2015

Indirect association rule mining for crime data analysis

Riley Englin
Eastern Washington University

INDIRECT ASSOCIATION RULE MINING FOR CRIME DATA ANALYSIS

A Thesis

Presented To

Eastern Washington University

Cheney, Washington

In Partial Fulfilment of the Requirements

for the Degree

Master of Science in Computer Science

By Riley Englin

Fall 2015

THESIS OF RILEY ENGLIN APPROVED BY

_____DATE_____

DAN LI, GRADUATE STUDY COMMITTEE

_____DATE_____

STU STEINER, GRADUATE STUDY COMMITTEE

MASTER'S THESIS

In presenting this thesis in partial fulfillment of the requirements for a master's degree at Eastern Washington University, I agree that the JFK Library shall make copies freely available for inspection. I further agree that copying of this project in whole or in part is allowable only for scholarly purposes. It is understood, however, that any copying or publication of this thesis for commercial purposes, or for financial gain, shall not be allowed without my written permission.

Signature_____

Date_____

Abstract

Crime data analysis is difficult to undertake. There are continuous efforts to analyze crime and determine ways to combat crime but that task is a complex one. Additionally, the nature of a domestic violence crime is hard to detect and even more difficult to predict. Recently police have taken steps to better classify domestic violence cases. The problem is that there is nominal research into this category of crime, possibly due to its sensitive nature or lack of data available for analysis, and therefore there is little known about these crimes and how they relate to others. The objectives of this thesis are 1) develop an indirect association rule mining algorithm from a large, publicly available data set with a focus on crimes of the domestic violence nature 2) extend the indirect association rule mining algorithm for generating indirect association rules and determine its impact.

Acknowledgements

I would like to thank a couple people for their help with this thesis. First I would like to thank Dr. Li for her guidance, patience, and understanding throughout this process. Without her expertise, this thesis would have never been possible. Next, I would like to thank Stu Steiner for challenging me and pushing me to succeed. I will be forever grateful for his endless support and encouragement throughout graduate school.

Table of Contents

List of Figures

1 INTRODUCTION

Crime is a serious problem throughout the world. In recent years, there has been an evolving effort to use data in order to combat it. There is an abundance of data pertaining to each crime that is collected and stored. The crime related data has been gathered for many years, so there is a massive amount of it in existence. However, without the necessary knowledge and tools to analyze this data, it is meaningless. Currently, one of the most frequently used methods to identify crime patterns involves reviewing crime reports each day and comparing those reports to past reports in order to determine if any patterns can be detected [5]. In addition to being highly prone to error, this method is extremely time consuming and inefficient. For this very reason, a technique called data mining is very useful, with proper training and research. Data mining is the process of discovering hidden patterns and relationships within large amounts of data [13]. This technique is beneficial when used with crime data because there is no need to know what is being searched for in order to use it. Instead, the process of analyzing and exploring the data with various data mining techniques gives way to vast amounts of important, useful and usable information. Data mining can also allow for pattern discovery and analysis in an automated manner that has the potential to "enhance and accelerate the efforts of local law enforcement" [2].

Crime data is very difficult to work with when using data mining for a couple reasons. First, crime data that has been collected over the years was never intended to be examined, so it was not collected in a form that is "friendly" to be used. This means that it first needs to be processed into a form that can be used, and often times this task is more extensive and difficult than the actual process of analyzing the data. Additionally,

the nature of crime data poses a large challenge in and of itself. It presents issues that are delicate to deal with but need to be addressed, such as data storage, warehousing, and privacy [2]. These aspects can make accessing crime data difficult because sensitive information, such as victim name, address, etc., are not available to the public but are often times the focus of research projects. For this very reason, the data chosen for this research is from a data portal that is accessible by anyone and provides none of this sensitive information. The data being examined for this study is from the City of Chicago data portal and provides basic data about reported crimes [6]. At the time of data collection, there were over four million records in the data set with each record containing twenty-two attributes.

In the dataset there are two Boolean attributes of "Arrest" and "Domestic" that state whether or not the crimes committed were domestic in nature or resulted in an arrest. The research will focus on generating indirect association rules when the crime either resulted in an arrest or was domestic, or both. Indirect association rule mining is one technique that is used for discovering value from infrequent patterns by indirectly connecting two rarely co-occurring items through some deemed mediator [15]. By doing this effectively there is the possibility to identify interesting item sets from a database that may appear to be "uninteresting" by another algorithm. The goal of the research is to show that significant relationships can be mined from public, unclean data by employing and extending indirect association rule mining on the attributes available.

2 BACKGROUND

There is a large amount of publicly available crime data, but there is no benefit to having this data without the ability to analyze it. By mining the data, useful information can be found to help combat crime and aid police personnel in discovering patterns for future use. The data set that was used for this work came from the City of Chicago data portal. The data is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system [6]. Any personal data and specific addresses are withheld from the data set in order to protect the privacy of the suspects and victims. The data set was acquired on October 1, 2014. There were 4,556,343 records in the data set at that time, and the date range spanned from January 1, 2001 to October 1, 2014. This set is updated every seven days with the most up-to-date cases and there is no guarantee that the data in the set is clean or without error.

The data set was downloaded in a comma-separated values (.csv) format. Once this file was downloaded, the data cleansing process was begun. Each record contained twenty-two possible categorical, quantitative, and Boolean attributes. The attributes with their descriptions reside in table 2-1.

1	ID	Unique identifier for each record
2	Case Number	Case number assigned to each case
3	Date	Date that the crime took place
4	Block	Approximate address of occurrence
5	IUCR	Illinois Uniform Crime Reporting codes
6	Primary Type	Type of crime committed
7	Description	Further description of the type of crime committed
8	Location Description	Describes type of place crime took place
9	Arrest	Yes or no if culprit was arrested
10	Domestic	Yes or no if crime was a domestic crime
11	Beat	Code corresponding to the territory and time of police patrol
12	District	22 Districts
13	Ward	50 Wards
14	Community Area	77 Community areas divided by the Social Science Research Committee at the University of Chicago
15	FBI Code	Code assigned to case based off the primary type of crime
16	X Coordinate	X-Coordinate of crime
17	Y Coordinate	Y-Coordinate of crime
18	Year	Year the crime took place
19	Updated On	Date that the case was last updated
20	Latitude	Latitude of crime
21	Longitude	Longitude of crime
22	Location	(Latitude, Longitude)

Table 2-1: Initial Data Set Attributes

3 RELATED WORK

The crime community is rich with data and over recent years this data has begun to be mined for useful information in a large array of ways. By using data mining techniques and analyzing this crime data, there is the ability to discover crime patterns, identify when and where crimes may take place, and determine how to efficiently employ police personnel to be the most productive at combating crime while making use of police budgets. In order to analyze crime data, the proper technique must be used based on desired outcome and the data set that is being used. The limitation of using association rule mining is that when needing to generate rules for data that is categorical, such as types of crimes, or quantitative, such as number of crimes, some additional data preprocessing is needed to establish some kind of numeric identifier for the categorical value in order to efficiently develop association rules or item sets.

