Dakota State University

Beadle Scholar

Faculty Research & Publications   College of Business and Information Systems

2013

Discovering Predictive Event Sequences in Criminal Careers

Carl A. Janzen
Dakota State University

Amit Deokar
Dakota State University

Omar F. El-Gayar
Dakota State University

Follow this and additional works at: https://scholar.dsu.edu/bispapers

Recommended Citation

Discovering Predictive Event Sequences in Criminal Careers

Carl A. Janzen, Amit Deokar and Omar El-Gayar

Abstract—**In this work, we consider the problem of predicting criminal behavior, and propose a method for discovering predictive patterns in criminal histories. Quantitative criminal career analysis typically involves clustering individuals according to frequency of a particular event type over time, using cluster membership as a basis for comparison. We demonstrate the effectiveness of hazard pattern mining for the discovery of relationships between different types of events that may occur in criminal careers. Hazard pattern mining is an extension of event sequence mining, with the additional restriction that each event in the pattern is the first subsequent event of the specified type. This restriction facilitates application of established time based measures such as those used in survival analysis. We evaluate hazard patterns using a relative risk model and an accelerated failure time model. The results show that hazard patterns can reliably capture unexpected relationships between events of different types.**

Index Terms— **Predictive analytics, Event sequence mining, Criminal behavior prediction**

I. INTRODUCTION

WHEN EVALUATING alternatives for sentencing policy and rehabilitation programs, there is a recurring question of whether or not existing approaches are effective over the long term. One popular approach for quantitative analysis of criminal careers is to cluster offenders according to their offending patterns over time. This approach, called group trajectory modeling, usually results in an offender typology grouping consisting of two to four categories, to which descriptive labels are attached (e.g., short term juvenile, long term chronic offender). Comparisons between these groupings are then made, with attention to correcting for selection bias and exposure time. The concern of selection bias arises when the treatment of interest, such as arrest and incarceration, cannot be randomly assigned. Thus, treatment outcomes may be reflective of factors that influenced assignment to treatment, such as an individual propensity to commit a crime. Corrections for exposure time, or street time, are intended to address the changes in opportunity that may be expected when an individual is incarcerated. This forms the basis for state of the art quantitative studies designed to measure the long-term

effectiveness of a particular program of treatment, such as arrest and incarceration. However, approaches based on clustering individuals according to rate of certain events have not yet addressed how an arbitrary number of different types of events throughout the criminal career may affect the possibility of future offenses. Key events in a criminal career include arrest, conviction, sentencing, parole, and discharge. Each of these may be further broken down into sub-types. Since existing quantitative analyses do not facilitate ad-hoc discovery of relationships between events of many different types, unexpected relationships between various different event types remain undiscovered.

Event sequence mining can be used to discover patterns consisting of many different types of events. However, a number of challenges arise with the use of existing measures of interest when used to describe predictive relationships. The most fundamental measure of interest for event sequence patterns is support. This measure indicates the number of pattern occurrences, and is borrowed from association rule mining. For each identified support counting method, at least one of the following limitations applies: (A) length of patterns influences support counting (B) an occurrence may or may not be counted depending on the characteristics of other occurrences of the same pattern (non-independence), and (C) unrelated sub-pattern occurrences unduly inflate support counts of some patterns. These problems do not arise in association rule mining, where there is no dimension of time. A related challenge arises during the analysis of a partial event stream or an event stream with censored observations (observations that are unknown because data is missing or because the observation period ended before the event may have occurred). The challenges raised above need to be addressed in a manner that specifically takes into account the nuances that come with the introduction of the dimension of time.

Insofar as there is an interest in discovering relationships between events of multiple different types, there is a need for a method for the ad-hoc discovery of such relationships. Measures based on occurrences of these patterns should not be unduly affected by pattern length, other occurrences of the same pattern, or unrelated occurrences of sub-patterns. This article includes the following the key contributions from the domain and methodology standpoints. For the criminology domain, we demonstrate that hazard patterns based on occurrences of distinct events can be used to make a statement about expected changes in the probability of certain future events as well as expected changes in time to event. For the

C. A. Janzen is with University of the Fraser Valley, Abbotsford, BC V2V 7B1 Canada (e-mail: carl.janzen@ufv.ca).

A. Deokar and O. El-Gayar are with Dakota State University, Madison SD 57042 USA (e-mail: {amit.deokar, omar.el-gayar}@dsu.edu).

event sequence mining methodology, we address limitations

$$s: <(p, 1), (b, 2), (b, 3), (b, 4), (c, 5), (c, 6), (p, 7)>$$

b: burglary, c: conviction, p: parole

Fig. 1. Illustrative event sequence database.

that apply to existing pattern support counting methods, and demonstrate how event hazard patterns address these limitations. We evaluate the usefulness of the event hazard patterns from real data using two time-based models.

The remainder of the paper is organized as follows. In Section II, we introduce the literature in the problem context and highlight unaddressed challenges involved in criminal career analysis and event sequence mining. The gaps identified in this section form the motivation for our design. In Section III, we define the objectives of a solution. These objectives form the guidelines for our evaluation. Section IV includes the design and development of the core algorithms and data structures. In section V, we demonstrate and evaluate the proposed solution, and finally in section VI we conclude with some implications and directions for future research.

II. PROBLEM IDENTIFICATION AND MOTIVATION

In this section, we discuss the problem context and the motivation for this work. We provide a review of relevant criminology literature with attention to predictive patterns in criminal careers. We then provide an overview of literature related to event sequence mining and note the needed developments for effective prediction of events in criminal careers.

A. Domain: Criminal Career Analysis

There are a few notable studies that have addressed the challenge of making long-term predictions about criminal history event patterns using a combination of group trajectory modeling and predictive indicators. The group trajectory modeling technique was introduced in [1]. This technique involves clustering offenders into trajectory groups according to offense rate over a period of time. The following three recent studies involve the use of this method.

