

Dakota State University Beadle Scholar

Masters Theses & Doctoral Dissertations

Fall 10-1-2013

Crime Analytics: Mining Event Sequences in Criminal Careers

Carl A. Janzen
Dakota State University

Follow this and additional works at: <https://scholar.dsu.edu/theses>

Recommended Citation

Janzen, Carl A., "Crime Analytics: Mining Event Sequences in Criminal Careers" (2013). *Masters Theses & Doctoral Dissertations*. 284.
<https://scholar.dsu.edu/theses/284>

This Dissertation is brought to you for free and open access by Beadle Scholar. It has been accepted for inclusion in Masters Theses & Doctoral Dissertations by an authorized administrator of Beadle Scholar. For more information, please contact repository@dsu.edu.

CRIME ANALYTICS: MINING EVENT SEQUENCES IN CRIMINAL CAREERS

A dissertation submitted to Dakota State University in partial fulfillment of the
requirements for the degree of

Doctor of Science

in

Information Systems

October 2013

By

Carl A. Janzen

Dissertation Committee:

Amit V. Deokar, Ph.D. (Chair)

Omar F. El-Gayar, Ph.D.

Viki Johnson, Ph.D.

Darryl Plecas, Ed.D.



DISSERTATION APPROVAL FORM

This dissertation is approved as a credible and independent investigation by a candidate for the Doctor of Science in Information Systems degree and is acceptable for meeting the dissertation requirements for this degree. Acceptance of this dissertation does not imply that the conclusions reached by the candidate are necessarily the conclusions of the major department or university.

Student Name: Carl Andrew Janzen

Dissertation Title: Crime Analytics: Mining Event Sequences in Criminal Careers

Dissertation Chair:  Dr. Amit Deokar

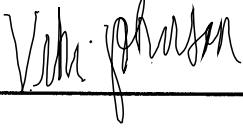
Date: Nov 29, 2013

Committee member: 

Date: Nov 29/13

Committee member: 

Date: 12/2/13

Committee member: 

Date: Nov 29, 2013

ACKNOWLEDGMENTS

I would like to knowledge several people who have contributed to the success of this work, and who have supported and encouraged me throughout my studies. I would like to thank Dr. Amit Deokar, for giving much of his time, for engaging in many thoughtful discussions, and above all for his encouragement and support. I would also like to thank Dr. Omar El-Gayar for his guidance and helpful insights. I would like to thank Dr. Viki Johnson for serving on the committee and for her kind words of encouragement. I would like to also thank Dr. Darryl Plecas for sharing his insights and for providing much encouragement and support.

I would also like to thank my wife, Karen, for her love and for her patience. Her unwavering support has been integral to all of my undertakings, particularly for this work

ABSTRACT

This dissertation addresses the pressing and difficult problem of assessing the risk of re-offending for parolees. The prison system in the state of California has been given a strong mandate to dramatically reduce the prison population. Before final discharge, prisoners often serve a portion of their sentence on parole release, but they are at high risk to re-offend. A number of systems have been developed to aid practitioners in parolee risk assessment, but the recommendations of these systems have not been consistently followed. Field practitioners were skeptical that recommendations adequately accounted for repeat offending histories, and did not believe that the recommendations were logical. We propose a hazard pattern based risk assessment approach to address these concerns. In this work, we demonstrate this approach using real world data, and rigorously evaluate the discovered patterns.

The design science nominal process flow was selected as the methodological framework for this undertaking. The motivating case is a business problem, in context. The search for and development of a solution is documented, including the careful evaluation of existing technologies and development of novel approaches and artifacts where necessary. An IT artifact is developed, demonstrated and evaluated within the context of the motivating case.

The driving question behind this work is this: How can we assess risk of future offending? A substantial body of work has explored this question, reflecting the importance of the question and the difficulty of finding an answer. A number of risk assessment tools have been developed but their accuracy has been moderate and their acceptance by practitioners has been lukewarm. We are thus faced with a need for a way to make accurate risk assessments that can be justified to field practitioners.

As necessary components of a solution, two key contributions are highlighted in this work: a) hazard patterns, which extend existing work in event sequence patterns, and b) a method of selecting and presenting a relatively small number of interesting patterns that codify the rationale underlying the assessment of risk.

The solution was evaluated according to design objectives of parsimony, generalizability across data sets, meaningfulness, and predictiveness over time. We satisfied the objective of parsimony by selecting only those hazard patterns showing statistically significant differences in relative risk. To demonstrate generalizability and guard against over-fitting, ten-fold cross validation testing was performed. The selected patterns were consistent indicators of increase or decrease in arrest risk across folds in cross-validation trials. To test for meaningfulness, pattern discovery and selection was repeated with the underlying data randomly shuffled. The differences in the resulting output empirically demonstrate that the patterns were dependent on the input rather than on the pattern discovery process. Finally, to test for predictiveness over time, hazard patterns discovered in one time frame were compared to arrest outcomes in subsequent time frames. A moderate relationship between antecedent hazard patterns and future outcomes was observed, with lower accuracy near the beginning and end of criminal careers.

DECLARATION

I hereby certify that this project constitutes my own product, that where the language of others is set forth, quotation marks so indicate, and that appropriate credit is given where I have used the language, ideas, expressions or writings of another.

I declare that the project describes original work that has not previously been presented for the award of any other degree of any institution.

Signed,

A handwritten signature in black ink, consisting of a series of fluid, connected strokes that form a stylized representation of the name Carl A. Janzen.

Carl A. Janzen

TABLE OF CONTENTS

ACKNOWLEDGMENTS.....	II
ABSTRACT	III
DECLARATION	V
TABLE OF CONTENTS	VI
LIST OF FIGURES.....	VIII
INTRODUCTION	1
1.1 BACKGROUND AND MOTIVATION	1
1.2 PROBLEM STATEMENT	1
1.3 RESEARCH OBJECTIVES	2
1.4 SCOPE OF THE STUDY.....	2
LITERATURE REVIEW	4
2.1 RISK ASSESSMENT	4
2.2 CRIMINAL CAREER ANALYSIS.....	5
2.3 EVENT SEQUENCE MINING.....	6
2.4 SUMMARY	8
RESEARCH METHODOLOGY	9
3.1 DESIGN SCIENCE RESEARCH METHODOLOGY	9
3.3 OBJECTIVES OF A SOLUTION	11
THEORY AND ARTIFACT DESIGN	15
4.1 HAZARD PATTERNS.....	15
4.2 ALGORITHM DESIGN.....	21
DEMONSTRATION AND EVALUATION.....	25
5.1 DEMONSTRATION: PATTERN DISCOVERY USING REPRESENTATIVE DATA	25
5.2 EVALUATION.....	28
5.2 CONCLUSION.....	34
REFERENCES	36

LIST OF TABLES

Table 1: Research guidelines	9
Table 2: Months until re-arrest (contrived data set)	17
Table 3: Months until paroled (contrived data set)	18
Table 4: Arrests after parole (data set A)	26
Table 5: New offense after parole contact (data set B)	27
Table 6: Cross validation (data set A)	29
Table 7: Cross validation (data set B)	29
Table 8: Arrests after parole (data set A shuffled)	31
Table 9: New offense after parole contact (data set B shuffled)	31
Table 10: Two year arrest, based on releases 2-4 years earlier (data set A)	32

LIST OF FIGURES

Figure 1: Nominal DSRM process model (Peffer et al., 2007)	11
Figure 2: Contrived histories.....	15
Figure 3: Example offset and constraint lookup tables	22
Figure 4: Ordinal and constraint indexes	22
Figure 5: Pattern discovery algorithms	23

CHAPTER 1

INTRODUCTION

1.1 Background and Motivation

The facilities in California's prison system were designed to house approximately 85,000 inmates. These facilities held approximately 156,000 inmates in 2011, when the Supreme court upheld an order that would require the state to decrease the prison population by 46,000 (Newman & Scott). In the face of California's growing prison population, policy makers are under pressure to reduce the number of individuals housed in the prison system, as mandated by the U.S. Supreme court.

One key way to reduce the number of individuals serving their sentence in prison is through parole release. Individuals may serve a portion of their sentence outside of prison, provided they abide by the terms of their release. However, identifying candidates for successful parole release is no easy task when recidivism rate is high and the number of life-long desisters is low. In the context of a criminal career, recidivism is the re-occurrence of an arrest charge or conviction, while desistance is the absence of such a re-occurrence. The rate of recidivism will vary depending on whether the subject of interest is an arrest or a conviction. In California, 84% of individuals released from prison during the fiscal year 2007-2008 were re-arrested within three years of release, and 60% were convicted (Cate et al.).

1.2 Problem Statement

The California Parole Violation Decision-Making Instrument (PVDMI) is a risk assessment tool that utilizes risk and severity scores from the California Static Risk Assessment Instrument (CSRA). This tool was recently deployed to select locations in a pilot study. A process evaluation showed that the tool as used by the pilot sites did not result in consistency in parole release decisions and did not lead to a reduction in recidivism. This may be due to the deviation of practitioners from the recommendations of this tool. Two key concerns were that practitioners did not see the recommendations as logical, and that criminal

histories with repeat offending did not appear to be adequately accounted for (Turner, Braithwaite, Kearney, Murphy, & Haerle, 2012). Thus, in addition to the difficult task of assessing risk of recidivism, we are presented with the challenge of justifying the risk assessment to the decision maker.

1.3 Research Objectives

We propose hazard pattern analysis both for the discovery of patterns leading to recidivism, as well as for communicating the risk of recidivism to a decision maker. Hazard patterns can capture commonalities in the order as well as the time between many different types of events in criminal histories.

There are two primary goals for this work. The first goal is to find how we can assess risk of recidivism based on past offending behavior. The second goal is to codify the risk, as well as the basis for the risk in simple terms.

In this work, we make two key contributions. First, we demonstrate that hazard pattern mining can be used to discover patterns that can reliably predict differences in risk of re-arrest following parole release. Second, we propose and demonstrate a test of meaningfulness for hazard patterns. Without such a test, it can be difficult to differentiate between patterns that occur by chance and genuine meaningful patterns.

1.4 Scope of the Study

This study examines the relationship between prior events and parole violations, and provides a way to summarize and codify that relationship in a concise manner. Two data sets were analyzed for this purpose.

The first data set (data set A) consists of arrest charge, disposition type, parole, and discharge data for a group of young male offenders in 1964 and 1965 entering the California Youth Authority (CYA). For 3,652 of the original 4,165 of these individuals, criminal histories were collected for both their juvenile record and the subsequent 20 years. Dates were discretized to the nearest 15th day of the month. In total, 54,175 arrest records were evaluated. This data set is available through the Inter-University Consortium for Political and Social

Research at the University of Michigan (ICPSR). This data was originally collected for the study of recidivism rates. For further details, see (Wenk, 2006).

The second data set (data set B) consists of intensive parole supervision records for 146 parolees in Milwaukee, released in 1980-1981. This data set consists of parole officer contacts as well as violation data, including the method of contact, over a two year period. Given the short time span, only a descriptive evaluation was performed for this data set.