Association rule mining is the process of finding relationships among different attributes in a data set. It was originally introduced as a way to discover frequent items that were bought together in a supermarket transaction. This algorithm generates association rules in the form of X implies Y, or X → Y, from a frequent item set of {X,Y} [13]. One of the most popular algorithms for generating these frequent item sets is called the Apriori algorithm. Apriori uses an approach that makes use of a property that states that any subset of a frequent item set must also be frequent. To do this, a set of candidate items of length n + 1 are generated from a set of items of length n [16]. Then, each of these candidate sets is checked to see if they meet the minimum support threshold and can be considered frequent. This process is very inefficient, especially on large amounts of data. For this very reason many improvements have been made, resulting in

many algorithms that have emerged from Apriori, such as the FP-Growth algorithm which uses a structure called an FP-tree to discover frequent item sets [12] and the Partition algorithm that uses intersections to determine support values of items rather than the Apriori method of counting [9].

Association rule mining algorithms use some parameters that are specified by the user in order to generate rules or item sets that the user would deem as useful or important, usually based on the kind of data that is being analyzed. These are generally used in some form or another across all rule mining algorithms, so it is important that they be introduced. The common parameters used are support, confidence, and lift of the item set in question. The *support* of an item set is the number of times the item set appears throughout the transaction database, or dataset.

Transaction ID	Items
1	{A, D}
2	{B, C, E}
3	{A, B, C}
4	{D}
5	{A, B, C, E}

Table 3-1: Example Database

For example, looking at Table 3-1, item set {B, C} has a support of 3 because it appears in transactions 2, 3, and 5. This value can be represented as a simple number, such as 3, a decimal value, such as 0.6, or a percentage, in this case 60%. The support value would then be used in the algorithm to determine which item sets would be considered frequent by the user because only the item sets that have a support higher than a pre-defined minimum threshold value would be selected. Similarly, the support of a rule X → Y is defined as the number of transactions that contain X U Y. For example, the support of item B is 60% because it appears in 3 of the 5 transactions, and the support of a rule, say B → E, is 40%. The *confidence* of a rule X → Y is the number of transactions that contain X U Y divided by the number of transactions that contain X. Again, looking at

the rule B → E, the confidence of the rule would be 2/3, or about 0.67. Similar to support, the user will specify some minimum confidence value that they are looking for a rule to have in order for it to be considered important enough for the final rule set. The *lift* of a rule is used to determine if the confidence value calculated is one that should be considered. Lift is a calculation that takes into account the overall transaction database, while the confidence of a rule only looks at the item sets that are a part of the given rule, which can result in a "false positive". The lift calculation is as follows:

$$Lift\ (X,Y) = \frac{\frac{Sup(X\ U\ Y)}{N}}{\frac{Sup(X)}{N} * \frac{Sup(Y)}{N}}$$

where N = the number of transactions in the database

If the resulting lift value is equal to 1 then X and Y are independent of one another. If lift is greater than 1 then X and Y are positively correlated. If lift is less than 1 then X and Y are negatively correlated. Generally, the minimum lift value set by a user is 1 in order to remove any of those negatively correlated rules that pass the confidence threshold. Looking back at the rule B → E, the lift would be as follows:

$$Lift\ (B,E) = \frac{\frac{Sup(B\ U\ E)}{5}}{\frac{Sup(B)}{5} * \frac{Sup(E)}{5}} = \frac{\frac{2}{5}}{\frac{3}{5} * \frac{2}{5}} = 1.667$$

This means that the items B and E are positively correlated within the dataset. *Cosine* is a symmetric measure that shows how closely related two items, or rules, may be [11]. The closer the cosine value of a rule X → Y is to 1, the more transactions containing X also contain Y. Cosine also has a null-invariant property, meaning that the transactions in the

dataset that do not contain X or Y have no influence on the result of cosine(X → Y). Cosine is defined as:

$$Cosine\ (X,Y) = \frac{P(X,Y)}{\sqrt{P(X)*P(Y)}}$$

Again, revisiting the rule B → E, the cosine value is as follows:

$$Cosine\ (B,E) = \frac{P(B,E)}{\sqrt{P(B)*P(E)}} = \frac{0.4}{\sqrt{0.6*0.6}} = 0.667$$

This shows that B and E are more closely related than unrelated and could be of interest within the dataset. *Interest* is a measure that can be used to quantify the strength between items [15]. The interest between items X and Y is defined as:

$$Interest\ (X,Y) = \frac{P(X,Y)}{P(X)*P(Y)}$$

$$Interest\ (B,E) = \frac{P(B,E)}{P(B)*P(E)} = \frac{0.4}{0.6*0.6} = 1.111$$

There have been many data mining techniques employed on crime data, as it is a large area of interest and there has been vast amounts of data collected [2,3,5,6]. Quantitative association rule mining is one technique that has been investigated. It handles categorical and quantitative values by partitioning the values of the attributes and then combining adjacent partitions when deemed necessary [14]. Quantitative rule mining uses a mapping of categorical and quantitative attributes to a set of consecutive integers that can then be used to develop rules [8]. However, this technique has the potential to result in information loss and high execution time, especially when

performed on large data sets. These two issues are trade-offs - as the number of intervals is increased less information is lost but execution time increases, and if the number of intervals is reduced, then the data integrity is lost but execution time decreases.

Another technique that has been used is called fuzzy association rule mining. To generate fuzzy association rules, the Apriori algorithm was extended to a Fuzzy Apriori algorithm that is more easily understood by humans [2]. For each item, the algorithm decides if it is a member or not of each set, and this allows for a smooth transition for each element between membership and non-membership of every set generated. The process involves defining "fuzzification" membership functions for each variable that then produces the membership values for each of the data items. Next, the fuzzy Apriori algorithm is employed on the data set, which includes initial pruning of the generated rules based on some constraint. This method was used on an open-source Communities and Crime dataset and produced promising results [2]; however, exploring this algorithm requires a subject matter expert to determine the "fuzzification" membership functions, which is not available for this work.