Group trajectory modeling was used in [4] to cluster individuals trajectory groups, with the goal of predicting membership in chronic (life-long offender) or high rate (frequent offender) groups. Demographic variables as well as the number of early juvenile offenses were considered as candidate predictors of membership in these groups. The sample consisted of all prisoners convicted in the Netherlands in 1977. However, there were no risk factors that were found to be good predictors of trajectory group membership.

The same method was also used to cluster the members of a cohort of adolescent boys in Montreal into groups, in a study examining the effects of adolescent first-time gang joining at the age of 14 [3]. In this case, propensity score matching was used to balance the treatment (joiners) and control (non-joiners). Propensity scores are calculated based on known predictors of group membership, and comparisons between the two groups are made only between individuals with matching propensity scores. The effect of first-time adolescent gang membership at age 14 was associated with a short-term increase in violence, but no other effect was observed.

In [2], group trajectory modeling formed part of a strategy to predict increasing or decreasing offense rate following incarceration, in a cohort of American prisoners released from state prisons in 1994. The researchers included a variable to represent the heterogeneity of the individual offense history in relation to the rest of the trajectory group. Individual offense rate micro-trajectories were estimated for each released prisoner. After a 3 year follow-up period, 40% of the prisoners had an offense rate that was significantly lower than estimated, and 4% of the prisoners had an offense rate that was significantly higher than estimated. However, the analysis did not address arrest hazard beyond the first post-release arrest, or the different types of subsequent events that may occur.

In addition to group trajectory based approaches, where behavior is modeled according to group characteristics, a number of researchers have focused on the predicting the location of the crime. One example of such work is the Blue CRUSH system used by Memphis police [5]. This system is designed to direct enforcement efforts to geographical areas where there is a high likelihood of a crime. Another case is the prediction of hotspots using data from monthly crime reports [6]. Hotspot prediction is based on aggregate figures where the unit of observation is geographical, such as a district. Although the history of a particular area provides useful information for predictive analytics, this approach does not take the histories of individuals into account.

Individual criminal histories are comprised of discrete event occurrences of various types along a timeline, often separated by long periods for which no events of interest occur. Behavior patterns from similar data have been successfully captured using event sequence mining approaches. In [7], event sequence mining was used to effectively capture patterns involving the type and order of activities in a door event log. Event sequence patterns are patterns of events that frequently occur in the same order. These patterns were used to identify five cluster groups within a building, three of which exhibited a strong group membership. However, as far as we know, there is no work applying event sequence mining to the problem of predicting the behavior of individuals.

B. Methodology: Event Sequence Mining

The use of event sequences for predictive modeling poses some unique challenges. Since event sequence mining is an extension of association rule mining, measures of interest commonly used in association rule mining are natural candidates for use in event sequence mining. Two common examples of such measures from association rule mining are support and confidence [8]. However, the use of such measures in event sequence mining is complicated by some fundamental differences between association rule mining and event sequence mining brought about by the dimension of time.

Support is a measure of pattern occurrence frequency, usually expressed as a count. *Confidence* is a measure of association between occurrences of an antecedent pattern and occurrences of a consequent pattern and is calculated as support of consequent / support of antecedent. Confidence that approaches 1.0 shows that when the antecedent is present, the consequent is expected to also be present. Thus, the presence of the antecedent might be used to determine the probability that the consequent is also present.

In the event sequence database in Fig. 1, one individual event sequence is represented. Each event occurrence is associated with the number of months since some point in the past. For our discussion of support counting, we will not consider censored events or relationships between events in one sequence and events in another sequence.

There are two main approaches to event sequence mining: sequence mining, and frequent episode mining. With sequence mining, pattern support is based on the number of input sequences that contain at least one occurrence of a given pattern. With frequent episode mining, it is the prevalence of the pattern without respect to different input sequences that determines support. Since we are looking for predictive relationships within pattern occurrences, we focus on the frequent episode mining approach.

We can apply frequent episode mining to the event database in Fig. 1. Discussion of frequent episode mining with window based support counting can be found in [9]. Using this technique, the event database is subdivided into all possible windows of some specified size ω. Support count is based on the number of fixed size windows that contain at least one pattern occurrence. Using a window size of five months, we see that there are four windows that contain at least one occurrence of *b* and there are two windows that contain *b* followed by *p*. A simple calculation of confidence gives us a 50% confidence that *b* leads to *p* within five months. In this case, the discovered relationship is as follows: 50% of windows of opportunity that contain a burglary event also contain a subsequent parole event. Note that this does not mean that 50% of burglaries are followed by parole. The relationship is with respect to the windows of opportunity.

A number of alternative methods of support counting have been explored in addition to window-based counting. Examples of support counting also include: minimal occurrence based, non-interleaved, non-overlapping, head frequency, total frequency, and distinct occurrence based. Some of these are also commonly combined with an expiry

TABLE I
SUMMARY OF DOMAIN RELATED GAPS

Approach	Gaps
Group Trajectory Modeling [1-4]	Models expected behavior of individuals over time but does not use multiple event types.
Hotspot based analysis [5, 6]	Relates aggregate counts per geographic region, but does not address individual histories
Event Sequence Mining [7]	Clustering shown to be useful but no examples of event prediction found.

time constraint. For a comprehensive discussion of these variations in support counting, including window-based counting see [10].

However, the use of these support counting methods for event sequences is hampered by counting that is unduly influenced by pattern length, non-independence of pattern occurrences, and the inclusion of unrelated occurrences. These limitations are detailed in Section IV-B.

Event hazard patterns do not have the above mentioned limitations. These patterns are a specialization of event sequence patterns and are comprised of frequently occurring event sequences, wherein each event in the event sequence is the first subsequent occurrence of that event type [11].