For both data sets, pattern mining yielded a large collection of hazard patterns. Of these patterns, a subset was highlighted as indicating a statistically significant increase in relative risk of parole violation. The data mining process was repeated with the same data, but with the ordering of the events shuffled. The discovered patterns were evaluated for robustness using ten-cross validation, and for significant differences between ordered and shuffled inputs.

Additionally, data set A was evaluated to determine whether patterns discovered in one period also corresponded to patterns in a subsequent time period. We noted a moderately strong relationship across different time periods, suggesting that other time related covariates also play an important role in determining risk of parole violations.

Further, in both data sets, a test for over-fitting was performed by ten-cross validation. We show that the discovered hazard patterns were consistent between validation folds, supporting the conclusion that such patterns may be generalizable to other similar data.

Finally, a test for meaningless output was performed. The design of this test follows from (Keogh & Lin, 2004) where a clustering technique used in a large number of publications was shown to produce output that was independent of the input. We demonstrate support for the conclusion that hazard patterns are meaningful based on dramatic differences in quantity and content of patterns discovered in ordered versus shuffled data.

CHAPTER 2

LITERATURE REVIEW

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 recidivism prediction in parolees.

2.1 Risk Assessment

Risk assessment instruments may draw on static predictors of recidivism, dynamic factors, and theory. Static predictors are those characteristics that cannot be changed, such as prior offenses. Dynamic factors are those which can be changed, such as the attitude of an individual. In the most advanced systems, these are supplemented by theory and the integration of needs assessments.

Due to the many differences in the populations and correctional systems of different regions, it is not surprising to see development of state-specific risk assessment tools. Examples of state specific systems are Ohio's progressive sanction grid (Martin & Dine, 2008), the Minnesota Screening Tool Assessing Recidivism Risk (MnSTARR) (Duwe, 2013), and the California Parole Violation Decision Making Instrument (PVDMI) (Turner et al., 2012). Ohio's progressive sanction grid and the Minnesota Screening tool both make use of static and dynamic predictors, but the PVDMI relies only on static predictors as rated by the California Static Risk Assessment Instrument (CSRA).

A commonality of these systems is the incorporation of multiple predictive factors and the use of statistical methods to produce an objective decision. Further, both the Ohio screening tool and the PVDMI encountered considerable resistance from practitioners in the field. Parole officers using the Ohio screening tool questioned whether the decisions of the tool were logical, and parole officers in California consistently escalated the recommended sanction for parolees with significant prior criminal behavior. Turner et al. (2012) suggested

that parole officers may not have been confident that criminal histories were properly taken into account by the system. It remains to be seen whether the deployment of MnSTARR will fare better.

Resistance to actuarial tools is not altogether surprising, since predicting recidivism is very difficult. The accuracy of available tools has been moderate. A meta-analysis of risk assessment instruments found these produced area under curve scores ranging from 0.65 to 0.71 (Min, Wong, & Coid, 2010).

2.2 Criminal career analysis

Research in the area of quantitative criminal career analysis often makes use of group trajectory modeling. First introduced by (Nagin & Land, 1993) and since used in many other studies, this technique can be used to make predictions of criminal behavior over the life course of the individual. Criminal career analysis using group trajectory modeling involves clustering offenders by their offense rate over the span of their criminal careers. Comparisons can then be made across cluster groups. For instance, chronic offenders might be compared with early desisters to identify demographic or early offending patterns that predict which group a young offender will eventually belong to.

A Canadian study by (Haviland, Nagin, & Rosenbaum, 2007) examined the relationship between adolescent gang joining and future offending. For this comparison, the treatment group consisted of gang joiners and the control group consisted of non-joiners. Since assignment to either group was non-random, propensity score matching (PSM) was used to adjust for selection bias. PSM is a technique where comparisons are made between matched pairs of individuals across groups, where each individual has the same propensity to belong to the treatment group. PSM has been shown to be useful, but is not applicable when the significant predictors of group membership are unknown.

In (Bhati & Piquero, 2007), 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. In order to more effectively characterize important predictive characteristics of offender histories, the analysis included variables to represent age at first arrest, number of previous arrests, whether the previous arrest resulted in confinement, and a variable that characterizes the amount of time between preceding arrests.

Offenders were clustered according to life-long offense rates. A measure of individual heterogeneity was calculated based on the selected variables. Based on this heterogeneity score, a confidence interval was estimated for offense rate in the subsequent three years, in relation to the rest of the trajectory group. After a three 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 (Bersani, Nieuwbeerta, & Laub, 2009), group trajectory modeling was used to cluster a cohort of Dutch offenders to find predictive demographic factors for group membership. Overall predictive accuracy was informative, with a 71% accuracy rate, but accuracy for low-rate offenders, classic desister and chronic offender groups was under 10% (names of groups are qualitative descriptions of the shape of the offense rate plot over the span of the criminal career). Researchers cautioned against the use of risk assessment tools to support policy, and expressed skepticism that better results could be obtained using new analysis methods.

2.3 Event Sequence Mining

There are two major approaches to event sequence mining: (a) *sequential pattern mining*, and (b) *frequent episode mining*. (Blanchard, Guillet, & Gras, 2008). Sequential pattern mining is the discovery of subsequences that are deemed frequent if they occur in many input sequences. The approach was first introduced by Agrawal and Srikant (1995b) and is also commonly called “sequence mining” (Abraham, 2006b; Eichinger, Nauck, & Klawonn, 2006a; Spiliopoulou, 1999; Mohammed J Zaki, 2000). Frequent episode mining is the discovery of frequent subsequences in a single sequence, within a window of opportunity and was introduced by Mannila and Toivonen (1995a).

Data mining tasks shown to be suitable for sequence mining include classification (Eichinger et al., 2006a; Ferreira & Azevedo; Srivastava, Sural, & Majumdar; M.J. Zaki, Lesh, & Ogihara), clustering (Abraham, 2006b), and pattern discovery (Zhang, Zhou, Yang, & Zhong). Similarly, there are also examples of frequent episode mining used for

classification (Qin & Hwang, 2004b), clustering (Bathoorn, Welten, & Richardson, 2010a), and pattern discovery (Fujikawa, Kida, & Katoh, 2011a) tasks.

To guide the selection of an appropriate event sequence mining method, we return to the research problem and to the characteristics of the data. Key to expressing criminal histories as event sequences is the relationship between the antecedent and the subsequent. The approach to counting sequential patterns depends on the number of input sequences rather than number of occurrences. Thus, the frequency of a shorter pattern, when compared to a longer pattern does not capture the proportion of antecedents that lead to the subsequent. Rather, the relationship simply captures the proportion of individuals for whom the subsequent occurred at least *once*. However, frequent episode mining does not include a notion of multiple input sequences. A review of the frequent episode mining literature showed considerable disagreement on how to count the number of episodes (pattern occurrences). (Achar, Laxman, & Sastry, 2011) describes ten different methods of frequent episode support counting.

Event sequence mining is a problem with a complexity of $\Theta(m^k)$ where m is the number of possible itemsets and k is the number of elements in the discovered patterns. A number of researchers have addressed the challenge of mining the explosive number of possible frequent patterns by enhancing the efficiency of the algorithm with pattern growth (Pei, Han, & Wang, 2002) and vertical database (Gouda, Hassaan, & Zaki, 2007; Mohammed J Zaki, 2001) approaches. To narrow the focus to only the most meaningful rules, there has been some exploration of rule interestingness for sequence mining and frequent episode mining (Blanchard et al., 2008; Spiliopoulou, 1999) There have also been efforts to reduce the overall number of discovered patterns, such as the reduction to closed and maximal patterns (Yan, Han, & Afshar, 2003), as well as numerous efforts to introduce various domain and gap constraints (Leleu, Rigotti, Boulicaut, & Euvrard; Masseglia, Poncelet, & Teisseire, 2009; Pei et al.; Wang & Han, 2004).

With the introduction of constraints comes the challenge of finding appropriate parameters that might be suitable for a particular dataset. The current practice is to use operator-specified parameters for domain constraints, gap constraints, and rule interestingness parameters. To make event sequence mining accessible to more practitioners, there is a need to reduce this requirement for operator specified parameters.

The VOGUE algorithm combines sequence mining for pattern discovery with a hidden Markov model (HMM) that represents various gaps and elements as states (M J Zaki, 2010). This approach is suitable for protein sequences, where the size of the alphabet is relatively small. However, as the size of the alphabet and number of gap constraints grows, the HMM grows exponentially. Furthermore, as the number of sequence segments grows, precision of the Markov model suffers. Nonetheless, the success of this hybrid approach for biological sequence discovery, and the success of the aforementioned T-pattern discovery method (Magnusson, 2000) both support the case for a versatile gap constraint discovery mechanism.

2.4 Summary

There is a substantial body of literature in the field of developmental criminology involving criminal career trajectory analysis, but there is a need for a method to discover ad hoc relationships between different types of criminal life course events. Existing approaches to criminal career analysis involve classification and clustering, and focus on offense rate predictions. Research in behavioral pattern analysis has shown the usefulness of pattern discovery in sequential interactions using event types and interval constraints rather than event rate. Event sequence mining is applicable to classification, clustering and pattern discovery, but existing notions of pattern support and constraints are not well suited for expressing sequential relationships. Further, although event sequence mining does include the notion of gap constraints, where events must be separated by a given minimum and maximum time interval, it does not include the notion of periods of time during which an event does not occur.

In this work we address each of these limitations by designing and implementing a new type of event hazard pattern designed to capture periods during which an event does not occur (periods of desistance), and to accurately encode sequential relationships using a new measure of support.

CHAPTER 3

RESEARCH METHODOLOGY

This chapter describes the research methodology employed in this work. It also includes a brief review of the design science research guidelines as they shaped the methodological framework.

3.1 Design Science Research Methodology

Design Science is a problem solving process that produces knowledge and understanding. Hevner, March, Park, and Ram (2004) describe seven key guidelines for design science research:

Table 1: Research guidelines

1) Design science research must produce a viable artifact in the form of a construct, a model, a method, or an instantiation.	This work describes the design of a software instantiation designed to discover hazard patterns used to assess risk of recidivism and to justify the risk by concisely representing the antecedents leading to increases in relative risk of recidivism.
2) The objective of design science research is to develop technology-based solutions to important and relevant business problems.	The problem identified in this work is both pressing and important. Practitioners are faced with a Supreme Court mandate to reduce the California prison population by tens of thousands of individuals.
3) The utility, quality, and efficacy of a design artifact must be rigorously demonstrated via well-executed evaluation methods	Based on a review of the literature, a number of key requirements for a technological solution were identified. The discovered patterns should not be over-fitted to the training data. They should be meaningful and should codify the logic supporting a particular risk assessment. Finally, they should be generalizable over time. These form the design objectives and the evaluation criteria of the artifact.