Finally, there is an algorithm for generating indirect association rules that ultimately stemmed from Apriori [15]. It's based off the idea that there may be insight to be gained from the item sets that most algorithms would deem uninteresting or would consider to be negatively associated, and therefore would disregard in the result set [1]. This concept is best described with an example. Suppose there are two items in a data set, X and Y, which rarely occur in the same transaction. The item set {X, Y} would not pass the minimum support threshold designated for most algorithms, such as Apriori. However, X and Y are both highly dependent on another item set in the dataset, Z. As a

result, the item set {X, Y} is considered to be indirectly associated through Z, which would now be called the mediator of {X, Y}.

There has been additional work using indirect association rules for web recommendations [10], text mining, and stock market data mining [15]. However, there has been no found work done with indirect association rule mining incorporating the lift, cosine, and interest thresholds explained earlier in this section. The goal of this research is to introduce these additional parameters and examine the impact that it has on the resulting association rules when using crime data.

4 METHODS

4.1 Data Cleansing

In order to perform data cleansing and analysis on this data set, the set needed to be cut down to a subset of its original size. Analysis of the entire data set with the indirect association rule mining algorithm would take an extensive amount of time to complete. By taking a subset of the original set, the time for analysis is cut down, but the goal is to preserve the structure of the data set in conjunction with a proper representation of the crimes that were recorded within a given time span. A random sampling of the data set was considered but ultimately not used because the goal was to look at the crimes taking place throughout the span of an entire year, and a random sampling would have affected this analysis goal because there would not be a proper representation of the crimes that took place during each month throughout an entire year. Instead, the subset was produced by looking at a specific time frame within the set in order to preserve a proper representation of crimes that occurred in a year. After analyzing the crimes recorded within various date time frames, it was decided to look at crimes that were recorded between October 1, 2011 and October 1, 2014, because the data was obtained on October 1, 2014, thus giving a set of 932,436 crimes over a three year period in the city of Chicago with a proper representation of the dispersion crimes that took place throughout those years.

After this reduction, the data cleansing process was begun. The dataset did contain missing information and have some anomalies that came along with the file being in .csv format that needed to be addressed before anything further could be done with it. Firstly,

some of the attribute descriptions contained commas and were consequently split when being parsed. For instance, the "Location Description" attribute contained a few descriptions in which there were more than one location listed, such as "Boat, Watercraft" or "Hotel, Motel." Additionally, under the "Description" attribute that expands of the "Primary Type" of crime, there were descriptions like "Theft by Lessee, Motor Veh" and "Truck, Bus, Motorhome." All of these descriptions were split into separate columns instead of being kept within their single attribute column when the file was saved in csv format. This meant that additional consideration and parsing techniques needed to be employed in order to keep these descriptions and locations all as one. When one of these descriptions or locations occurred in the data set, the commas were replaced with "/" and kept together as the whole field for the record attribute when placed into the final file to be used for analyzing. For example, the final result would look something along the lines of a "Location Description" as "Boat/Watercraft" or a "Primary Type" of "Theft" and "Description" of "Theft by Lessee/Motor Veh."

For records that contained missing attributes, there were a couple different methods used for filling in those records depending on the attributes. Some records contained empty "Location" fields. When these records were encountered, the value of "NONE" was entered to make analysis simpler later on. Primarily, records that contained missing "Location" values were crimes such as "Deceptive Practice" with a "Description" attribute of "Financial Identity Theft Over $300" or "Theft" with a "Description" attribute of "$500 and Under." Records that contained missing "Latitude" and "Longitude" values were assigned "0" to keep the fields from being null. The same practice was initially applied to records with missing "District," "Ward," and

"Community Area" attributes. However, filling in the missing "District," "Ward," and "Community Area" attributes was taken one step further when it was discovered that all records in the data set contained a "Beat" attribute. A police beat refers to a location patrolled and a given time that the specific location is patrolled by the specified police officer. When looking at the records that were recorded under a single beat, it was observed that the "District," "Ward," and "Community Area" codes were all very similar. For example, a given beat may have two differing district codes, two differing ward codes, and three differing community area codes for a large number of records. Given this, it was decided to take an approach to fill in these values with the discovered information. First, the data set was scanned, and for each record that did not have missing attribute values, the beat, district, ward, and community area codes were stored. For each beat, a count was kept for how many times each differing district, ward, and community area code appeared. Once the entire data set was scanned, the maximum of each of these values for the individual beat was stored. Then, the records with missing district, ward, and community area codes were filled in, according to the beat of the record, with the code that appeared the most within the rest of the data set.

Finally, the data set initially provided a "Date and Time" attribute in the form of "10/1/2011 10:32 AM". This attribute was very useful, but it was most useful when the individual parts of the attribute were used separately because the time, date, and AM/PM part of the attribute could each be considered different items in any association rules being built. Therefore, the attribute was split into three different attributes in the final data set so that it was easier for each individual attribute to be used for analysis. The result was three different attributes of "Date", "Time", and "AM/PM" for each record.

For the final data set, not all attributes were kept because not all attributes were going to be useful for further analysis in this research. Any attributes that were considered duplicates were removed. For example, “Year” was removed because the actual date of the crime was already provided and “Location” was removed because “Latitude” and “Longitude” were provided individually. Also, attributes specifying case identifiers, such as case number, were removed due to the lack of significance in association rule mining algorithms. Initially, “Latitude” and “Longitude” were selected to be used, but were eventually discarded due to the fact that there were already 4 different location attributes provided and the difficulty involved in determining a proper grouping and then mapping of the values. Also, after initial runs of the algorithm over the dataset, it was decided that the “AM/PM” variable could not be used. This was because it was a Boolean attribute, meaning that it would appear an overwhelming amount of time in the rules being generated, taking away from the focus of the “Domestic” and “Arrest” attributes. The reason for this is because the indirect association rule algorithm looks for the support of an attribute and by having a Boolean attribute, the value for each is, most likely, going to pass that support threshold to be included in the final rule that is formed. This aspect will be discussed further in the next section. Even with the “AM/PM” attribute removed, that metric could still be determined in the rule analysis stage. Table 4-1 shows the final variables chosen for analysis.