Table II contains event hazard patterns based on the contrived event sequence in Fig. 1. The first burglary charge following parole release leads to a higher proportion of subsequent burglary charges when compared to the remaining cases. Note that there are three opportunities for a burglary charge to be repeated. However, after accounting for the one burglary charge that immediately follows a parole release, only two remain. Thus we have a 100% confidence that parole followed by burglary will lead to more burglary, but we only have 50% confidence that subsequent burglaries will do the same (1/1 instead of 1/2 for the remaining burglary charges). Naturally, this does not give us an indicator of generalizability nor does it account for censored observations. We will address each of these in Section IV.

Since hazard patterns incorporate information about the interval that precedes the first occurrence of each subsequent event, we expect them to be well suited for time-based analysis. Time-based models are well suited for addressing ordering of events, and include methods to deal with censored events.

Two complementary time-based options are relative risk ratio (RR) and accelerated failure time (AFT) models. RR is an indicator of treatment impact that relates the number of failures in the treatment group to the number of failures in the control group [12], [13]. In contrast, AFT models the relationship between the expected time before failure in the treatment group, relative to the same in the control group [14].

There is a substantial body of literature in the field of developmental criminology involving criminal career trajectory analysis, but there is still a need for a method to discover interesting relationships between the many different types of events in a criminal history. Existing approaches to

quantitative criminal career analysis focus primarily on offense rate predictions.

Due to the limitations outlined above, measures adapted from association rule mining, such as support and confidence may not adequately capture predictive relationships between time-ordered events. Since event hazard patterns do not have these limitations, we expect them to be well suited for integration with time-based models such as RR and AFT.

III. OBJECTIVES OF THE PROPOSED PATTERN DISCOVERY SYSTEM

Existing quantitative approaches to criminal career analysis are not well suited to the ad hoc discovery of relationships between different event types. Event sequence mining is a potential method for the discovery of such relationships.

Event hazard patterns have not yet been used with time based measures of interest. In this work we empirically evaluate the predictive ability of event hazard patterns selected using time based models.

We propose a software instantiation to address the challenge of ad-hoc discovery of predictive event sequences in criminal careers. The main objectives of the proposed pattern discovery system are:

1) Discovery of frequent event sequences in a database of criminal career events
2) Selection of accurate predictive patterns

In this work, we make two key contributions:

3) A domain contribution: facilitating the ad-hoc discovery of relationships between various different event types in a criminal history
4) A methodology contribution: introduce the use of time-based models with event hazard patterns

IV. DESIGN AND DEVELOPMENT

The proposed pattern discovery system builds on an event hazard pattern discovery algorithm. A crime analytics system that will utilize this pattern discovery system is currently under development. In this work we adapt the pattern discovery system for use with time-based measures of interest.

The pattern discovery algorithm is designed to facilitate discovery of event patterns in an event history database expected to contain frequent event sequence patterns separated by both short and long time intervals during which each subsequent event does not yet occur. Such patterns are a specialization of event sequence patterns and are referred to here as event hazard patterns. Given that we expect time based event occurrences that are independent from each other to occur at intervals that follow an exponential distribution, we apply hazard constraints to approximate intervals of exponentially increasing size. This strategy is described as heterogeneous constraints in [11].

A. Definitions

Except where specifically noted, the following are definitions of terms commonly used in event sequence mining. For further details on these terms see [9] and [10].

TABLE II
EVENT HAZARD PATTERNS

Pattern	Opportunities	Support
b	-	-
$b \rightarrow b$	3	2
$p \rightarrow b \rightarrow b$	1	1

Event Type: An event-type refers to a class of discretely identifiable events with common characteristics.

For example, when an individual is arrested charged with a burglary offense, an event type of burglary arrest charge occurs. Additionally, it can be said that a more general event type of arrest charge, or property crime related arrest charge occurs at the same time. An event type is alternatively referred to as an event.

Event Occurrence: The occurrence of an event is denoted (e, t), where e represents the event type and t represents the time of the event occurrence. The unit of discretization for t, such as second, minute, hour, day, etc. is an important consideration when selecting constraints that must be satisfied by t. For example, $(c, 5)$ is the occurrence of event (or event type) c at time 5.

Event Sequence: An event sequence of length n is denoted $<(e_1, t_1),(e_2, t_2),...(e_n, t_n)>$ where e_i represents the type of the ith event, t_i represents the time of the ith event, and $t_{i-1} < t_i$. An event sequence is a time oriented arrangement of event occurrences. For example, $<(b,4),(c,5)>$ is an event sequence. In this work we address only serial event sequences.

Event Sequence Pattern: A frequently occurring event sequence, as defined by a minimum support threshold. An event sequence pattern can be denoted as $<(e_1, T_1), (e_2, T_2),... (e_n, T_n)>$, where e_i represents the type of the ith event occurrence, and T_i represents the collection of all occurrences of the ith event type. Each occurrence represented in T_i with $i>1$ corresponds to an antecedent occurrence represented in T_{i-1}. Alternatively, an event sequence pattern can be summarized in a more compact and intuitive form, as a sequence of events: $b \rightarrow c \rightarrow p$.

In the context of sequential pattern or sequence mining, an event sequence is frequent when it occurs in many input sequences. In the context of frequent episode mining, an event sequence pattern is frequent when there are many occurrences of the pattern. In this work, we consider event sequence mining in the context of frequent episodes.

Gap Constraint: The requirement that except for the initial event occurrence, for any event occurrence (e_i, t_i) in an event sequence, there exists at least one event occurrence (e_{i-1},t_{i-1}) where $mingap \leq (t_i-t_{i-1}) \leq maxgap$. For example, two events in an event sequence satisfy a minimum gap constraint if they are separated by at least $mingap$ and they satisfy a maximum gap constraint if they are separated by at most $maxgap$.