<p>4) Effective design science research must provide clear and verifiable contributions in the areas of the design artifact, design foundations, and/or design methodologies.</p>	<p>The developed artifact is a contribution in its own right, extending the sub-field of event sequence mining in a new direction. Existing event sequence mining techniques were individually examined and found to be missing one or more key design requirements. The most fundamental of these is the cross-fertilization of survival analysis with event sequence mining. Hazard patterns are event sequence patterns that are frequent occurrences of time-to-event sequences (Janzen, Deokar, & El-Gayar, 2013a).</p>
<p>5) Design science research relies upon the application of rigorous methods in both the construction and evaluation of the design artifact.</p>	<p>The design and construction of the artifact was guided by gaps identified in the literature. In the domain space, there was an identified need for a risk assessment tool that both incorporates prior offending history and codifies that risk in a logical manner. In the solution space, there was a need for a way to express durations of desistance, and to express the relationship between antecedent event sequences and subsequent events.</p>
<p>6) The search for an effective artifact requires utilizing available means to reach desired ends while satisfying the laws in the problem domain</p>	<p>The design of the artifact was conducted through an iterative generate/test cycle. The first iteration lead to the exploration of event sequence mining as a potential tool to demonstrate and explain the sequential relationships in criminal histories. However, existing approaches to event sequence mining did not account for periods during which events do not occur. Further, existing event sequence support counting methods were not well suited for quantifying the relationship between antecedent sequences and their subsequent extensions. The existing methods were examined in detail and formed the requirements for a new approach to event sequence mining, developed in a second iteration of the development cycle. These were then evaluated according to identified domain space objectives. Limitations identified during this latest iteration will form the requirements of future work.</p>
<p>7) Design science research must be presented effectively both to technology-oriented as well as management-oriented audiences</p>	<p>In the course of the development of a solution to the presenting problem, two conference presentations were made. Hazard patterns with heterogeneous constraints were introduced in a presentation at a technology-oriented conference in (Janzen, Deokar, & El-Gayar, 2013b). Subsequently, the use of time oriented pattern selection mechanisms was presented to a technical and managerial audience in (Janzen et al., 2013a).</p>

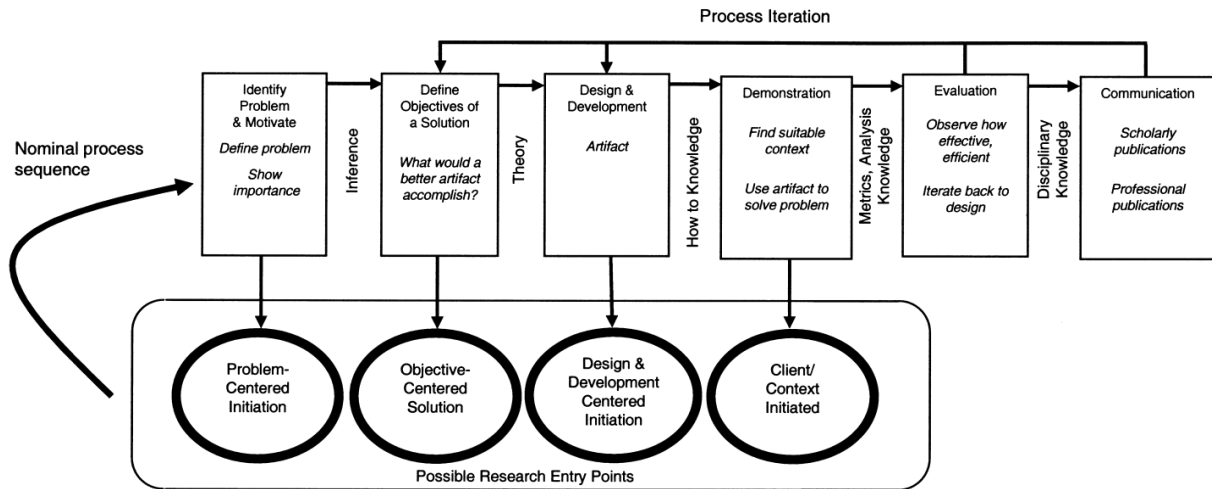


Figure 1: Nominal DSRM process model (Peffer et al., 2007)

The research process for this work follows the Design Science Research Methodology (DSRM) proposed by Peffer, Tuunanen, Rothenberger, and Chatterjee (2007). A DSRM nominal process sequence, with a problem-centered initiation, is applicable to this research study. The problem importance has been demonstrated as an important and current question asked by researchers and practitioners in the field of criminology. The literature was consulted to learn what has already been accomplished toward addressing the problem, and to what extent existing solutions have been found effective. This has led to the identification of specific missing advances. Based on the motivation from the problem domain, and based on the key missing advances identified in the literature review, specific design objectives have been formulated. Based on these design objectives, a new algorithm and an encompassing crime analytics prototype system was designed and implemented, then demonstrated in the application domain, and subsequently evaluated. Finally, the results were communicated through conference presentations and scholarly publications.

3.3 Objectives of a Solution

Based on the process review of the PVDMI pilot deployment in California, practitioners lacked confidence in the logic supporting the tool's risk determinations, and did not believe the tool properly accounted for changes in risk associated with repeat offending behavior (Turner et al., 2012). We are presented with the challenge, not only of assessing risk,

but of justifying that risk assessment to a decision maker, particularly with respect to prior offending record.

We identified two key requirements that a recidivism risk assessment tool should satisfy:

1. Incorporate salient characteristics of prior record to determine risk.
2. Concisely present the logic leading to the determined risk level.

Each individual's prior record consists of a series of discrete events over the course of the criminal career. Such events include arrest and charge, disposition, parole, and discharge. A risk assessment based on criminal history is a static risk assessment – a determination based on factors that cannot be changed. The PVDMI utilizes the California Static Risk Assessment Instrument (CSRA). This instrument incorporates indicators for repeat offending behavior by including counts for number of incidents of various types, such as convictions, sentences, and supervision violations. More details about this instrument are available in a working paper (Turner, Ph, Hess, & Jannetta, 2009).

One way to represent a criminal history is as an ordered sequence of many different types of events. Event sequences that occur frequently can be represented as patterns for classification, clustering, or prediction tasks. Hazard patterns are frequent sequences of events where each successive event in a pattern represents the first subsequent event of that type, and where the time between events in a pattern represents time-to-failure or time-to-event (Janzen et al., 2013b). As already noted in (Bhati & Piquero, 2007), time between preceding arrests is a useful predictor of future arrest risk.

A hazard pattern representing a history of many arrest charges for various offenses will also capture periods of desistance, during which no arrest occurred. Hazard patterns draw on survival analysis techniques, and allow the analyst to include potentially significant information about time between events. However, to demonstrate usefulness and reliability of these patterns for risk assessment, we must address some important concerns:

Over-fitting:

Are the patterns generalizable to other similar data sets?

Meaningfulness:

Are patterns found even when there are no patterns in the data?

Predictiveness:

Can patterns discovered in the past be useful predictors of future behavior?

Parsimony:

Can we produce output with minimal redundancy?

Each of these concerns must be addressed if a risk assessment tool is to be useful and credible.

A common way to address the concern of over-fitting is to rely on some form of validation on a hold-out sample. One portion of the data set is set aside only for validation, while the remainder of the dataset is used to train the model. A straightforward strategy is to split the data in half. In cases where there is little available data, setting aside half of the data for validation may substantially disadvantage the model training process. For this reason, it can be advantageous to perform k -fold validation. The data is divided into k subsets, each of which serves as a validation set for a model trained using the remainder of the data. If the discovered patterns are characteristic of the population, they will also be found in the hold-out sample. If the discovered patterns are merely characteristic of the training data, few patterns found in the training data will be found in the hold-out sample.

The second issue of meaningfulness is both subtle and important. We applied a test for *meaningless* results, to support the belief that the discovered patterns are meaningful. Keogh and Lin (2004) presented the surprising result that a subsequence clustering technique used in dozens of published papers produced meaningless results. For our test, we draw on their formal definition: “We call an algorithm *meaningless* if the output is independent of the input.” We can prepare a minimally differentiated data set where all of the same events occur, but their order is randomly shuffled. If the discovered patterns are dependent on the ordering of the underlying events, the tool should not discover any patterns in the shuffled data.

The third concern of predictiveness is of vital importance. We can demonstrate predictiveness by showing that patterns discovered in one time period can reliably predict arrest risk in a subsequent time period. This is also the most difficult test, since patterns learned in the past cannot account for future changes in the environment.

Finally, to avoid producing an output of thousands or even millions of patterns that may or may not be useful, we must apply a pattern selection strategy. One approach is to apply a test of statistical significance to determine whether complex patterns predict risk that is different from simpler alternatives. However, there is a danger that, in a large enough sample, some patterns will appear to be significant due to chance alone. If we rely on

statistical tests of significance for pattern selection, we must also guard against, and if necessary, correct for multiple testing bias.

CHAPTER 4

THEORY AND ARTIFACT DESIGN

In this section we describe the system design by example, using a collection of three contrived criminal histories. We first present the events on a time line, and then refer to this example as we describe our search for a solution.

4.1 Hazard Patterns

Figure 2 contains three contrived example criminal histories for individuals a, b , and c . The data set used for demonstration of this pattern discovery tool includes many more event types including a range of arrest charges and dispositions. This example is simplified for the sake of illustration. However, even in this simplified case, it is difficult to see whether there might be a pattern between antecedent events and arrest after parole release. Keeping in mind that these are contrived histories; can we find a relationship between past behavior and risk of arrest after parole release?

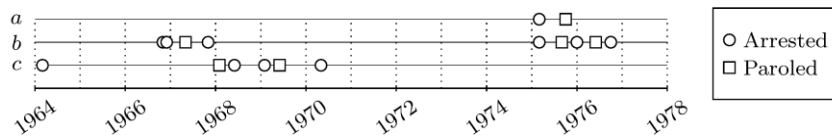


Figure 2: Contrived histories

Frequently occurring patterns of events might be used to discover behavior patterns that are characteristic of a particular type of offender or that are indicative of increased re-arrest risk.

An **event occurrence** is denoted (e, t) , where e represents the event type and t represents the time of the event occurrence. For example, $(Paroled, 909)$ is the occurrence of event (or event type) *Paroled* at time 909 (in this case, the number of months since the beginning of the 20th century).

An **event sequence** of length n is denoted $\langle (e_1, t_1), (e_2, t_2), \dots, (e_n, t_n) \rangle$ where e_i represents the type of the i^{th} event, t_i represents the time of the i^{th} event, and $t_{i-1} < t_i$. An event sequence is a time oriented arrangement of event occurrences. For example, $\langle (Paroled, 833), (Arrested, 844) \rangle$ is an event sequence. Note that in this work we address only serial event sequences, where each subsequent event occurs after the preceding event.