1	Date	Date that the crime took place
2	Time	Time that the crime took place
3	Primary Type	Type of crime committed
4	Description	Further description of the type of crime committed
5	Location Description	Describes type of place the crime took place
6	Arrest	Yes or no if suspect was arrested
7	Domestic	Yes or no if crime was a domestic crime
8	Beat	Code corresponding to the territory and time of police patrol
9	District	22 Districts
10	Ward	50 Wards
11	Community Area	77 Community areas divided by the Social Science Research Committee at the University of Chicago

Table 4-1: Final Attributes

4.2 Integration

Once the final dataset was cleansed and produced, it was loaded into a SQL database in order to map the attribute values in the data set to unique identifiers for the final data file. The reason for this was because the indirect association rule mining algorithm used requires that each item in the data set be represented as an integer, and therefore, each of the values in the data set need to be linked to a unique integer value. The dataset was loaded into a single table and then individual tables were created for each attribute. Next, a unique identifier was assigned to each possible value in the entire dataset. Once the mapping was complete, a final table was made that joined all the records back together with the identifiers for each attribute. This table was then output to a data file to run through the algorithm. The last step in this process was to convert that file into the format that was desired by the algorithm and the data was ready to be analyzed.

4.3 Data Set Generation

The final data set that spanned over three years of crime data was split into 2 individual data sets for the purpose of rule generation and prediction. The initial rules were generated from the data set that spanned from October 2011 to September 2013, or the training set, containing 655,309 records. Next, rules were generated from the smaller data set, or test set, spanning from October 2013 to September 2014 containing 276,209 records. This allowed for the ability to determine if the rules generated from the data would actually be applicable to future data and be used for rule prediction in the future. These sets of data were ultimately the same – both coming from the same larger set of data – but containing a different number of records and no overlapping records in order to compare the results of the two during analysis and determine if they would produce the same sets of rules, therefore determining if these rules would hold throughout the entire dataset and future data to be stored.

4.4 Rule Generation

To complete this work, an open-source data mining library was chosen. The library chosen is called SPMF [7]. SPMF is written in Java and offers implementations of 93 data mining algorithms distributed under the GPL v3 license. This library worked well because all of the code is well documented and it contains a program that allows users to interact with a user interface very easily. The algorithm produces the associations in the form of {X, Y} | M, where X and Y are single items in the dataset and M is an item set that is the mediator between X and Y. In order to determine these indirect association

rules there are three parameters that must be provided for the indirect association rule algorithm being used from the SPMF library. These three parameters are defined below:

1. minsup - the minimum support threshold between each item and the mediator
2. ts - the minimum support for the item pair
3. minconf - the minimum confidence required for the indirect associations

The rules that are generated must satisfy these support and confidence thresholds specified by the user. Using this algorithm and the dataset ranging from October 2011 through September 2013, indirect association rules were generated for all differing values of minsup, ts, and minconf. There were varying values for each of these parameters used to generate sets of these indirect association rules. To demonstrate how these values are computed and used, it is easiest to use an example based on table 4-2.

Transaction ID	Items
1	{A, D, E}
2	{B, C, D}
3	{A, B, D, E}
4	{E}
5	{A, B, D, E}

Table 4-2: Example Database

If the user were to specify a minsup of 60%, ts of 50%, and minconf of 10%, 3 indirect association rules would be generated. One of those rules would be {A, E | {D}} because $\frac{Sup\ (A\ U\ D)}{5} = 0.6$ and $\frac{Sup\ (E\ U\ D)}{5} = 0.6$ satisfying the minsup constraint, $\frac{Sup\ (A\ U\ E)}{5} = 0.6$ satisfying the ts constraint, and the confidence of A in terms of D is $\frac{Sup\ (A\ U\ D)}{Sup\ (A)} = 1$ and the confidence of E in terms of D is $\frac{Sup\ (E\ U\ D)}{Sup\ (E)} = 0.75$, both satisfying the minconf constraint. Once these constraint values are set, the algorithm works to generate rules based off of them. First, the algorithm counts the number of times each item occurs in the data set to determine if the item is considered frequent based off of the user's constraints. Next, it uses an Apriori-style generation of frequent item sets, starting

from individual frequent items and working its way up to generate larger sets from those frequent sets until no more candidate sets can be generated. Then, for each item set of size k, for k > 2, the algorithm compares that item set against all other item sets of size k looking for item sets in which the two sets only have one differing item. It is important to note that the algorithm only looks for one differing item because this means that the rules that are generated will contain two single items that are indirectly associated through the found mediator. Next, for all item sets found, the algorithm then removes those items, for example, A and B, which are different. Finally, it checks to see if the remaining item set could be a mediator for A and B by determining if the support of {A, B} is higher than the ts threshold and if the confidence of A with respect to the mediator and B with respect to the mediator pass the minconf threshold. If the items and mediator pass these determined threshold, then the indirect association rule is established for the user.

4.5 Algorithm Extension

After the initial set of indirect association rules were generated, the algorithm was extended to take into account the lift, cosine, and interest values for the association rule. These would also be user defined metrics like the other three already provided. Once implemented, the same tests were run.