The selection of appropriate $mingap$ and $maxgap$ are domain specific. Gaps are chosen by a human operator to reduce the number of irrelevant patterns that are discovered.

Hazard Constraint: The requirement that except for the initial event occurrence, for any event occurrence (e_i, t_i) in an
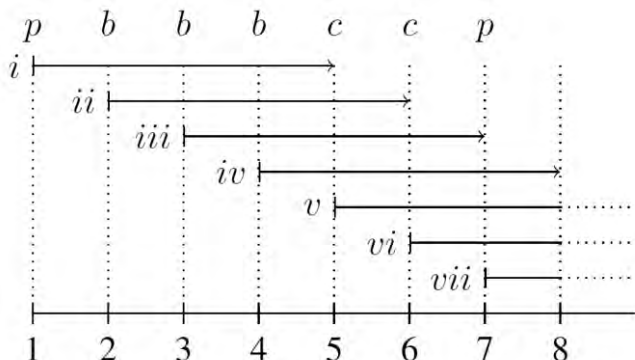
Fig. 2. Windowing and unrelated occurrences.

event sequence there exists no event occurrence (e_r, t_r) where $e_r = e_i$ and $t_{i-1} < t_r < t_i$. Furthermore, each antecedent occurrence (e_{i-1}, t_{i-1}) is a unique occurrence. In other words, each successive event occurrence in an event sequence is the first successive occurrence of the specified type, relative to occurrences of the specified antecedent event type in the same event sequence. Hazard constraint is a new term first introduced in [11].

As with *mingap* and *maxgap*, a hazard constraint can be expressed as a *minhaz* and *maxhaz* for similar effect. A hazard constraint $h((min,max])$ specifies an interval during which the subsequent event may first occur, such that for all event occurrences of a given event hazard pattern, $minhaz < (t_i - t_{i-1}) \leq maxhaz$. In other words, the first occurrence of each subsequent event type, relative to the antecedent occurrence, takes place after *minhaz* and may occur as late as *maxhaz*.

Hazard Pattern / Event Hazard Pattern: An event sequence pattern wherein each subsequent event occurrences the first subsequent occurrence of that particular event type. A specific hazard constraint of *minhaz* and/or *maxhaz* may also be specified, to select only those cases where the subsequent event first occurs after *minhaz* but no later than *maxhaz*. A given event hazard pattern $a \rightarrow b$ can be expressed with a hazard constraint as *a (minhaz, maxhaz] b*. Hazard Patterns were recently introduced in [11]

Relative Support: The number of unique antecedent event occurrences that are followed by a subsequent event type in an event hazard pattern. For example, in Fig. 2, $c \rightarrow p$ occurs twice, but $c \rightarrow c$ occurs only once. Relative support was proposed for event hazard patterns in [11].

Relative Risk: The ratio of the risk within a treatment group over the risk of the control group. It is used to measure the cumulative treatment effect at the end of a period of time. For a discussion of practical application of relative risk ratios, see [13].

B. Justification for Relative Support

There are ten methods of counting support described in [10]. However when using these support counts to describe sequential relationships, a number of challenges arise. Relationships between events in an event sequence are not representative when support count is affected by (a) length of patterns, (b) non-independence between pattern occurrences,

and (c) side effects of unrelated pattern occurrences on support count. Further, in the case of incomplete or censored observations, we need to draw on time based analysis methods. We discuss each of these challenges in detail below.

Length of patterns: When the number of pattern occurrences is directly dependent on the length of the pattern, it is difficult to use differences in support count to construct sequential relationships between shorter and longer extensions of those patterns. We generally expect longer patterns to occur less frequently, but window-based counting methods further penalize the support count of longer sequences. One example of this phenomenon is with event sequences that are constrained by a window of opportunity or expiry time. For instance, the event sequence $<(b,3),(c,4)>$ in Fig. 2 appears in four windows of opportunity of size five whereas the event sequence $<(b,3),(c,4),(c,5)>$ occurs in only three windows of opportunity.

Independence: An alternative to window based counting is occurrence-based counting. Examples of these are minimal occurrence based, non-interleaved patterns, non-overlapping patterns, and distinct occurrence based counting. Each of these suffers from a lack of independence between pattern occurrences. This is because in each of these cases, some pattern occurrences are not counted based on the position and ordering of events in other occurrences of the same pattern. Examples of such missed counts are detailed in [15]. Violation of the independence assumption makes it more difficult to describe relationships between patterns using statistical methods.

Unrelated Occurrences: A solution to the problem of non-independent occurrences is to use head or total frequency. With head frequency, the number of pattern occurrences is based on the number of windows of a specified size that start with the head (first event) of the pattern [16]. However, the same challenges described for other window-based counting methods still apply to head frequency. In addition, head frequency has the undesirable side effect of over-representing the number of occurrences of patterns with a frequent head.

For example, in Fig. 2 the pattern $b \rightarrow c \rightarrow c$ has a head support of three (windows *ii, iii*, and *iv*). However, the support count is unduly inflated by the relationship represented in the initial $b \rightarrow c$ sub-pattern. Total frequency is a partial remedy to this problem whereby the support is equal to the lowest head frequency of any sub-pattern [17].

However, this measure can still be unduly inflated by unrelated occurrences of sub-patterns. For instance, in Fig. 2 c and p both have a head frequency of two, so the total support of $c \rightarrow p$ is two even though the support count is affected by an occurrence of p that is unrelated. Thus, we cannot use this measure to describe relationships between antecedent patterns and their subsequent extensions.