Event sequence mining has been widely used to discover characteristic patterns for classification (Eichinger et al., 2006b; Ferreira & Azevedo, 2005; Qin & Hwang, 2004a; Srivastava, Sural, & Majumdar, 2006b; M.J. Zaki, Lesh, & Ogihara, 1998a), clustering (Abraham, 2006a; Bathoorn, Welten, & Richardson, 2010b), and pattern discovery (Fujikawa, Kida, & Katoh, 2011b; Zhang, Zhou, Yang, & Zhong, 2010b). In each of the examples of classification and clustering, event sequence mining was used indirectly, to produce the input for classification or clustering algorithms. Common application domains are the analysis of biological sequences and malicious activity.

Constraints between event occurrences

To discover time-based relationships between events in a criminal history, it is useful to apply a constraint that limits the amount of time between the events of interest. In event sequence mining, *mingap* and *maxgap* constraints can be applied for this purpose.

A **gap constraint** is 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*. For a more detailed discussion of gap constraints, see (Leleu et al., 2003).

The practical application of gap constraints brings with it the challenge of selecting useful minimum and maximum allowable gaps. The discovered patterns will vary greatly depending on the operator-specified parameters. Further, some patterns with both short term and long term relationships will only be found when multiple different constraints are specified. For instance, Han and Dong (1999) describe the strategy of combining of weekly and yearly periodicity in patterns, and Giannella, Han, Pei, Yan, and Yu (2004) applied a tilted time window framework, mining for patterns in windows of 15 minutes, 24 hours and 31 days.

We can apply multiple different gap constraints to a criminal event history to characterize clustering of events over time. However, gap constraints do not include information about whether the event of interest occurs additional times prior to *mingap*. To describe a relationship between previous events and time to re-arrest, hazard patterns with heterogeneous constraints were proposed in (Janzen et al., 2013b).

A **hazard pattern** is a frequently occurring sequence of events where each subsequent event occurrence is the first subsequent occurrence of that particular event type. A hazard pattern can be denoted as *Paroled* \rightarrow *Arrested*. For all occurrences of this pattern, *Arrested* refers to the first arrest after parole release.

We can also apply a constraint whereby the period of time between two events in a pattern must fall within a specified minimum and maximum time interval. A given hazard pattern *Paroled* \rightarrow *Arrested* can be expressed with a hazard constraint as *Paroled* $\overrightarrow{(\minhaz, \maxhaz)}$ *Arrested*. For instance, occurrences of *Paroled* $\overrightarrow{(3,6]}$ *Arrested* satisfy the condition that more than three and at most six months elapsed between parole release and the next arrest. We can also apply hazard constraints of increasing sizes, to capture patterns that reflect both short term and long term relationships.

Table 2: Months until re-arrest (contrived data set)

Antecedents	$\overrightarrow{(0,3]}$	$\overrightarrow{(3,6]}$	$\overrightarrow{(6,12]}$	$\overrightarrow{(12,24]}$	$\overrightarrow{(24,96]}$	Total
RR ($p \leq$)						
<i>Arrested</i>	0.09	0.00	0.36	0.09	0.18	0.73
s - / d 11 / i 3	1/1/1	0/0/0	4/4/2	1/1/1	2/2/2	8/8/2
RR ($p \leq$)		1.50(<i>ns</i>)				
<i>Arrested</i> $\overrightarrow{(3,6]}$ <i>Paroled</i>	0.00	0.75	0.25	0.00	0.00	1.00
s 5 / d 4 / i 2	0/0/0	3/3/1	1/1/1	0/0/0	0/0/0	4/4/2
RR ($p \leq$)						
<i>Paroled</i>	0.00	0.67	0.17	0.00	0.00	0.83
s - / d 6 / i 3	0/0/0	4/4/2	1/1/1	0/0/0	0/0/0	5/5/2

Counting pattern occurrences

A straightforward way to describe relationships between antecedent patterns and subsequent events is to describe the proportion of antecedents that lead to the subsequent. For instance, the relationship can be expressed as the occurrences of *Paroled* that also occur in *Paroled* $\overrightarrow{(3,6]}$ *Arrested* as shown in Table 2.

Table 3: Months until paroled (contrived data set)

Antecedents	$\overrightarrow{(0,3]}$	$\overrightarrow{(3,6]}$	$\overrightarrow{(6,12]}$	$\overrightarrow{(12,24]}$	$\overrightarrow{(24,96]}$	Total
RR ($p \leq$)						
<i>Arrested</i>	0.00	0.45	0.18	0.00	0.18	0.82
s - / d 11 / i 3	0/0/0	5/4/2	2/2/2	0/0/0	2/2/2	9/8/3

There are six occurrences of *Paroled*. Of these six *Paroled* events, we see in Table 3 that five lead to re-arrest within 4-12 months, aggregated as a proportion of 0.83 re-arrests per parole release. When more than one antecedent parole release leads to the same subsequent arrest, counting varies based on the approach to event sequence mining.

Event sequence mining includes both sequential pattern (or sequence) mining, and frequent episode mining. The support of an event sequence pattern is a measure of pattern frequency. Sequential patterns are frequent if they occur in many input sequences (Agrawal & Srikant, 1995a). Event sequences are frequent if they occur in many windows of opportunity (Mannila & Toivonen, 1995b). Sequential patterns may be useful for discovering commonalities between offenders, but since our primary interest is to predict future risk of recidivism, we primarily examined support counting methods based on opportunities, as used in frequent episode mining.

There is no agreed upon way to count the number of event sequence pattern occurrences. For instance, (Achar et al., 2011) describes 10 different support counting methods, each of which was evaluated for the presenting problem. For all support counting methods we encountered, one or more of the following were true: (a) counts were non-independent of other occurrences of the same pattern (non-overlapping, non-interleaved, distinct occurrence, and minimal window based), (b) longer patterns were unduly penalized (window and expiry time based) and (c) unrelated event occurrences can inflate support counts (head frequency, total frequency).

In the case of non-overlapping, non-interleaved, and distinct occurrence based patterns, a pattern occurrence that would otherwise have been counted might not be counted due to the existence of other pattern instances. *Distinct* patterns may not share any events in common. For instance, in event sequence

$\langle (c, 1), (a, 2), (a, 3), (b, 4), (a, 5), (b, 6), (c, 7), (d, 8) \rangle$, there are three potential occurrences of $a \rightarrow a$: $\langle (a, 2), (a, 3) \rangle$, $\langle (a, 3), (a, 5) \rangle$, and $\langle (a, 2), (a, 5) \rangle$. Since each of these shares an event with common. For instance, in event sequence

$\langle (c, 1), (a, 2), (a, 3), (b, 4), (a, 5), (b, 6), (c, 7), (d, 8) \rangle$, there are three potential occurrences

of $a \rightarrow a$: $\langle(a, 2), (a, 3)\rangle$, $\langle(a, 3), (a, 5)\rangle$, and $\langle(a, 2), (a, 5)\rangle$. Since each of these shares an event with at least one of the other potential pattern occurrences, we only count one distinct occurrence. *Non-interleaved* patterns maintain the relative order of their events with events in at least one of the other potential pattern occurrences. *Non-interleaved* patterns maintain the relative order of their events with events in other pattern occurrences. For instance, in the above example, $\langle(a, 2), (b, 4)\rangle$ and $\langle(a, 3), (b, 6)\rangle$ are non-interleaved but $\langle(a, 2), (b, 6)\rangle$ and $\langle(a, 3), (b, 4)\rangle$ are interleaved. *Non-overlapping* patterns occur in distinct time spans. For instance, $\langle(a, 2), (b, 4)\rangle$ and $\langle(a, 5), (b, 6)\rangle$ are non-overlapping but $\langle(a, 2), (b, 4)\rangle$ and $\langle(a, 3), (b, 6)\rangle$ are overlapped. *Minimal occurrence-based* support includes only those occurrences during which there exist no occurrences of the same pattern over a smaller time window. For instance, $\langle(a, 2), (b, 4)\rangle$ is not a minimal occurrence because it contains $\langle(a, 3), (b, 4)\rangle$ in a sub-window. The interdependencies between pattern occurrences make it difficult to establish a relationship between a shorter pattern and an extension of that pattern. For instance, even the simple relationship between an antecedent a and the subsequent extension to $a \rightarrow c$ is not accurately represented by the differing support counts for each of these patterns. In fact, for each of these counting methods, there are three occurrences of a and only one occurrence of $a \rightarrow c$, leading us to conclude that these support counting methods cannot be used to express a sequential relationship between an antecedent a and a subsequent c . For each of these approaches, any single occurrence of $a \rightarrow c$ is not independent of another occurrence of the same pattern.

Window and expiry-time constraint based counting methods disproportionately penalize longer patterns. In the case of an expiry-time constraint, the time between the first and last events in a pattern occurrence must not occur farther apart than a specified expiry time. This has no impact on patterns consisting of a single event, such as a , but for $a \rightarrow c$ in the preceding example, the choice of expiry time constraint dramatically affects the support count. For window-based support, the number of windows that contain at least one pattern occurrence are counted. In addition to the impact on longer patterns as seen with an expiry time constraint, window-based counting adds a further distortion by over-inflating the prevalence of very short patterns. For instance, there are 5 windows of size 2 that contain an occurrence of a , and only 1 window of size 2 that contains an occurrence of $a \rightarrow c$. Given the penalty against longer patterns and given the inflation of support for shorter patterns, we were

unable to use window-based and expiry-time based counting methods for modelling sequential relationships.

We also considered head frequency and total frequency counting methods. Head frequency is the number of windows of opportunity that begin with the first event in the pattern. This is quite effective for patterns consisting of up to two events, but is problematic for longer patterns, particularly when the first event is very frequent. For instance, the head frequency of $a \rightarrow c$ for a window size of 5 is 3 and the head frequency of $a \rightarrow c \rightarrow d$ is 2 for the same window size. The frequent a over-inflates the number of potential $a \rightarrow c$ that might lead to a subsequent d . Total frequency partially addresses this problem by counting support as the lowest head frequency of any sub-pattern. The support of $a \rightarrow c$ is limited by the support of c . However, as we see in the above example, the support of sub-patterns may depend on completely unrelated occurrences, in this case, $(c, 1)$. Thus, in addition to the penalty against long sequences introduced by use of a window of opportunity, both head frequency and total frequency cannot be used to accurately describe the relationship between an antecedent and subsequent because the support counts can be distorted by unrelated events (frequent head, sub-pattern occurrences without the antecedent of interest).

To be able to adequately express the relationship between the antecedent and the subsequent, a new measure of support was proposed in Janzen et al. (2013b). **Relative Support** is the number of distinct or unique antecedent event occurrences that are followed by a subsequent event of a particular type in a hazard pattern. For instance, in Table 3, of 11 occurrences of *Arrested*, 5 are eventually followed by *Paroled* in 4-6 months. However, there are only 4 distinct occurrences of *Paroled* that participate in this relationship. Two of the *Arrested* events lead to one *Paroled* event (see individual b in Figure 2). Further applying this concept, in Table 2, the antecedent pattern $Arrested \overline{(3,6]} Paroled$ has a support of 5, but we only consider 4 distinct antecedents when calculating the proportion that participates in $Arrested \overline{(3,6]} Paroled \overline{(3,6]} Arrested$.