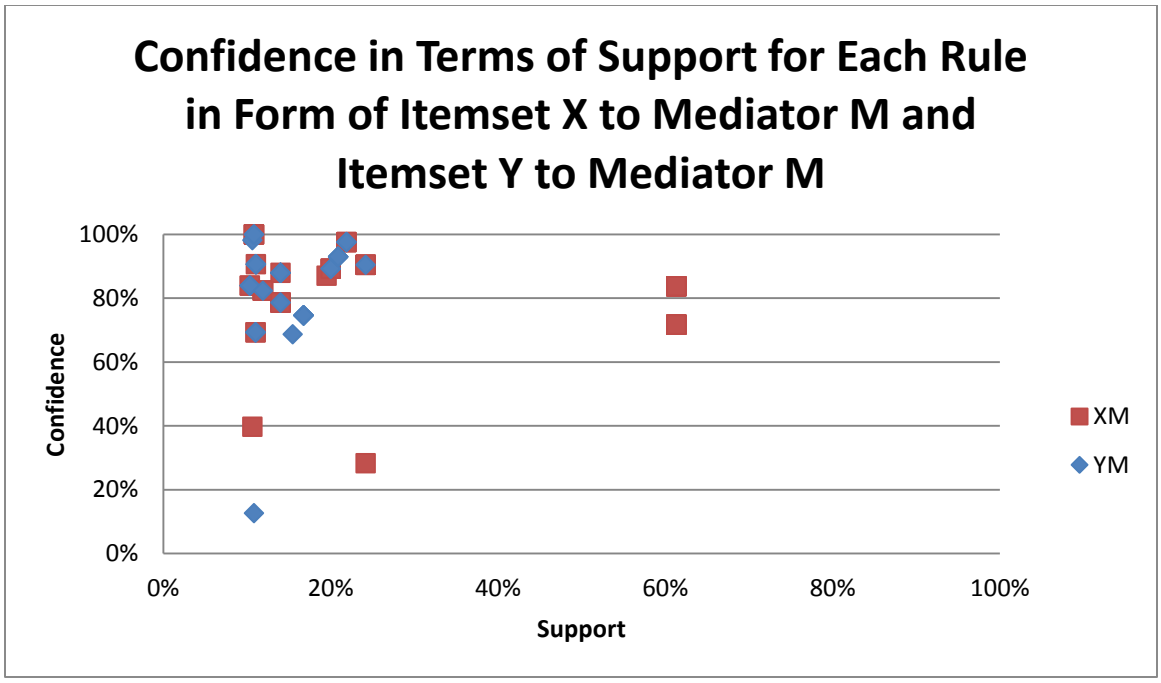
Additionally, it was clear after the initial results that the algorithm needed to be tweaked to account for crime data, extended to allow for the generation of more indirect association rules, and customized to ensure that the rules produced followed the form desired for this research. For this reason, the algorithm was extended further to allow for indirect association rules that could demonstrate an indirect relationship between item

sets versus single items. Also, there was a need to be able to specify which items to use in the potential mediator set, allowing for the ability to ensure that “Domestic” and “Arrest” appeared only within the mediator set, if applicable. In the following section, the effect of these extensions on the rules produced from the algorithm and dataset will be discussed.

5 RESULTS & ANALYSIS

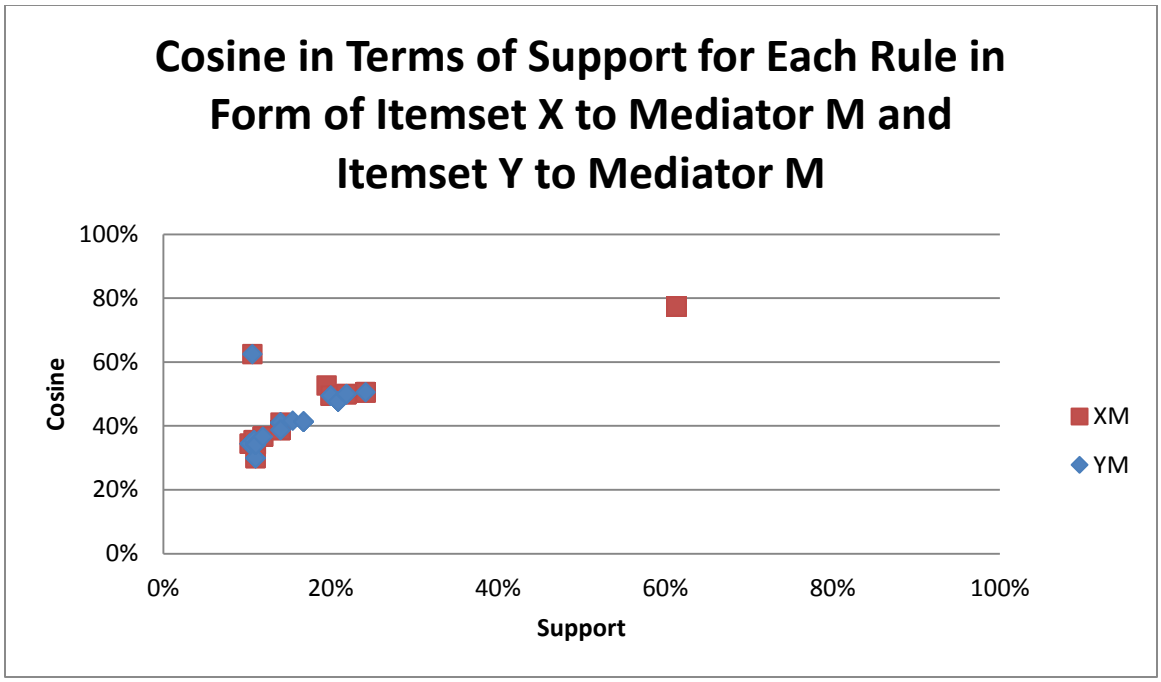
5.1 Initial Findings

The algorithm produced a set of 45 rules from the October 2011 to September 2013 training data set. Graph 5-A below shows the support value of each item set in the rule in relation to the confidence value of item set in the rule.



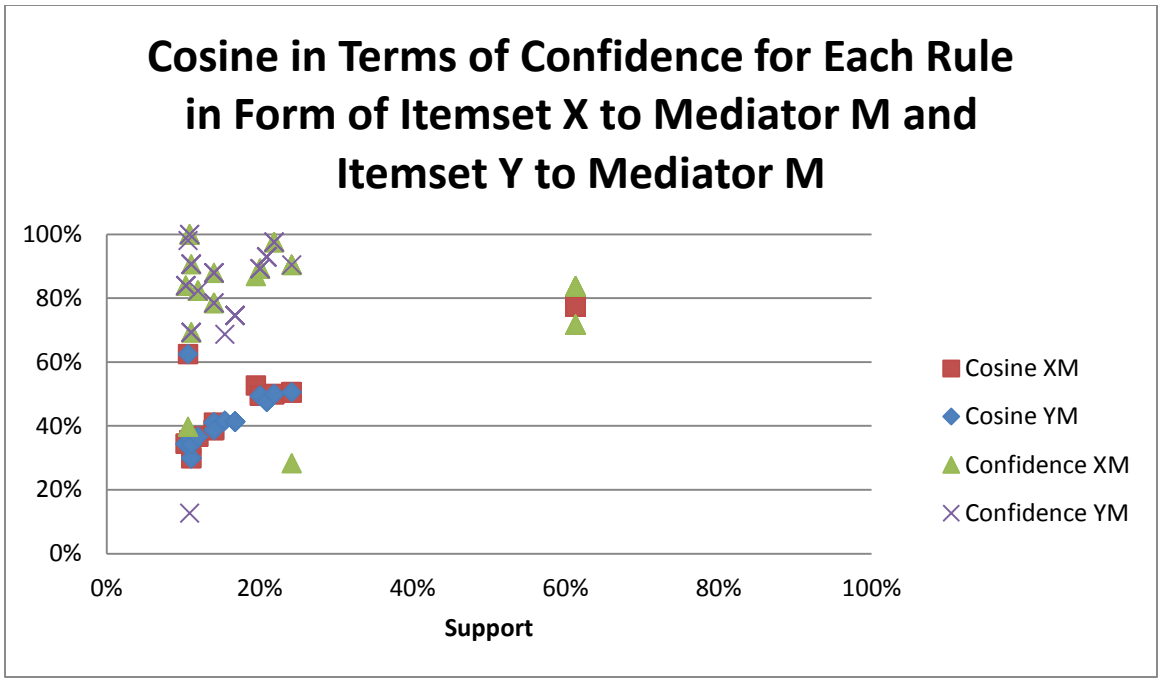
Graph 5-A: Confidence vs support for the rules generated from the training data set