Censoring: In Fig. 2 windows *v, vi* and *vii* are all censored. If the patterns are not sufficiently short relative to the event database, this missing data may adversely affect the interpretation of support counts. Missing or incomplete event data, if not accounted for, can be misleading. In the real database of criminal histories used in our analysis, the

Require: $\forall (event, ordinal) \in D : event \in M$
Require: Constraints G
Ensure: $\forall Pat \in Freq : |Pat| \geq sup$
 1: **function** GROW($Pat, Ords$)
 2:    **for all** $event \in M$ **do**
 3:       **for all** $constrt \in G$ **do**
 4:          **if** $|Ords_{constrt.event}| \geq sup$ **then**
 5:             $NewPat \leftarrow (Pat, constrt, event)$
 6:             append $NewPat$ to $Freq$
 7:             $Nextords \leftarrow$ NEXT($Ords_{constrt,event}$)
 8:             GROW($NewPat, Nextords$)
 9:          **end if**
10:       **end for**
11:    **end for**
12: **end function**

Fig. 3. Pattern discovery.

database is considered to be complete, such that events past the end of each criminal history are believed to be negligible. However, the matter of censoring in event sequence patterns is problematic for cases where pattern length is not negligible relative to the length of input sequences. Event hazard patterns do not specifically address event censoring, but these can be addressed with time-based models. For a related discussion of censored events in event sequence mining with variable window sizes, see [18].

Some additional limitations of existing support count methods are also discussed in [15] along with a complex proposed support metric. The proposed support metric relies on non-redundant occurrences (occurrences with no events in common). This non-redundancy requirement introduces dependencies between different occurrences of the same pattern.

With consideration to the limitations outlined above, Hazard patterns are counted by relative support. Given a number of distinct opportunities, support is the number of those distinct opportunities or distinct antecedent events that are followed by the subsequent event. The number of opportunities is less than or equal to the support of the antecedent pattern. Note also that, unlike the event sequence mining approaches described above, each subsequent event in an event hazard pattern is the first such subsequent event, and that it is not necessarily distinct (it may participate in more than one pattern occurrence).

For instance, in Fig. 2 there are three occurrences of b. All three of them are followed by c, so support for $b \rightarrow c$ is three. However, since these three antecedents all converge on the same c occurrence at position four, there is only a single distinct opportunity to extend $b \rightarrow c$ to the subsequent c or p events. The support of $b \rightarrow c$ is three (out of three distinct opportunities), and the support of $b \rightarrow c \rightarrow p$ is one (out of one distinct opportunity). Using this approach, confidence is the proportion of successes given a number of distinct opportunities, meaning the confidence of $b \rightarrow c$ is 1.0 and the confidence of $b \rightarrow c \rightarrow p$ is also 1.0.

In this way, hazard patterns address all of the challenges discussed above except censoring. Length of patterns does not unduly affect support count, pattern occurrence counting treats

Require: $[R_{event,ordinal}]_{(m \times n)}; [I_{event,ordinal}]_{(m \times n)}$
Require: alphabet of all event types M
 1: **function** NEXT($Antecedents$)
 2:    $Antecedents \leftarrow$ UNIQUE($Antecedents$)
 3:    **for all** $event \in M$ **do**
 4:       **for all** $ord \in Antecedents$ **do**
 5:          $c \leftarrow R_{event,ord}$ ▷ get constraint
 6:          $nextOrd \leftarrow I_{event,ord}$
 7:          append $nextOrd$ to $Subsequents_{event,c}$
 8:       **end for**
 9:    **end for**
10:    return $Subsequents$
11: **end function**

Fig. 4. Get next ordinals by event, constraint.

each pattern occurrence independently, and events that are unrelated to the relationship do not affect relative support counts. However, although the confidence measure provides an indicator of the proportion of success, it does not provide information about the generalizability of the pattern. To this end, and to account for censored observations, we draw on time-based models in our analysis.

C. Algorithm Design

It is expected that some frequent event sequences will include events that occur close together and others that occur far apart. One way to capture such patterns is to use a windowing strategy, first described in [19], to create item sets, alternatively presented as a partial order or as parallel episodes [20]. However, this approach may discard potentially valuable information, and relies on the analyst to specify optimal windowing and gap constraints.

Windowing and gap constraints capture all events that fall within the constraint boundaries, but do not differentiate between them and do not capture the non-occurrence of an event. Instead, hazard constraints specify a period during which an event does *not* occur, followed by a period during which the first occurrence an event *does* take place.

The proposed algorithm iteratively applies hazard constraints of exponentially increasing sizes, similar to the use of multiple periodic constraints in [21]. The proposed implementation uses progressively larger intervals to represent the number of months before the first occurrence of the subsequent event, such as re-arrest following discharge or parole release.

The GROW function shown in Fig. 3 uses pairs of event and ordinal values to represent a database of known event occurrences. Ordinals are translated to offsets at $O(1)$ cost as needed for constraint calculations. Input ordinals are supplied in a matrix indexed by *event,constraint* where each $M_{event,constraint}$ represents the antecedent ordinals for the current pattern growth step. In Line 4, those antecedents with cardinality that is high enough to meet a specified support threshold are added to the frequent pattern database in line 4, and are passed to the NEXT function, where a new matrix of candidate event occurrences is created, and passed to the subsequent recursive GROW attempt on line 8.

Fig. 4 contains the NEXT function, which takes as input a

collection of antecedent ordinals, grouped by event, and produces the Ordinal matrix *NextOrds* needed in line 7 of Fig. 3. This function uses two indexes: $R_{event,ordinal}$ and $I_{event,ordinal}$. See Fig. 5(c) and 5(d) for the R and I indexes corresponding to event sequence τ shown in Fig. 5(a). Ordinals in I have corresponding offsets in Fig. 5b and constraint identifiers in R have corresponding constraint intervals in Fig. 5(e). R and I are matrices of dimension $(m \times n)$ where m is the alphabet of all possible events, and n is the number of distinct offsets. Multiple events may occur at the same offset. I contains the ordinal of the subsequent occurrence of a given event type. The value stored at the intersection specified by an ordinal and an event type corresponds to the ordinal of the first subsequent occurrence of that event type. R contains the constraint that is satisfied at a given event offset (represented as an ordinal), relative to its immediate antecedent event.