Selecting interesting patterns

An additional problem we faced, particularly in a large data set, is the large number of patterns discovered. To determine whether a particular pattern might convey useful information, we can calculate a measure of interest and apply a statistical test of significance.

Relative Risk is 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 (Bewick, Cheek, & Ball, 2004).

We evaluated the use of pattern selection using significance tests on Relative Risk (RR). Patterns shown to significantly affect the RR coefficient in training data were also shown to have a similar effect in test data. For further details, see (Janzen et al., 2013a).

RR expresses the ratio between survival proportion in a treatment group compared to the same in a control group. Since we value parsimony, to reduce the number of patterns that a decision maker might need to review, we compare the RR for a presented pattern with the RR for the same pattern with the first antecedent removed. In Table 2, RR could only be calculated in this way for one of the patterns. The risk of arrest in the four distinct antecedent parole releases in *Arrested* $\overline{(3,6]}$ *Paroled* ($\frac{3}{4}$) is compared against the risk of arrest in the two distinct parole releases in *Paroled* that are not already counted in *Arrested* $\overline{(3,6]}$ *Paroled* ($\frac{1}{2}$). The RR of 1.50 indicates that the risk of re-arrest during the subsequent 4-6 months is one and a half times higher if parole release is 4-6 months after arrest. A RR of 1 indicates no change. To see whether the increase in risk might be generalizable to the broader population, we draw on the statistical significance of RR. In this case, as we can expect with such a small sample, the resulting Z-score of 0.32 ($p > 0.05$) indicates that we do not have enough evidence to conclude that the RR is different than 1.

4.2 Algorithm design

Data structures

To facilitate indexing, constraint and offset values were stored in a lookup table. Events were encoded as integers, constraints of increasing sizes were represented as successive integers, and offset values were represented as ordinals. Offsets were kept separated per individual, as illustrated in Figure 3.

		ordinal	offset	individual		
		0	902	0		
		1	909	0		
		2	802	1		
		3	803	1		
		4	808	1		
		5	814	1		
		6	902	1	encoded	constraint
encoded	event	7	908	1	0	(0, 0]
0	Arrested	8	912	1	1	(0, 3]
1	Paroled	9	917	1	2	(3, 6]
		10	921	1	3	(6, 12]
		11	770	2	4	(12, 24]
		12	817	2	5	(24, 96]
		13	821	2		
		14	829	2		
		15	833	2		
		16	844	2		

Figure 3: Example offset and constraint lookup tables

Using these simplified representations of events, offsets and constraints, an index was constructed to enable easy lookup of both when the next event of a given type might occur, as well as what constraint is satisfied by that occurrence.

Next Ordinal Index																	
Ordinal	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Arrested	0	0	3	5	5	6	8	8	10	10	0	13	13	14	16	16	0
Paroled	1	0	4	4	7	7	7	9	9	0	0	12	15	15	15	0	0

Constraint Index																	
Ordinal	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Arrested	0	0	1	3	2	5	3	2	3	2	0	5	2	3	4	3	0
Paroled	3	0	2	2	6	5	2	3	2	0	0	5	4	3	2	0	0

Figure 4: Ordinal and constraint indexes

For instance, by referencing ordinal 6 in Figure 4 we see in the ordinal index that the next *Arrested* event occurs at ordinal 8 and we see in the constraint index that constraint 3 is satisfied for that next ordinal. A value of zero indicates that there is no applicable next ordinal. Note that histories of different individuals are indexed back to back. Ordinals 1, 10 and 16 contain zeroes because the subsequent ordinal belongs to a different individual's history.

We can also see convergence from multiple antecedents to a single subsequent. For ordinals 5 and 6 (columns 5 and 6), we see that two distinct *Arrested* events occurred. For each of the arrest events, the next *Paroled* event is the same occurrence. This convergence is also seen in Figure 2 in individual *b* at the end of 1966, and is the reason for the difference

between the support and the distinct count in $Arrested\overline{(3,6]}Paroled$ (see Table 3). The algorithms used to construct the above indexes and lookup tables are not detailed here.

Pattern discovery algorithm

The discovery of frequent patterns follows a depth first tree-traversal search pattern (though breadth first traversal is certainly possible due to lack of dependencies between search branches, affording an opportunity for parallel processing) Frequent antecedent ordinals are collected, and for each type of subsequent event, the subsequent ordinals are grouped according to the constraint they satisfy (each ordinal satisfies only one constraint). Within each type of subsequent event, constraint groupings that are larger than a specified support threshold become candidates for further extension.

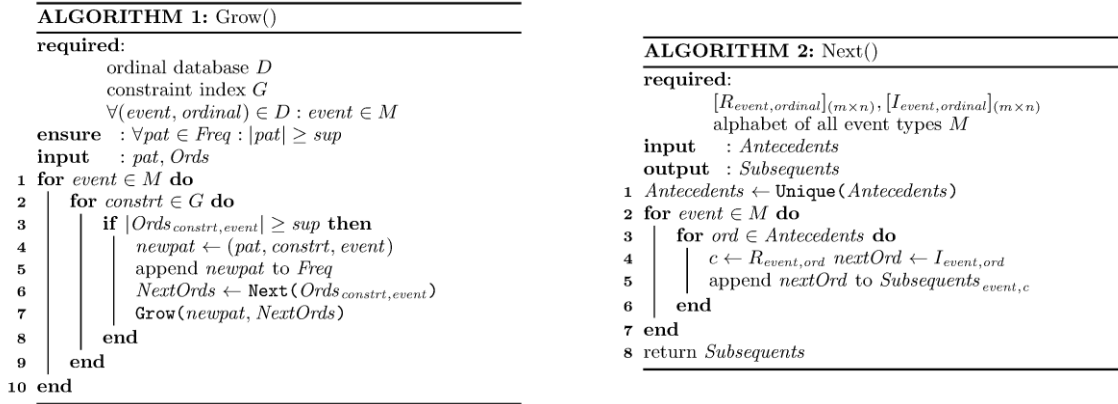


Figure 5: Pattern discovery algorithms

The Grow function shown in Figure 5 relies on the constraint and ordinal indexes. 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, and are passed to the Next function, where a new matrix of candidate ordinals is created, and passed to the subsequent recursive Grow attempt on line 7.

The Next function in Algorithm 2 takes as input a collection of antecedent ordinals, grouped by event, and produces the Ordinal matrix $NextOrds$ needed in line 6 of algorithm 1. This function uses two indexes: $R_{event,ordinal}$ and $I_{event,ordinal}$. See Figure 4 for

the R and I indexes corresponding to event histories shown in Figure 2. Ordinals in I and constraint identifiers in R have corresponding values in the lookup tables shown in Figure 3. 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 4 of the Next function pseudo-code in Algorithm 2, 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 4 the subsequent ordinals are retrieved from I and then grouped according to their matching constraints in line 5. 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.

CHAPTER 5

DEMONSTRATION AND EVALUATION

In this section, we demonstrate the results obtained using the pattern discovery system to mine a data set of real life criminal histories. We then evaluate the pattern discovery system according to the four design objectives described in chapter 3:

Over-fitting:

Are the patterns generalizable to other similar data sets?

Meaningfulness:

Are patterns found even when there are no patterns in the data?

Predictiveness:

Can patterns discovered in the past be useful predictors of future behavior?

Parsimony:

Can we produce output with minimal redundancy?

5.1 Demonstration: Pattern discovery using representative data

The pattern discovery system was used to discover patterns in two related data sets. Data set *A* consisted of complete criminal histories from a non-random sample of offenders who entered the California Youth Authority's Deuel Vocational Institute in 1964 and 1965. The event database contains 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 (Wenk, 2006).

For this analysis, the individual histories in the data set were randomly assigned to either the training set or the testing set. 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.

Table 4: Arrests after parole (data set A)

Antecedents	$\overrightarrow{(0,3]}$	$\overrightarrow{(3,6]}$	$\overrightarrow{(6,12]}$	$\overrightarrow{(12,24]}$	$\overrightarrow{(24,96]}$	$\overrightarrow{(96,384]}$	Total
RR ($p \leq$)	1.13(0.001)	1.04(<i>ns</i>)	1.01(<i>ns</i>)	1.02(<i>ns</i>)	0.95(<i>ns</i>)	0.85(<i>ns</i>)	
<i>Convicted</i> $\overrightarrow{(6,12]}$ <i>Paroled</i>	0.30	0.21	0.20	0.15	0.10	0.01	0.96
s 4115 / d 3497 / i 2364	1053/1053/863	717/717/643	704/704/645	528/528/496	334/334/330	29/29/29	3365/3365/2261
RR ($p \leq$)	1.11(0.001)	1.05(<i>ns</i>)	1.00(<i>ns</i>)	0.79(<i>ns</i>)	0.89(<i>ns</i>)	1.61(0.05)	
<i>Convicted</i> $\overrightarrow{(12,24]}$ <i>Paroled</i>	0.30	0.21	0.20	0.13	0.09	0.01	0.94
s 5885 / d 4087 / i 2720	1210/1210/981	839/839/755	815/815/739	536/536/508	376/376/368	47/47/47	3823/3823/2563
RR ($p \leq$)	0.87(<i>ns</i>)	0.97(<i>ns</i>)	0.93(<i>ns</i>)	0.99(<i>ns</i>)	1.18(0.05)	0.73(<i>ns</i>)	
<i>Convicted</i> $\overrightarrow{(24,96]}$ <i>Paroled</i>	0.25	0.20	0.19	0.15	0.11	0.01	0.91
s 6871 / d 3042 / i 2038	774/774/642	595/595/529	579/579/528	451/451/428	332/332/328	23/23/23	2754/2754/1885
RR ($p \leq$)	1.48(0.001)	1.18(0.001)	0.97(<i>ns</i>)	0.62(<i>ns</i>)	0.58(<i>ns</i>)	0.45(<i>ns</i>)	
<i>Paroled</i> $\overrightarrow{(12,24]}$ <i>Paroled</i>	0.38	0.23	0.19	0.10	0.06	0.00	0.97
s 1504 / d 1504 / i 1065	577/577/473	345/345/306	292/292/269	149/149/146	92/92/91	7/7/7	1462/1462/1042
RR ($p \leq$)	1.04(<i>ns</i>)	1.02(<i>ns</i>)	0.93(<i>ns</i>)	1.00(<i>ns</i>)	0.89(<i>ns</i>)	0.89(<i>ns</i>)	
<i>Paroled</i> $\overrightarrow{(24,96]}$ <i>Paroled</i>	0.29	0.20	0.19	0.15	0.09	0.01	0.93
s 2886 / d 2886 / i 1682	833/833/655	584/584/489	549/549/478	430/430/388	263/263/258	25/24/24	2684/2683/1592

Table 4 contains the patterns with a minimum support threshold of 500, plus their neighboring stubs. Stubs are those patterns that would otherwise be excluded due to low support, but which are siblings of a frequent pattern. For instance, if the subsequent event occurs frequently in the follow-up period of (0,3], we also tabulate the number of occurrences in the adjacent follow-up periods, and calculate a total. Table 4 shows only patterns with an antecedent ending with *parole* and a subsequent event of *Arrest*. To reduce redundancy, antecedents with an *Arrested* antecedent event were also excluded from the table. Patterns with a RR value that is significantly different from 1 are presented in bold. We see that the recidivism is generally high in this group. These rates do not represent the general population. There are three relevant considerations to keep in mind when interpreting these patterns. First, the data set contains only male offenders who were young offenders in 1963/1964. In other words, late onset offenders and females were not included. Second, since the mining process specifically selected frequently occurring patterns, it is not surprising that these patterns would reveal sub-groups with high recidivism rates. Third, since there are 54,175 arrest records for 3,652 individuals, each individual had on average 14.83 arrests, all but one of which was their final arrest, so a high recidivism rate is not surprising in this data set.