Graph 5-B below shows the support value of each item set in the rule in relation to the cosine value of item set in the rule. While the confidence values for the majority of the rules landing within the lower support values are fairly high, the cosine values appear clustered in the middle of the range from about 0.3 to 0.65, essentially meaning that the values are not related but also are not independent of one another, thus showing some kind of relationship between the values.



Graph 5-B: Cosine vs support for the rules generated from the training data set

Graph 5-C below shows a combinations of the two graphs shown previously in order to display any relationship between the two values. The two calculations appear to cluster in the same general pattern, however, there is no real overlap in the values computed or the range in which the values land.



Graph 5-C: Cosine/Confidence vs support for the rules generated from the training data set

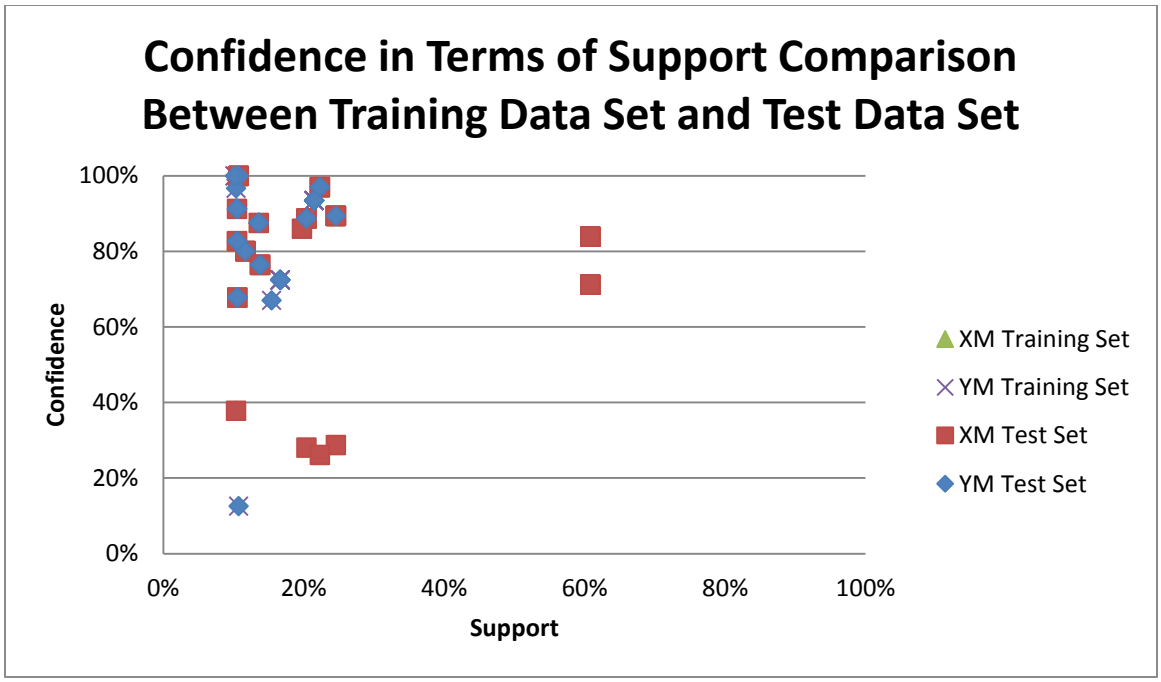
Unlike the above values seen for cosine, the lift and interest values computed for this data set were not as valuable. The lift values generated for all rules were essentially 0, showing that the values in question appear less often together than may have been expected and are negatively correlated. The interest values generated for all rules were also essentially 0, meaning that the items in the item set do not have a strong dependency on one another. These parameters are used to measure how related items are, so when considering them in an algorithm that examines items that are indirectly associated, these insignificant 0 values would be expected. However, by computing the values from one item to the mediator and the other item to the mediator, it was expected that these parameters may produce values of more interest, i.e. values other than 0, but that was not the outcome. This could be due to the fact that both interest and lift do not have the null-

invariant property and cosine does, meaning that unrelated items to the data of interest do not affect the cosine association.