On line 5 of the NEXT function pseudo-code in Fig. 4, for each antecedent event occurrence, the constraint $R_{event,ordinal}$ that is satisfied for each potential subsequent event is retrieved. Given the half-open interval topology used to describe the different constraints, each subsequent event can satisfy one constraint. In line 6 the subsequent ordinals are retrieved from I and then grouped according to their matching constraints in line 7. The creation of R and I are not described here, but are straightforward. Their purpose is to pre-compute comparisons and look-ups that are frequently repeated during candidate generation. Simply put, the index serves to reduce the number of calculations required during candidate generation at the cost of increasing memory usage up front. Optimization strategies to take advantage of redundancies in R and I are currently being evaluated.

V. EVALUATION

The goals in this undertaking involve both a domain and a methodology contribution. For the problem domain, this work provides a method for discovering predictive sequences of events. The methodology contribution is the use of a time-based measure of interest to demonstrate the generalizability of discovered relationships.

A. Data Preparation

The pattern discovery system was used to discover patterns in a data set of complete criminal histories. The histories were collected from part of a non-random sample of offenders who entered the California Youth Authority's Deuel Vocational Institute in 1964 and 1965. The event database contains of 54,175 arrest records and associated dispositions, parole, and discharge events for 3,652 individuals from the time of first arrest through 1983. Dates were discretized to the nearest 15th day of the month [22].

For this analysis, the individual histories in the dataset were randomly assigned to either the training set or the testing set. Each arrest event was associated with up to five arrest charges. Additionally, the nature of the disposition and judgment date was also recorded for each arrest event, as were parole and discharge events. Arrest charges were encoded both as the

$$\langle (A,2), (A,3), (B,5), (B,6), (C,10), (B,11) \rangle$$
(a) event sequence τ

ord offset		0 1 2 3 4 5	0 1 2 3 4 5	h constrt
0 2				0 0
1 3		A 1 0 0 0 0 0	A 1 0 0 0 0 0	1 (0, 3]
2 5		B 2 2 3 5 5 0	B 1 1 1 2 1 0	2 (3, 6]
3 6		C 4 4 4 4 0 0	C 3 3 2 2 0 0	3 (6,12]
4 10		(c) I	(d) R	4 (12,24]
5 11				(e) constrt
(b) offsets				

Fig. 5. Ordinal, constraint, and offset indexes

TABLE III
OCCURRENCE FREQUENCIES
FOR PATTERN P AND BASELINE B

	Occurrences	Non-occurrences	Total
P	(a) 775	(b) 914	1689
B	(c) 1146	(d) 2648	3794
Total	1921	3562	(n) 5483

specific arrest charge as well as a general arrest event. Disposition events were similarly encoded, with the additional adaptation that arrest dates were used for disposition events. Note that due to the discretization of the data, the relationship between an arrest and a conviction for that same arrest is not represented. All dispositions (including convictions) were recoded to the arrest charge date. Any patterns showing both arrests and convictions have nothing to do with conviction rates.

Hazard pattern mining was performed on the training data with multiple different hazard constraints per pattern at increments of 0, 3, 6, 9, 12, 24, 48, 96, 192, and 384 months to generate hazard constraints of (0,3], (3,6], (6,12], and so forth (see heterogeneous constraints described in [11]). Only patterns with a support count of at least 500 were mined. This process yielded 44 frequent events and 1085 hazard patterns (single events are not considered hazard patterns). For each hazard pattern the number of opportunities, the support count (number of opportunities that were successful) and the number of unique subsequent events were recorded. The same patterns were also mined from the test set.

Of the 1085 hazard patterns, 305 patterns involved an event sequence of three or more events. Each of these was compared against the equivalent patterns from the test set and against the baseline or control pattern with the first antecedent and constraint removed. For example: $a \rightarrow p \rightarrow a$ from the training set would be compared with the same pattern in the test set, as well as against a minimally differentiated baseline of $p \rightarrow a$ from the training set. There should be no statistically significant difference between the training set and the test set. Further, there should be agreement between the test set and the training set about the expectations implied by the discovered patterns.

B. Relative Risk

One measure considered for this evaluation was Relative Risk Ratio (RR) [12], [13]. Since the RR measure is applied only at the end of a follow-up period, the measure is inherently sensitive to the choice of follow-up. Further, since relative risk does not account for left truncated data, we use this measure when there is no minimum hazard constraint. For this reason we evaluate only those patterns that end with a hazard constraint of greater than zero and up to three months, expressed as (0,3].

Relative Risk (RR) is defined as follows:

$$RR = \frac{\text{Estimated risk in the exposed group}}{\text{Estimated risk in the unexposed group}}$$

For our analysis, we consider the set difference between the baseline pattern and the pattern of interest to be the unexposed group. Consider the following two patterns:

P: "Arrest" (0,3] "Continue probation"(0,3] "Arrest"
S: "Continue probation" (0,3] "Arrest"

To construct a baseline, we calculate the measures for *B = (S-P)*. Since *S* includes all pattern occurrences that participate in *P*, we subtract *P* from *S* to create a baseline *B* to compare against. In other words, *B* contains all occurrences in *S* that do not also occur in *P* (see Table III). Thus, we can measure the relative risk of "Arrest" within three months associated with a disposition of "Continue probation" occurring within three months of an arrest compared to those cases where it did not occur within three months of a preceding arrest.

Referring to Table III, we can calculate

$$RR = \frac{(a/(a+b))}{(c/(c+d))} = \frac{775/1689}{1146/3794} \approx 1.52$$

We see that the risk of re-arrest within three months in the final stage of pattern *P* (0.46) is 1.52 times the risk of re-arrest within three months in pattern *B* (0.30) for an absolute risk difference of 0.16%.