We note several relationships between criminal history and recidivism. Of all the follow-up periods, even though the (0,3] time interval is the smallest, it also tends to be the time period with the highest support counts. Over all, there is only a small amount of variation between the groups represented by each pattern. RR values for shorter follow-up periods are closer to 1, with larger Z-scores, and RR values for longer follow-up periods are farther from 1, with smaller Z-scores. The patterns provide more generalizable information about the short

follow-up periods. Past repeat offending over (0,3] increases risk of repeating the same when released on parole after (12,24]. Generally, those who are released on parole sooner also re-offend sooner than others. The single strongest relationship is shown in the last two patterns. Time since previous parole release has a large and significant impact on recidivism. Individuals released on parole (12,24] after their previous parole release are almost 1.48 times as likely to re-offend within 3 months when compared to all others released on parole, and are significantly less likely to wait to re-offend until the subsequent follow-up periods when compared to all other parolees. Using the same data set, mined at a lower minimum support threshold, we observed other patterns relating to specific arrest charges, dismissals, and convictions.

Table 5: New offense after parole contact (data set B)

Antecedents	$\overrightarrow{(0,7]}$	$\overrightarrow{(7,30]}$	$\overrightarrow{(30,180]}$	$\overrightarrow{(180,720]}$	Total
RR ($p \leq$)	2.602(0.001)	0.933(ns)	1.194(ns)	2.463(0.001)	
$maxf \overrightarrow{(0,7]}maxf$	0.034	0.006	0.027	0.102	0.168
s 2165 / d 2165 / i 716	73/33/33	12/10/10	59/46/42	220/105/100	364/194/134
RR ($p \leq$)	0.735(ns)	1.730(0.001)	3.585(0.001)	6.201(0.001)	
$maxf \overrightarrow{(7,30]}maxf$	0.005	0.029	0.087	0.031	0.152
s 5930 / d 5930 / i 980	30/30/29	170/135/106	517/171/153	182/50/50	899/386/188
RR ($p \leq$)	1.071(ns)	1.020(ns)	2.999(0.001)	3.442(0.001)	
$maxf \overrightarrow{(7,30]}maxf \overrightarrow{(7,30]}maxf$	0.005	0.027	0.081	0.030	0.143
s 3660 / d 3660 / i 784	19/19/19	97/77/67	295/120/115	111/32/32	522/248/144
RR ($p \leq$)	0.405(ns)	1.233(ns)	1.712(0.001)	3.294(0.001)	
$maxf \overrightarrow{(7,30]}maxf \overrightarrow{(7,30]}maxf \overrightarrow{(7,30]}maxf$	0.003	0.026	0.069	0.030	0.129
s 2351 / d 2351 / i 592	8/8/8	62/49/45	163/72/70	71/23/23	304/152/103
RR ($p \leq$)	0.617(ns)	1.153(ns)	2.602(0.001)	3.441(0.001)	
$maxf \overrightarrow{(7,30]}maxf \overrightarrow{(7,30]}maxf \overrightarrow{(7,30]}maxf \overrightarrow{(7,30]}maxf$	0.017	0.041	0.105	0.031	0.192
s 2555 / d 1799 / i 742	30/30/29	73/73/65	188/117/111	55/33/33	346/253/155
RR ($p \leq$)	2.282(0.001)	2.343(0.001)	2.506(0.001)	2.348(0.001)	
$mxph \overrightarrow{(0,7]}mxph$	0.011	0.047	0.111	0.035	0.204
s 2275 / d 1939 / i 731	21/19/18	91/73/59	216/124/113	68/34/34	396/250/149
RR ($p \leq$)	1.641(0.05)	1.631(0.001)	2.336(0.001)	4.888(0.001)	
$mxph \overrightarrow{(0,7]}mxph \overrightarrow{(0,7]}mxph$	0.031	0.052	0.122	0.049	0.254
s 1325 / d 1325 / i 421	41/26/23	69/42/34	161/72/67	65/23/23	336/163/107
RR ($p \leq$)	1.133(ns)	1.465(0.05)	2.608(0.001)	4.324(0.001)	
$mxph \overrightarrow{(7,30]}mxph \overrightarrow{(7,30]}mxph$	0.025	0.050	0.114	0.037	0.226
s 1329 / d 1329 / i 499	33/33/28	67/55/46	152/86/81	49/26/26	301/200/119

We also applied the same process to the probation data set *B* originally collected for an evaluation of intensive probation in Milwaukee. Hazard pattern mining was performed on chronological records of violations and probation contacts of 1781 probationers. There were a total of 47,169 contacts, under a minimum (5396), medium (7977), and maximum (33738) intensive parole supervision (58 contacts did not include information on supervision level). Contacts included face to face, phone call, and mail with either the probationer or a collateral. We focused on face to face and phone contact with the probationer, and on their relationship with subsequent new offenses. There were 23,276 contacts with 1,744 probationers that satisfied these criteria: Under maximum supervision, there were 10,839 face to face (mxmf) contacts and 4,739 phone (mxph) contacts. Under medium supervision, there were 3,068 face

to face (mdf) contacts and 1,170 phone (mdph) contacts. Under minimum supervision, there were 1,370 face to face (mnf) contacts and 549 phone (mnph) contacts. There were also 434 rules violations and 144 new offenses. The remaining contacts were coded as other, blank, or missing. Since contacts were very frequent and coded by number of days since a fixed point, we selected correspondingly granular constraints of (0,7], (7,30], (30,180], and (180,720] days between events. With the exception of 5 outliers, all of these contacts took place within a span of 780 days. Overall, there were about 13 contacts per probationer, but only 1 new offense per 12 probationers. With the relatively rare occurrence of new offenses, this data set provides a strong contrast with the criminal career data from California (data set A).

Based on the patterns shown in Table 5, we note several relationships between antecedent patterns and subsequent events. Parolees who were in contact with their parole supervisor at intervals of 8-30 days were not at increased risk of violation within 7 days. However, parolees with supervisor contacts of 7 or less days apart were significantly more likely to commit a violation within 7 days. A plausible explanation for this relationship, given that all antecedents involved maximum supervision, is that very high risk parolees are simply contacted more frequently by their parole supervisor.

5.2 Evaluation

Over-fitting

In the case of a very complex pattern discovery system, it may be possible to over-fit the characteristics of the training set. The discovered patterns may describe the training data perfectly, but they may not be generalizable to other similar data. To test against this, we performed a k-fold cross-validation with ten folds. Each fold consisted of a 90% training split and a 10% testing split. We selected patterns based on a RR Z score outside ± 1.96 . For each fold, we considered contradictions to be those cases where the training split and the testing split each reported a significant Z score of opposite sign. We recorded consistency where a significant Z score in the testing split corresponded to a Z score of the same sign in the training split. Ten-fold cross validation was performed on a randomly selected sub-sample of 500 individuals without replacement. For each fold, pattern mining was performed with a minimum support threshold of 100 (note that support is determined by number of pattern occurrences, not number of individuals). We tabulated the above indicators for each of the ten

folds, and repeated the process for the same 500 individuals with all events shuffled. The results are shown in the first half of Table 6.

Table 6: Cross validation (data set A)

Shuffled	Fold	Patterns	Sig	Contradict	Consistent
False	0	11756 3551		0.00	0.97
False	1	12163 3510		0.00	0.97
False	2	11633 3766		0.00	0.98
False	3	11788 3649		0.00	0.97
False	4	11420 3682		0.00	0.97
False	5	11770 3690		0.00	0.97
False	6	11925 3652		0.01	0.97
False	7	12031 3549		0.01	0.97
False	8	11837 3549		0.00	0.97
False	9	11508 3503		0.00	0.97
True	0	5903 1362		0.01	0.96
True	1	5919 1493		0.01	0.94
True	2	5872 1556		0.01	0.95
True	3	5877 1552		0.01	0.95
True	4	5741 1351		0.01	0.95
True	5	5880 1521		0.01	0.95
True	6	5980 1446		0.01	0.96
True	7	5995 1347		0.01	0.96
True	8	5860 1500		0.01	0.95
True	9	5752 1428		0.01	0.95

Table 7: Cross validation (data set B)

Shuffled	Fold	Patterns	Sig	Contradict	Consistent
False	0	460 61		0.02	0.97
False	1	437 66		0.02	0.95
False	2	426 78		0.00	0.95
False	3	440 73		0.00	0.96
False	4	420 84		0.00	0.95
False	5	447 81		0.05	0.89
False	6	435 58		0.00	1.00
False	7	459 77		0.03	0.90
False	8	444 79		0.01	0.95
False	9	438 73		0.00	0.99
True	0	648 65		0.02	0.94
True	1	683 83		0.01	0.90
True	2	639 92		0.01	0.93
True	3	628 83		0.01	0.87
True	4	664 83		0.00	0.94
True	5	648 74		0.01	0.91
True	6	685 65		0.02	0.92
True	7	642 81		0.00	0.86
True	8	643 92		0.02	0.91
True	9	682 89		0.00	0.96

Intuitively, the sign of significant patterns in each training split should predict the sign of the Z-score in the test split. However, since each training split was much larger than each corresponding test split, it was more appropriate to compare, for each pattern, the sign of the significant Z-scores for patterns in the test split with the sign of the corresponding patterns in the training split.

For the non-shuffled data, we found that almost all of the patterns shown to have a significant increase in the testing split also had an increase in the training split ($Z \pm 1.96$). This demonstrates that, given a representative sample, we can select a small number of significant patterns based on a test of significance, and can reliably claim that the direction of difference in the population RR for a particular pattern.

Meaningfulness

Our next concern was whether the discovered patterns were meaningful. In other words, would we find similar patterns even if the order of the events were randomly shuffled? More generally, how will we know whether the discovered patterns are simply an artifact of the mining process?

We mined for patterns in a randomly shuffled transformation of data set A, and discovered a small number of patterns with significant RR, as shown in the second part of Table 6. We again observed good consistency between test and train splits. In addition, the patterns in the randomly shuffled data were almost as consistent as the patterns in the original data. Since the shuffled train and test splits came from the same sample, some hazard patterns would have been formed as a result of the frequency distribution of the event types in the sample.