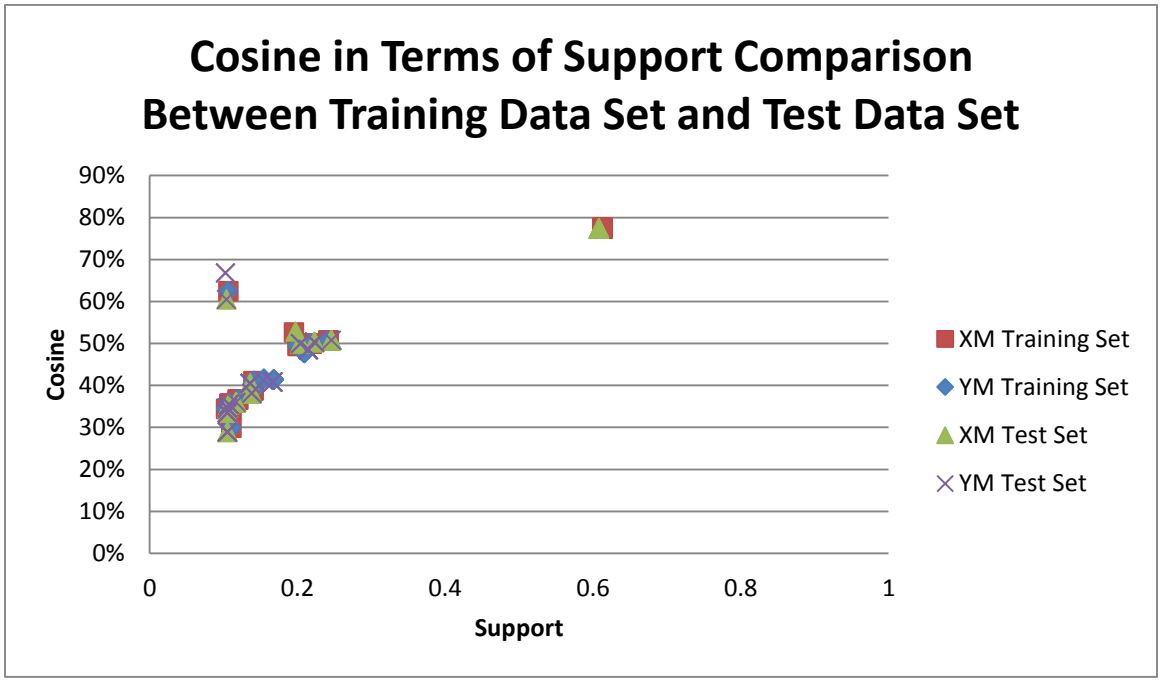
The indirect association rule mining algorithm has the potential to produce rules in which the mediator is a set of items rather than just a single item, and from the test set only one of those rules was generated. This rule was {“Theft” “Street” | “NOT Domestic” “NOT Arrest”} and it could be considered a rule that could have been deduced without the help of an association rule mining algorithm.

5.2 Training Set vs Test Set

Next, the training set findings were compared to the test set findings in order to determine if the initial analysis held true for other sets of data within the set. The algorithm produced 45 rules from the training set of data in comparison to the 47 produced from the test set, where all 45 from the training set were present within the test set. Graph 5-D and 5-E below compare the confidence values and cosine values found from each of the data sets and shows a near exact overlap between the two.



Graph 5-D: Confidence vs support for the rules generated from the training and test data sets



Graph 5-E: Cosine vs support for the rules generated from the training and test data sets

These graphs show that within the data sets, the associations that have been found are represented nearly equally throughout, meaning that these rules could be trusted against previous, un-analyzed data from the set and future crime records from the data set.

5.3 Indirect Association Crime Rules

Of the 931, 518 records analyzed over the two data sets, there were only 110,472 records, or about 11.9%, that were identified as "NOT Arrest" and "Domestic". For this reason, when running the algorithm over the data set, the minsup threshold was set to 0.08, the ts threshold was set to 0.07, and the minconf threshold was set to 0.05. These values were chosen in order to gather a large set of rules from the data set to analyze, but due to the large distribution of attributes, the support values could not be very high. For example, the "Primary Type" attribute only had 32 possible values to it while the "Location Description" attribute had 130 possible values. By setting the thresholds low, the aim was to be able to capture those attributes that had a larger number of possible values. There were 5 indirect association rules discovered that demonstrated this behavior and they are displayed in Table 5-1.

Item X	Item Y	Mediator
Domestic	Street	NOT Arrest
Domestic	Theft	NOT Arrest
Domestic	Battery	NOT Arrest
Domestic	Residence	NOT Arrest
Domestic	Apartment	NOT Arrest

Table 5-1: Indirect Association Rules discovered that incorporate the "Domestic" and "NOT Arrest" attributes.

Alone, these rules do not give much insight into the crime that took place, however, these incidents occurred often enough that the algorithm identified them as ones that have

value. As it stands currently, these rules do not show much that is interesting to the user. Additionally, some rules outside of the five rules shown here would not be considered to be viable. For example, one rule found was in the form of {“NOT Domestic” “Domestic” | “NOT Arrest”} because the algorithm simply looks for a single value in each set that differs from the other frequent set, so this rule is valid for the algorithm but not for the context of the crime data.

In order to make the rules produced more interesting with respect to a crime data set, further analysis needed to be done. This is where the additional extensions mentioned in the previous section were employed. The values of “Arrest”, “Domestic”, “NOT Arrest”, and “NOT Domestic” were restricted to only occurring within the mediator item set because they were the focus of the work. This allowed for rules to be generated that gave some more insight into the indirect associations existing within the data set with relation to these values. Also, the algorithm was altered to allow for indirect relations between an item and another item set. The algorithm initially builds these rules from frequent item sets that are generated. It would pick two of these frequent item sets, and for each value in the first set it would compare it with each value in the second set trying to find only one item that is different between the sets. This method was altered so that for each item in the first set, it would find the set of items in the second set that differ from it, removing the restriction of only finding a single item difference. Table 5-2 below shows the new rules that were found in common after running this new algorithm back over each of the data sets.

Item X	Item Set Y	Mediator
Battery	Theft, $500 and Under	NOT Arrest
Criminal Damage	Battery, Domestic Battery Simple	NOT Arrest
Theft	Battery, Domestic Battery Simple	NOT Arrest
$500 and Under	Battery, Domestic Battery Simple	NOT Arrest
Apartment	Battery, Domestic Battery Simple	NOT Arrest
Apartment	Theft, $500 and Under	NOT Arrest
Residence	Battery, Domestic Battery Simple	NOT Arrest
Street	Battery, Domestic Battery Simple	NOT Arrest
Narcotics	Battery, Domestic Battery Simple	Arrest
Narcotics	Theft, $500 and Under	Arrest
Battery	Theft, $500 and Under	NOT Domestic
Criminal Damage	Battery, Domestic Battery Simple	NOT Domestic
Narcotics	Battery, Domestic Battery Simple	NOT Domestic
Narcotics	Theft, $500 and Under	NOT Domestic
Theft	Battery, Domestic Battery Simple	NOT Domestic
$500 and Under	Battery, Domestic Battery Simple	NOT Domestic
Residence	Battery, Domestic Battery Simple	NOT Domestic
Sidewalk	Battery, Domestic Battery Simple	NOT Domestic
Sidewalk	Theft, $500 and Under	NOT Domestic
Street	Battery, Domestic Battery Simple	NOT Domestic
Street	Theft, $500 and Under	NOT Domestic
Criminal Damage	Theft, $500 and Under	NOT Arrest, NOT Domestic
Residence	Theft, $500 and Under	NOT Arrest, NOT Domestic
Street	Theft, $500 and Under	NOT Arrest, NOT Domestic
Narcotics	Theft, $500 and Under	Arrest, NOT Domestic

Table 5-2: Indirect Association Rules generated from the extended algorithm over both data sets.