It is important to note at this stage that the pattern does not provide enough information to state that the initial "Arrest" and delay before the "Continue probation" disposition event were the key predictors. Although this may seem to be a perfectly intuitive conclusion, it would overlook the unrelated occurrence problem described in Section IV-A. To put it another way, even though all occurrences of *P* contain an occurrence of *B*, some occurrences of *B* may have occurred without a coinciding occurrence of *P*. There may be such a relationship, but it is not represented by these counts.

Instead, what is represented is the relationship between the entire antecedent pattern and the last constraint plus event combination. We can compare the effect represented by *P* to the effect in the shorter pattern *B* to determine whether the additional information provided by the longer pattern may be of use. To this end, we estimate the confidence interval for the RR calculated above. The standard error is symmetrical about the logarithm of RR as follows:

$$SE(\ln RR) = \sqrt{\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} + \frac{1}{c+d}}$$

A 95% confidence interval is then estimated by taking the antilog:

$$e^{\ln(RR) \pm SE(\ln RR) * 1.96}$$

Similarly, a Z score can be estimated as follows:

$$Z = \frac{\ln RR}{SE(\ln RR)}$$

This makes it possible to perform a test to see if the difference in relative risk ratios is due to chance. See [13] for an example. For patterns *P* and *B*, the resulting z-score is 11.56, supporting a rejection of the null hypothesis that the difference between *P* and *B* is due to chance.

Further, we can show the robustness of the pattern by comparing the risk ratio of *P* in the training set with the risk ratio of *P* in the test set. The risk ratio for *P* in the test set was 748 / 1693. Using the same process outlined above, we estimate a relative risk ratio of 1.02, showing that the two risk ratios are almost the same (RR=1.00 would indicate no difference at all). We then calculate a z-score of 0.48 showing that we are unable to reject the null hypothesis that the differences between *P* from the training set and *P* from the test set are due to chance.

For this portion of the analysis, we concern ourselves only with the patterns ending with a constraint of (0,3] due to the limitations of the relative risk measure discussed above. Of the 305 patterns that could be paired with a baseline, 66 patterns have an ending constraint of (0,3]. In other words, for this evaluation, we consider only patterns with at least three events, and with the constraint between the last two events being greater than zero and up to three months.

Based on a 95% confidence interval, 47 of the 66 selected patterns were shown to represent a statistically significant difference in risk between discovered patterns and their corresponding baselines. Further, 2 of the 66 selected patterns were found to have statistically significant differences between the risk ratios discovered in the training set when compared to the risk ratios discovered in the test set and there were no instances where patterns in the training set indicated a significant increase while the corresponding patterns in the training set indicated a significant decrease and vice versa. In other words, the training and test set did not contradict each other.

The risk difference between the selected 66 patterns and their corresponding baselines ranged from -0.07 to 0.16. The range of values for risk (support/opportunities) for the discovered patterns was 0.08 to 0.50.

C. Accelerated Failure Time

Whereas RR is a measure indicating the difference in the number of people affected by treatment when compared to a control group, accelerated failure time (AFT) models show the difference in expected time to event for the two groups.

AFT models are described in detail in [14]. For this analysis we used the Survival package for the R statistical software platform [23]. However, since the data being analyzed is discretized to the nearest month, and AFT expects continuous values, the events were analyzed as interval censored. In other words, the event occurred within a one month interval, but the exact time is unknown. It may be fair to state that all measurements are discretized to some degree, but in this case, a conservative approach was taken. An AFT model was constructed without regard to age, and the results were evaluated in the same manner as with RR above. Although we can expect improved results by taking into account time-varying coefficients such as age, and non-varying predictors such as demographics, our focus for this work is on the efficacy of the discovered patterns themselves.

The same 66 patterns described above were each individually used to fit an AFT model, using a logistic distribution. Other distributions that were tested were Weibull, lognormal, exponential, and Gaussian. In all cases, patterns in the training set were not contradicted by patterns in the test set. The logistic distribution was selected because it produced the most consistent results across training and test sets. Of the 66 patterns, based on a 95% confidence interval, 39 patterns showed a significant decrease in time to next event, and three patterns showed a significant increase. In 24 cases, there was no significant difference between pattern and baseline. AFT models were also fitted to compare the patterns found in the training data with the patterns found in the test data. Only one of the 66 patterns was found to have differences that cannot be attributed to chance. Furthermore, as we found with RR, there were no cases where the models fitted to the training and the testing data presented directly contradictory results.

VI. CONCLUSION

In this work we explored the use of time based measures for rule selection to address the problem of predicting criminal behavior. Although there has been some limited use of time based measures in event sequence mining, some characteristics of existing methods of event sequence counting make it difficult to accurately discover predictive relationships. These characteristics are (A) support count methods that unduly penalize longer patterns, (B) support count methods that involve dependencies between occurrences of the same pattern (an occurrence may or may not be counted depending on characteristics of other occurrences of the same pattern), and (C) support count methods that include unrelated event occurrences (sub-pattern or event occurrences that do not participate in a relationship with the super-pattern). Hazard patterns do not suffer from these limitations. We demonstrate the utility of hazard patterns for discovering sequential relationships between diverse event types. Patterns were

selected and evaluated using two time-based methods: relative risk ratio, and accelerated failure time models.

We note a number of important limitations. First, the relative risk ratio is not suitable for multiple follow-up periods. Patterns with follow-up periods other than (0,3] were excluded from the relative risk analysis. Further, since the event data is discretized to the nearest 15th day of the month, some short term patterns, such as an arrest event leading to a disposition event in less than one month were excluded during data preparation. Additionally, relative risk is sensitive to choice of follow-up period. For instance, the outcomes may have been different if we had selected (0,6] or (0,12] as the follow-up period.

