the impact of the (k-1)-size subsequence on a given event type, rather than considering the impact of the (k-1)th event type to the given event type. The minimum function used in the definition of sequence index ensures that a given sequential pattern is only as strong as the weakest link in the pattern. This takes care in rejecting spurious sequential patterns where the significance of an intermediate follow relation is small.

Since the sequence index is not anti-monotonic, it does not guarantee that every subsequence of a significant sequence $S$ is necessarily significant. For example, if $AB$ is significant and $BC$ is not
significant, ABC may or may not be significant. This can be attributed to the fact that we are looking at the density ratio of tail event of B to C in computing the sequence index, as opposed to the density ratio of event types B and C. Such non-adherence to the downward closure property motivates to look for newer algorithms to address this interest measure property. However, the sequence index does satisfy weak anti-monotone property which can be exploited in the development of the algorithm. The weak anti-monotone property states that all subsequences of a significant sequence S, which begins from S(1) and ends at S(k) where k is less than the size of sequence S, are significant.

Utilizing the weak anti-monotone property, a depth first expand based algorithm is proposed, termed as STS-Miner. It starts with an empty sequence and generates all possible candidate sequences of size (k+1) by attaching one more event type to a significant sequence of size k. A sequential pattern tree is maintained where each node represents a pattern and each edge represents an expansion of the pattern by appending another event type. The depth of a node in this pattern tree represents the size of the current pattern. At every step of expansion at one of the leaf nodes, density ratio computations are performed with the current pattern and all possible event types that can be appended to the pattern. If the resulting pattern is significant, the tree is expanded further in a depth-first approach. However, if the resulting pattern is non-significant, the pattern is declared as a terminal node, hence pruning the search space appropriately.

For mining sequential patterns in large datasets which cannot be placed in memory taken as a whole, a variant of the algorithm is developed, termed as Slicing-STS-Miner. In this method, they partition the large dataset into temporal slices of events, using the unidirectional property of time. At every step, the algorithm processes a single slice of event dataset which can fit into the current memory. Using temporal-slicing ensures that each slice is processed only once, the algorithm proceeds forward in time, processing one slice in time, in a piece-meal fashion. However, creating temporal-partitions of the data using overlaps between the time-slices poses the following two additional challenges to be addressed in the intermediate overlapping regions:

- Sequences falling in the overlapping areas of two consecutive slices might duplicate as they can be counted twice in both the slices.
- Sequences might get broken erroneously at the boundaries of consecutive slices, since the dataset has been partitioned.

Keeping track of all the event sequences that visit the overlapping areas of two consecutive slices for addressing the above 2 problems is computationally tedious. Hence, Slicing-STS-Miner introduces the concept of a crossing tail event set, which is structured as a queue containing all the overlapping patterns currently in memory between two consecutive slices, which need to be addressed first. A depth-first expand method in a similar pattern tree is thus developed using the concept of crossing tail event set. However, pruning cannot be done as an intermediate step in the new algorithm until we have observed the influences of a particular event type on another in all times, and hence is done as a final operation.

**Validation Methodology and results**

Computational complexity analysis for both in-memory processing and for large datasets has been performed for both STS-Miner and Slicing-STS-Miner. The computational costs and results on synthetic datasets reveal that STS-Miner performs faster in the case when dataset is small and can be processed in-memory. However, Slicing-STS-Miner provides significant reduction in computational time when the dataset is large and partitioning the dataset is necessary considering the memory constraints.
Performance comparisons were performed on synthetic datasets and the effects of varying the parameters in data generation was studied using simulations. Results on climate data were also obtained by looking at the event data of temperature, evaporation, precipitation and other climate variables. Events in the time series of such climate variables were defined as change points with statistically significant abnormality from the entire time series. Some interesting patterns that were mined by using the climate data in this fashion are: High Temperature leads to High Evaporation, Low Evaporation leads to High Temperature, and Low Evaporation leads to High Temperature leads to High Evaporation.

Assumptions

- The density ratio for independently occurring event type pairs is assumed to be 1, which corresponds to the same frequency of occurrence of an event type inside and outside the neighbourhood of another event type. The statistical sequence of this assumption can however be verified.
- For each event, a continuous neighbourhood in space and time is developed around the event (the spatial size of the neighbourhood may vary with time). However, this may not account for the temporal seasonality observed in spatio-temporal datasets, where neighbourhood may consist of disjoint spaces scattered around the event. For example, climate data follows a cyclical pattern in time, where the climate variables may exhibit similar trends in the same month or season of the year. In such cases, the neighbourhood definition of a particular event can be cyclical in time and may vary depending on the season category it falls into. Such disconnected neighbourhoods in time can be addressed by considering disconnected spatio-temporal neighbourhoods.
- In the computation of density ratio between two events, a count of the occurrence of events in the neighbourhood of a given event is considered. This disregards the quantity of closeness of a particular event to another event by considering discrete binary values.
- A serial ordering of event types in a particular sequential pattern is assumed. A linear model for sequential patterns representing a chain reaction may not always hold true as at a given time stamp, an event type can give rise to a number of event types in its neighbourhood. Hence, there may be number of event type sequences that can exist between a given event and a different event. The influence of an event to another event then should be accounted for by considering all the event type sequences possibly occurring between them, considered as a whole. The use of linear patterns of event types to model chain reaction phenomena in such cases would be difficult.

Suggestions for Improvement

- For finding the statistical significance in the occurrence of density ratio = 1 between 2 event types, a rigorous and complete approach would be to used Monte Carlo simulations between independently generated events and analyze their density ratios over a number of runs.
- For developing discontinuous neighbourhoods around an event so as to account for the temporal seasonality in the data, better neighbourhood definitions can be developed which tries to capture the clustering of events around a given event in space and in cyclical time. Possible ways to address this problem can be using basis functions in time or kernel transformations, which creates a mapping of the data in a space where neighbourhoods are continuous.
- Real valued spatio-temporal distance measure values can be used in the computation of the density ratio between two given events and the sequence index in the general case. A real-valued alternate definition of a density ratio measure can be developed to address this problem.
- Dynamically expanding or contracting neighbourhood can be used for modelling more complex real-world scenarios where the neighbourhood definitions might change with the event type, the spatial location and the time the event occurred.
- Instead of a linear model, a network-based model can be developed for quantizing the influence of a particular event on another event when a number of event type sequences are possible between the two event types. Graph theoretic measures can be developed for measuring the significance of a particular graph pattern in this scenario, as an analog to the sequence index in the linear case. Network-analysis is required for mining such significant patterns in the new graph representation.
- Since different event types might have different influences on the same event type, occurrence of an event in different event neighbourhoods might have different significance. Some events may follow slowly but can lead to a fast spread of other events around it. In such cases, taking the minimum may penalize the sequence more than is required, and we may loose some meaningful patterns by doing so. A better interest measure between event pairs which accounts for this problem can be developed which may not require the minimum interest measure to be chosen for a sequence. If the influences between different event type pairs is known a priori, a weighted calculation of neighbourhood computation can be used for different event type pairs.
- A more detailed and clear time complexity and space complexity analysis of both the algorithms can be performed for both in-memory processing and for large datasets.