We also mined for patterns in a randomly shuffled transformation of data set B, and discovered a larger number of patterns than we had discovered in the ordered data. This was an unexpected result but it does not support a conclusion of meaningless output, since the results are clearly very different from the results obtained using ordered data. One explanation for the increase in patterns in shuffled data is that some events that were strongly concentrated in one portion of the ordered data set became dispersed enough to participate in more pattern combinations.

Two additional considerations are the selection of support threshold and the selection of hazard constraints. We may be able to estimate a suitable minimum support threshold based on the characteristics of the data, possibly relying on a different support threshold for each event type. Further, the hazard constraints for each event type might be similarly tailored. These options will be explored in future work. For the purpose of the presenting

problem, patterns can be considered meaningful if the number and content of patterns found in a shuffled transformation of the data set are substantially different than in the original data.

Table 8: Arrests after parole (data set A shuffled)

Antecedents	$\overrightarrow{(0,3]}$	$\overrightarrow{(3,6]}$	$\overrightarrow{(6,12]}$	$\overrightarrow{(12,24]}$	$\overrightarrow{(24,96]}$	$\overrightarrow{(96,384]}$	Total
RR ($p \leq$)	0.98(<i>ns</i>)	0.98(<i>ns</i>)	0.89(<i>ns</i>)	0.82(<i>ns</i>)	1.13(0.05)	1.57(0.05)	
<i>Convicted</i> $\overrightarrow{(24,96]}$ <i>Paroled</i>	0.19	0.13	0.16	0.18	0.18	0.01	0.85
s 6657 / d 3219 / i 2165	604/604/557	433/433/402	523/523/489	575/575/543	577/575/551	40/40/40	2752/2750/1921

Table 9: New offense after parole contact (data set B shuffled)

Antecedents	$\overrightarrow{(0,7]}$	$\overrightarrow{(7,30]}$	$\overrightarrow{(30,180]}$	$\overrightarrow{(180,720]}$	Total
RR ($p \leq$)	3.180(0.001)	2.169(0.001)	1.203(<i>ns</i>)	1.931(0.001)	
<i>mx</i> $\overrightarrow{(0,7]}$ <i>mx</i>	0.060	0.026	0.044	0.141	0.271
s 1443 / d 1443 / i 697	87/43/43	38/32/29	63/47/45	203/124/122	391/246/182
RR ($p \leq$)	0.889(<i>ns</i>)	1.325(0.05)	2.503(0.001)	3.295(0.001)	
<i>mx</i> $\overrightarrow{(7,30]}$ <i>mx</i>	0.014	0.045	0.129	0.044	0.232
s 2947 / d 2947 / i 1028	40/40/40	132/112/101	381/211/196	130/62/62	683/425/258
RR ($p \leq$)	0.961(<i>ns</i>)	1.692(0.001)	1.404(0.001)	1.706(0.001)	
<i>mx</i> $\overrightarrow{(7,30]}$ <i>mx</i> <i>ph</i>	0.015	0.050	0.117	0.040	0.222
s 1997 / d 1570 / i 828	24/24/23	79/78/73	183/138/130	63/47/47	349/287/197
RR ($p \leq$)	0.729(<i>ns</i>)	1.330(<i>ns</i>)	1.960(0.001)	1.613(0.05)	
<i>mx</i> $\overrightarrow{(7,30]}$ <i>md</i>	0.012	0.048	0.148	0.043	0.251
s 1442 / d 1097 / i 668	13/13/13	53/53/52	162/132/124	47/34/34	275/232/177
RR ($p \leq$)	0.810(<i>ns</i>)	0.868(<i>ns</i>)	1.530(0.001)	1.060(<i>ns</i>)	
<i>mx</i> $\overrightarrow{(30,180]}$ <i>mx</i>	0.013	0.036	0.103	0.025	0.177
s 2595 / d 2595 / i 1127	34/34/34	94/94/91	266/193/181	65/45/45	459/366/245
RR ($p \leq$)	0.713(<i>ns</i>)	0.922(<i>ns</i>)	1.308(0.001)	0.847(<i>ns</i>)	
<i>mx</i> $\overrightarrow{(30,180]}$ <i>mx</i> <i>ph</i>	0.012	0.038	0.112	0.030	0.192
s 3035 / d 1603 / i 956	20/20/20	61/61/58	179/147/141	48/35/35	308/263/196
RR ($p \leq$)	0.952(<i>ns</i>)	1.061(<i>ns</i>)	1.949(0.001)	2.402(0.001)	
<i>mx</i> $\overrightarrow{(7,30]}$ <i>mx</i>	0.014	0.042	0.137	0.047	0.239
s 1518 / d 1396 / i 783	20/20/19	58/55/50	191/152/143	65/47/47	334/274/198
RR ($p \leq$)	1.199(<i>ns</i>)	1.184(<i>ns</i>)	2.044(0.001)	1.393(<i>ns</i>)	
<i>mx</i> $\overrightarrow{(7,30]}$ <i>mx</i> <i>ph</i>	0.018	0.043	0.157	0.040	0.257
s 956 / d 956 / i 541	17/17/17	41/40/38	150/105/102	38/27/27	246/189/138
RR ($p \leq$)	0.837(<i>ns</i>)	1.127(<i>ns</i>)	1.357(0.001)	1.025(<i>ns</i>)	
<i>mx</i> $\overrightarrow{(30,180]}$ <i>mx</i> <i>ph</i>	0.014	0.042	0.112	0.029	0.196
s 1578 / d 1578 / i 826	22/22/22	66/66/65	176/132/125	46/34/34	310/254/184

In addition to evaluating the directionality of the RR across validation folds, we also compared the pattern content for meaningfulness. We repeated the pattern mining and pattern selection process used in Table 4, using the same data, but shuffled. For data set A, rather than six significant patterns discovered, there were only two significant patterns discovered in the shuffled data set (see Table 8) allowing us to conclude that the patterns shown in Table 4 are indeed meaningful. We followed the same procedure for data set B and found a reduction in significant patterns from 22 in the ordered data (Table 5) to 17 in the shuffled data (Table 9). We also noted that the antecedent patterns in the shuffled data included several patterns with hazard constraints of (30,180] and only one pattern with a hazard constraints of (0,7], whereas in the ordered data, there were no patterns with a hazard constraint of (30,180] and several

patterns with a hazard constraint of (0,7]. This provides further evidence to support the conclusion that the discovered patterns are meaningful.

Predictiveness

Table 10: Two year arrest, based on releases 2-4 years earlier (data set A)

Year	prop Train	prop Test	Z Prop
61	0.42	0.96	6.20
62	0.69	0.98	8.24
63	0.87	0.94	4.23
64	0.90	0.90	0.40
65	0.94	0.89	-7.82
66	0.97	0.90	-12.16
67	0.94	0.92	-3.39
68	0.92	0.93	0.35
69	0.91	0.92	0.49
70	0.90	0.92	1.67
71	0.90	0.89	-0.96
72	0.90	0.85	-3.45
73	0.92	0.85	-3.93
74	0.92	0.86	-3.10
75	0.90	0.91	0.52
76	0.89	0.95	4.53
77	0.87	0.94	4.34
78	0.86	0.87	0.61
79	0.89	0.79	-2.88
80	0.89	0.71	-4.25

To evaluate the predictiveness of the discovered patterns, we first compared the proportion of patterns that lead to arrest in one time period with the proportion of pattern occurrences that lead to arrest in a subsequent time period. A number of challenges limited the design of such a test. First, to use past patterns to predict future recidivism within two years, it is necessary to apply a two year lag to the training data. For instance, recidivism data from those released January 1964 or earlier can be used to estimate two year recidivism for those released after January 1966 but not for those released sooner. Since the data set consists of a cohort group born approximately the same year, any age-related covariates complicate generalization from an earlier time period to a later time period.

We evaluated the entire time period for the antecedent pattern $\overrightarrow{Paroled(12,24]Paroled}$ to predict re-arrest within a two year time span. If there is no significant difference between the arrest risk over a preceding time period and a subsequent time period, the hazard pattern might be a useful predictor of future risk. We then compared the proportion with the proportion of re-arrests within two years going forward (testing period). To reduce variance,

and to summarize the results, these were grouped according to year, as shown in Table 10. Some arrests before 63/64 correspond to juvenile offenses. This corresponds to the most difficult time period for predicting arrest risk for these individuals. We anticipate that prediction accuracy would further improve with a data set comprised of individuals with varying ages.

We noted that past risk of arrest is significantly different than future risk of arrest in 13 of 20 years. We repeated the test using a variety of different training, testing and follow-up periods. In each case, the results were similar. For this particular hazard pattern, we observed that past risk of arrest is not a reliable indicator of future risk of arrest. This analysis was not repeated with data set B because only two years were covered.

Although there is only a moderate correspondence of recidivism risk between antecedent patterns in two different time periods, some of the discrepancy may simply be because the data set consists of a cohort group, exacerbating the effect of age-related covariates. Additionally, hazard patterns may be more effective when combined with other risk assessment indicators, such as prior drug and alcohol abuse.

Parsimony

Based on the above pattern tables, we see that the use of the RR Z-score dramatically reduces the number of patterns of interest, highlighting a small proportion of significant patterns for the analyst to consider. This pattern selection approach favors short patterns over longer patterns. This is because each longer pattern is only checked for significance relative to a shorter baseline pattern. There are likely many more patterns that might be significant relative to a baseline of random chance, but these would likely overwhelm a human analyst.

In Table 4 we summarized all patterns related to recidivism after parole release with a minimum support threshold of 500, and logically arranged them together with indicators of effect direction and significance in bold. We note that there are 24 patterns that have a significant impact on risk of re-arrest.

5.2 Conclusion

Contributions

In this paper we demonstrated that hazard patterns can be used to identify individuals with increased parole violation risk. Although we did not find a direct link between past arrest risk and future arrest risk, we did find a significant relationship between past RR and future RR between a group exhibiting a hazard pattern and a group that exhibited a hazard pattern with the first antecedent removed.

We also tested the generalizability of the discovered hazard patterns through ten-fold cross validation. We further also demonstrated a simple test for meaningfulness of hazard patterns. If a similar amount of patterns is discovered when the order of the underlying data is shuffled, then the discovered patterns are meaningless. The need for a meaningfulness test is particularly relevant, given that meaningless patterns can pass a cross-validation test.

Limitations

A common theme in the investigation of criminal careers is the counterfactual history. Since we cannot randomly assign individuals to parole release, we must rely on other means of determining what would have happened to a particular individual if they had not been released on parole. We do not address this question in this work.

Another closely related limitation arises due to interaction with the decision maker. If a decision maker relies on the indicated risk level to determine parole release eligibility, then the accuracy of the system will be negatively impacted. For instance, the system may show a high risk to re-offend for some cases. If the decision maker does not grant parole release based on this recommendation, the risk to re-offend has been altered.