Further, although care was taken to ensure that opportunity and pattern occurrence counts did not violate the independence assumption, some questions affecting generalizability remain. For instance, since the individuals were not randomly assigned to patterns and their respective baselines, a concern over selection bias is justified. The pattern occurrences were not matched or balanced to correct for selection bias. However, given the stability of the patterns between the training and the test set, we did not find evidence of a significant selection bias effect. However, in this work we evaluate only the directionality of effect. We may encounter evidence of bias upon examination of predicted effect size.

A number of directions for further work have been noted. We plan to explore the use of hazard ratio and survival curves to describe the effect over time that is represented by a particular pattern. Other available covariates, particularly age, may improve the accuracy of event hazard patterns. Further, the existing pattern database was mined at a relatively high support threshold. It remains to be seen how robust these patterns are, and particularly how useful time based measures of interest will be when the minimum support threshold is lowered. Another natural extension of this work is the use of sensitivity analysis to address concerns of selection bias. Finally, the use of the techniques described in this work can reasonably be extended to other domains where there are many different types of antecedent events, and where time before the first subsequent event is important.

REFERENCES

[1] D. S. Nagin and K. C. Land, "Age, criminal careers, and population heterogeneity: specification and estimation of a nonparametric, mixed Poisson model," *Criminology,* vol. 31, pp. 327-362, 1993.

[2] A. S. Bhati and A. R. Piquero, "Estimating the Impact of Incarceration on Subsequent Offending Trajectories: Deterrent, Criminogenic, or Null Effect?," *The Journal of Criminal Law and Criminology,* vol. 98, pp. 207-253, 2007.

[3] A. Haviland, D. S. Nagin, and P. R. Rosenbaum, "Combining propensity score matching and group-based trajectory analysis in an observational study.," *Psychological methods,* vol. 12, pp. 247-67, September 2007.

[4] B. E. Bersani, P. Nieuwbeerta, and J. H. Laub, "Predicting Trajectories of Offending over the Life Course: Findings from a Dutch Conviction Cohort," *Journal of Research in Crime and Delinquency,* vol. 46, pp. 468-494, September 2009.

[5] J. Vlahos, "The Department of Pre-Crime," *Scientific American,* vol. 306, pp. 62-67, December 2011.

[6] C.-H. Yu, M. W. Ward, M. Morabito, and W. Ding, "Crime Forecasting Using Data Mining Techniques," *Proceedings of the 11th International Conference on Data Mining Workshops,* pp. 779-786, December 2011.

[7] T. Abraham, "Event sequence mining to develop profiles for computer forensic investigation purposes," in *Proceedings of the 2006 Australasian workshops on Grid computing and e-research-Volume 54*, ed: Australian Computer Society, Inc., 2006, pp. 145-153.

[8] M. Spiliopoulou, "Managing interesting rules in sequence mining," *Principles of Data Mining and Knowledge Discovery,* pp. 554-560, 1999.

[9] H. Mannila and H. Toivonen, "Discovering frequent episodes in sequences extended abstract," in *1st Conference on Knowledge*, ed, 1995.

[10] A. Achar, S. Laxman, and P. S. Sastry, "A unified view of the apriori-based algorithms for frequent episode discovery," *Knowledge and Information Systems,* May 2011.

[11] C. A. Janzen, A. V. Deokar, and O. F. El-Gayar, "Non-parametric discovery of event sequence patterns in criminal behavior," in *Proceedings of the 46th Annual Hawaii International Conference on Systems Science (HICSS-46 '13) Symposium on Credibility Assessment and Information Quality in Government and Business*, ed. Maui, HI: IEEE Computer Society, 2013.

[12] L.-A. McNutt, "Estimating the Relative Risk in Cohort Studies and Clinical Trials of Common Outcomes," *American Journal of Epidemiology,* vol. 157, pp. 940-943, 2003.

[13] V. Bewick, L. Cheek, and J. Ball, "Statistics review 11: assessing risk.," *Critical care (London, England),* vol. 8, pp. 287-91, 2004.

[14] R. Cook and J. Lawless, "The statistical analysis of recurrent events," 2007.

[15] M. Gan and H. Dai, "Reliable Knowledge Discovery," 2012.

[16] M. Gan and H. Dai, "Fast mining of non-derivable episode rules in complex sequences," *Modeling Decision for Artificial Intelligence,* vol. 1, pp. 67-78, 2011.

[17] K. lwanuma and H. Nabeshima, "On anti-monotone frequency measures for extracting sequential patterns from a single very-long data sequence," in *IEEE Conference on Cybernetics and Intelligent Systems, 2004.* vol. 1, ed: IEEE, 2004, pp. 213-217.

[18] N. Müller, M. Studer, G. Ritschard, and A. Gabadinho, "Extraction de règles d'association séquentielle à l'aide de modèles semi-paramétriques à risques proportionnels," *mephisto.unige.ch,* 2008.

[19] R. Agrawal and R. Srikant, "Mining sequential patterns," in *Proceedings of the Eleventh International Conference on Data Engineering*, ed: IEEE Computer Society, 1995, pp. 3-14.

[20] H. Mannila, H. Toivonen, and A. I. Verkamo, "Discovery of Frequent Episodes in Event Sequences," *Data Mining and Knowledge Discovery,* vol. 289, pp. 259-289, 1997.

[21] J. Han and G. Dong, "Efficient mining of partial periodic patterns in time series database," *Data Engineering, 1999. Proceedings.,* pp. 106-115, 1999.

[22] E. Wenk. (2006). *Criminal Careers, Criminal Violence, and Substance Abuse in California, 1963-1983*. Available: 10.3886/ICPSR09964.v1

[23] T. M. Therneau, "A Package for Survival Analysis in S," ed, 2012.