Future Work

With the introduction of hazard patterns comes a wide range of opportunities for further work. Application domains with time-to-event data are the most likely to benefit. Examples include health care histories, business process analytics, equipment failure events, and insurance claim histories. More immediately, we plan to develop a decision support tool to facilitate discovery of patterns in event histories, and a graphical data exploration tool to facilitate interactive navigation of the discovered patterns. We also believe that heuristic can be applied to automate the selection of a support threshold that yields the most meaningful patterns. Further, the selection of hazard constraints is a compromise between theoretical

properties of inter-arrival time distribution and familiar calendar based time frames. More investigation is needed to identify suitable collections of constraint mechanisms for other data sets.

REFERENCES

- Abraham, T. (2006a). Event sequence mining to develop profiles for computer forensic investigation purposes. *Proceedings of the 2006 Australasian workshops on Grid computing and e-research-Volume 54*: Australian Computer Society, Inc.
- Abraham, T. (2006b). Event sequence mining to develop profiles for computer forensic investigation purposes *Proceedings of the 2006 Australasian workshops on Grid computing and e-research-Volume 54* (pp. 145-153): Australian Computer Society, Inc.
- Achar, A., Laxman, S., & Sastry, P. S. (2011). A unified view of the apriori-based algorithms for frequent episode discovery. *Knowledge and Information Systems*. doi: 10.1007/s10115-011-0408-2
- Agrawal, R., & Srikant, R. (1995a). Mining sequential patterns. *Proceedings of the Eleventh International Conference on Data Engineering*: IEEE Computer Society.
- Agrawal, R., & Srikant, R. (1995b). Mining sequential patterns *Proceedings of the Eleventh International Conference on Data Engineering* (pp. 3-14): IEEE Computer Society.
- Bathoorn, R., Welten, M., & Richardson, M. (2010a). Frequent episode mining to support pattern analysis in developmental biology *Pattern Recognition in bioinformatics* (pp. 253-263).
- Bathoorn, R., Welten, M., & Richardson, M. (2010b). Frequent episode mining to support pattern analysis in developmental biology. *Pattern Recognition in bioinformatics*.
- Bersani, B. E., Nieuwbeerta, P., & Laub, J. H. (2009). Predicting Trajectories of Offending over the Life Course: Findings from a Dutch Conviction Cohort. *Journal of Research in Crime and Delinquency*, 46, 468-494. doi: 10.1177/0022427809341939
- Bewick, V., Cheek, L., & Ball, J. (2004). Statistics review 11: assessing risk. *Critical care (London, England)*, 8, 287-291. doi: 10.1186/cc2908
- Bhati, A. S., & Piquero, A. R. (2007). Estimating the Impact of Incarceration on Subsequent Offending Trajectories: Deterrent, Criminogenic, or Null Effect? *The Journal of Criminal Law and Criminology*, 98, 207-253.
- Blanchard, J., Guillet, F., & Gras, R. e. (2008). Assessing the interestingness of temporal rules with Sequential Implication Intensity. *Statistical Implicative Analysis*, 55-71.
- Cate, M. L., Hoshini, M., Seale, L., Grealish, B., Fitzgerald, T., Grassel, K., . . . Reyes, M. (2012). *2012 CDCR Outcome Evaluation Report Office of Research*.
- Duwe, G. (2013). The Development, Validity, and Reliability of the Minnesota Screening Tool Assessing Recidivism Risk (MnSTARR). *Criminal Justice Policy Review*. doi: 10.1177/0887403413478821
- Eichinger, F., Nauck, D. D., & Klawonn, F. (2006a). Sequence mining for customer behaviour predictions in telecommunications *Proceedings of the Workshop on Practical Data Mining at ECML/PKDD* (pp. 3-10).
- Eichinger, F., Nauck, D. D., & Klawonn, F. (2006b). Sequence mining for customer behaviour predictions in telecommunications. *Proceedings of the Workshop on Practical Data Mining at ECML/PKDD*.
- Ferreira, P., & Azevedo, P. (2005). Protein sequence classification through relevant sequence mining and bayes classifiers. *Progress in Artificial Intelligence*, 236-247.

- Fujikawa, J., Kida, T., & Katoh, T. (2011a). Extracting refrained phrases from music signals using a frequent episode pattern mining algorithm *IEEE International Conference on Granular Computing* (pp. 199-202).
- Fujikawa, J., Kida, T., & Katoh, T. (2011b). Extracting refrained phrases from music signals using a frequent episode pattern mining algorithm. *IEEE International Conference on Granular Computing*.
- Mining frequent patterns in data streams at multiple time granularities 191-212 (AAAI Press 2004).
- Gouda, K., Hassaan, M., & Zaki, M. J. (2007). PRISM: A prime-encoding approach for frequent sequence mining *IEEE International Conference on Data Mining* (pp. 487-492). Omaha, Nebraska, USA: IEEE Computer Society.
- Han, J., & Dong, G. (1999). Efficient mining of partial periodic patterns in time series database. *Data Engineering, 1999. Proceedings.*, 106-115. doi: 10.1109/ICDE.1999.754913
- Haviland, A., Nagin, D. S., & Rosenbaum, P. R. (2007). Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychological methods*, 12, 247-267. doi: 10.1037/1082-989X.12.3.247
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *Mis Quarterly*, 28, 75-105.
- Janzen, C. A., Deokar, A. V., & El-Gayar, O. F. (2013a). *Discovering Predictive Event Sequences in Criminal Careers*. Paper presented at the Annual Symposium on Information Assurance (ASIA), Albany, NY.
- Janzen, C. A., Deokar, A. V., & El-Gayar, O. F. (2013b). Non-parametric discovery of event sequence patterns in criminal behavior. *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*. Maui, HI: IEEE Computer Society.
- Keogh, E., & Lin, J. (2004). Clustering of time-series subsequences is meaningless: implications for previous and future research. *Knowledge and Information Systems*, 8, 154-177. doi: 10.1007/s10115-004-0172-7
- Leleu, M., Rigotti, C., Boulicaut, J. F., & Euvrard, G. (2003). Constraint-based mining of sequential patterns over datasets with consecutive repetitions. *Knowledge Discovery in Databases: PKDD 2003*, 303-314. doi: 10.1.1.134.3862
- Magnusson, M. S. (2000). Discovering hidden time patterns in behavior: T-patterns and their detection. *Behavior research methods, instruments, & computers : a journal of the Psychonomic Society, Inc*, 32(1), 93-110.
- Mannila, H., & Toivonen, H. (1995a). Discovering frequent episodes in sequences extended abstract *1st Conference on Knowledge*.
- Mannila, H., & Toivonen, H. (1995b). Discovering frequent episodes in sequences extended abstract. *1st Conference on Knowledge*.
- Martin, B., & Dine, S. V. (2008). Examining the Impact of Ohio's Progressive Sanction Grid, Final Report.
- Masseglia, F., Poncelet, P., & Teisseire, M. (2009). Efficient mining of sequential patterns with time constraints: Reducing the combinations. *Expert Systems with Applications*, 36(2), 2677-2690.

- Min, Y., Wong, S. C. P., & Coid, J. (2010). The Efficacy of Violence Prediction: A Meta-Analytic Comparison of Nine Risk Assessment Tools. *Psychological bulletin*, 136, 740-767.
- Nagin, D. S., & Land, K. C. (1993). Age, criminal careers, and population heterogeneity: specification and estimation of a nonparametric, mixed Poisson model. *Criminology*, 31, 327-362.
- Newman, W. J., & Scott, C. L. (2012). Brown v. Plata: prison overcrowding in California. *The journal of the American Academy of Psychiatry and the Law*, 40, 547-552.
- Peppers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, 24(3), 45-77.
- Pei, J., Han, J., & Wang, W. (2002). Mining sequential patterns with constraints in large databases *CIKM 02 Proceedings of the eleventh international conference on Information and knowledge management* (pp. 18-25): ACM Press.
- Qin, M., & Hwang, K. (2004a). Frequent Episode Rules for Intrusive Anomaly Detection with Internet Datamining. *USENIX Security Symposium*.
- Qin, M., & Hwang, K. (2004b). Frequent Episode Rules for Intrusive Anomaly Detection with Internet Datamining *USENIX Security Symposium* (pp. 1-15).
- Spiliopoulou, M. (1999). Managing interesting rules in sequence mining. *Principles of Data Mining and Knowledge Discovery*, 554-560.
- Srivastava, A., Sural, S., & Majumdar, A. K. (2006a). Database Intrusion Detection using Weighted Sequence Mining. *Journal of Computers*, 1(4), 8-17.
- Srivastava, A., Sural, S., & Majumdar, A. K. (2006b). Database Intrusion Detection using Weighted Sequence Mining. *Journal of Computers*, 1, 8-17. doi: 10.4304/jcp.1.4.8-17
- Turner, S., Braithwaite, H., Kearney, L., Murphy, A., & Haerle, D. (2012). Evaluation of the California Parole Violation Decision-Making Instrument (PVDMI). *Journal of Crime and Justice*, 35, 269-295. doi: 10.1080/0735648X.2012.683636
- Turner, S., Ph, D., Hess, J., & Jannetta, J. (2009). Development of the California Static Risk Assessment Instrument (CSRA) ... of California, Irvine,
- Wang, J., & Han, J. (2004). BIDE: Efficient Mining of Frequent Closed Sequences *International Conference on Data Engineering*.
- Criminal Careers, Criminal Violence, and Substance Abuse in California, 1963-1983, Inter-university Consortium for Political and Social Research (ICPSR) [distributor] (2006).
- Yan, X., Han, J., & Afshar, R. (2003). CloSpan: Mining Closed Sequential Patterns in Large Datasets. *IN SDM*, 166--177. doi: 10.1.1.12.3538
- Zaki, M. J. (2000). Sequence mining in categorical domains: incorporating constraints *Proceedings of the ninth international conference on Information and knowledge management* (pp. 422-429): ACM.
- Zaki, M. J. (2001). SPADE: An efficient algorithm for mining frequent sequences. *Machine Learning*, 42(1), 31-60.
- Zaki, M. J. (2010). VOGUE: A variable order hidden Markov model with duration based on frequent sequence mining. *ACM Transactions on Knowledge Discovery from Data*, 4(1(5)).
- Zaki, M. J., Lesh, N., & Ogihara, M. (1998a). PLANMINE: Sequence Mining for Plan Failures. *4th Intl. Conf. Knowledge Discovery and Data Mining*.

- Zaki, M. J., Lesh, N., & Ogihara, M. (1998b). PLANMINE: Sequence Mining for Plan Failures *4th Intl. Conf. Knowledge Discovery and Data Mining* (pp. 962-969).
- Zhang, Z., Zhou, D.-D., Yang, H.-J., & Zhong, S.-C. (2010a). A service composition approach based on sequence mining for migrating e-learning legacy system to SOA. *International Journal of Automation and Computing*, 7(4), 584-595.
- Zhang, Z., Zhou, D.-D., Yang, H.-J., & Zhong, S.-C. (2010b). A service composition approach based on sequence mining for migrating e-learning legacy system to SOA. *International Journal of Automation and Computing*, 7, 584-595. doi: 10.1007/s11633-010-0544-2