Again, these rules could be interesting, but there is still not much information about the crime other than the type of crime it was and its location. Taking the work one step further, the focus was shifted to look at the "Domestic" aspect of the data set. The test data set was stripped down to contain only records that contain the "Domestic" attribute, resulting in 94,885 records. As can be seen in table 5-3, many more rules were produced with the incorporation of additional data points.

Item X	Item Set Y	Mediator
Assault	Battery, Domestic Battery Simple	NOT Arrest
Assault	Battery, Apartment	NOT Arrest
Assault	Battery, Residence	NOT Arrest
Assault	Domestic Battery Simple, Apartment	NOT Arrest
Assault	Domestic Battery Simple, Residence	NOT Arrest
Other Offense	Battery, Domestic Battery Simple	NOT Arrest
Other Offense	Battery, Apartment	NOT Arrest
Other Offense	Battery, Residence	NOT Arrest
Other Offense	Domestic Battery Simple, Apartment	NOT Arrest
Other Offense	Domestic Battery Simple, Residence	NOT Arrest
Street	Battery, Domestic Battery Simple	NOT Arrest
Street	Battery, Apartment	NOT Arrest
Street	Battery, Residence	NOT Arrest
Street	Domestic Battery Simple, Apartment	NOT Arrest
Street	Domestic Battery Simple, Residence	NOT Arrest
Assault	Battery, Domestic Battery Simple	Domestic
Assault	Battery, Apartment	Domestic
Assault	Battery, Residence	Domestic
Assault	Domestic Battery Simple, Apartment	Domestic
Assault	Domestic Battery Simple, Residence	Domestic
Other Offense	Battery, Domestic Battery Simple	Domestic
Other Offense	Battery, Apartment	Domestic
Other Offense	Battery, Residence	Domestic
Other Offense	Domestic Battery Simple, Apartment	Domestic
Other Offense	Domestic Battery Simple, Residence	Domestic
Simple	Battery, Domestic Battery Simple	Domestic
Simple	Battery, Apartment	Domestic
Simple	Battery, Residence	Domestic
Simple	Domestic Battery Simple, Apartment	Domestic
Simple	Domestic Battery Simple, Residence	Domestic
Street	Battery, Domestic Battery Simple	Domestic
Street	Battery, Apartment	Domestic
Street	Battery, Residence	Domestic
Street	Domestic Battery Simple, Apartment	Domestic
Street	Domestic Battery Simple, Residence	Domestic
District7	Battery, Domestic Battery Simple	Domestic
District7	Battery, Apartment	Domestic
District7	Battery, Residence	Domestic
District7	Domestic Battery Simple, Apartment	Domestic
District7	Domestic Battery Simple, Residence	Domestic
CommArea25	Battery, Domestic Battery Simple	Domestic
CommArea25	Battery, Apartment	Domestic

CommArea25	Battery, Residence	Domestic
CommArea25	Domestic Battery Simple, Apartment	Domestic
CommArea25	Domestic Battery Simple, Residence	Domestic

Table 5-3: Indirect Association Rules generated from the extended algorithm over the "Domestic" data set.

This table shows that the changes made to the algorithm and data set have improved the rules produced based on the desired outcome. Further updates have the potential to continue to improve results and find more interesting indirect relationships.

6 CONCLUSION & FUTURE WORK

With any data mining algorithm, it is important to ensure that the data that is being mined "fits" the algorithm. It is important to know that the algorithm not only has the potential to produce rules that are interesting to the work in question, but also that it is able to interpret the data in the correct way, or that the data could be modified to fit the algorithm's desire. Indirect association rule mining for crime data has the potential to provide interesting relationships among data, but it requires more data manipulation and rules than what was provided for this work. Some of this data manipulation has been done, but it is easy to see how more would need to be done in order to extract more meaningful relationships that incorporate all data points into the resulting rule set. It was discovered that the data needed to be trimmed to better analyze the “Domestic” attribute, and further trimming or selecting of data could better improve what is mined based on the desired outcome.

Additionally, with indirect association rule mining, the type of attributes that are being mined plays a large role. For example, a Boolean attribute with two values versus a string attribute with 300 values is going to show up many more times in the data set, thus throwing off the support value and ensuring that that Boolean value is present in nearly every rule generated. Depending on the desired outcome of the algorithm from the data set, this could be valuable. However, for this data, extending the algorithm and cleansing the data set into a set that has attributes with a few number of possible values would have been interesting because this would allow for more potential rules to be produced. By grouping common crime descriptors or location descriptors, the support for these values would have been higher, therefore resulting in more rules with more of those attributes

present. This would have also allowed for the use of higher support values and the tweaking of those thresholds in the generation of the association rules.

Also, the indirect association rule mining algorithm from the SPMF library uses an Apriori-style approach to generate frequent item sets. This is very slow and restricts the number of records that can be analyzed using the algorithm. It would be beneficial to update this algorithm and use an approach that is more efficient in generating these sets. Doing this would allow for a larger amount of data to be processed in a more efficient manner, possibly leading to the generation of more rules and easing the process for the developer.

Overall, the work done gave a starting point for employing the indirect association rule mining algorithm to discover rare associations within crime data. Extensions and further points of analysis have been identified in order to make more use of what the algorithm has to offer. These extensions proved to be valuable within the crime data analysis performed and have the potential to be taken further to potentially produce additional, varying rules.

Bibliography

[1] A. Sacasere, E. Omiecinski, and S. Navathe. “Mining for strong negative associations in a large database of customer transactions.” In Proc. of the *14th International Conference on Data Engineering*, pages 494-502, Orlando, Florida, February 1998.

[2] Buczak, A.L., Gifford, C.M. “Fuzzy association rule mining for community pattern discovery,” In *ACM SIGKDD Workshop on Intelligence and Security Informatics*, 2010.

[3] C. McCue, “Connecting the Dots: Data Mining and Predictive Analytics in Law Enforcement and Intelligence Analysis,” Copyright held by the International Association of Chiefs of Police. The Police Chief, vol. 70, no. 10, October 2003. Accessed 12/4/2014.

[4] C.M. Kuok, A. Fu, and M.H. Wong, “Mining Fuzzy Association Rules in Databases,” *ACM SIGMOD Record* 27(1), pp. 41-46, New York, NY, 1998.

[5] C. Rudin, R. Sevieri, D. Wagner, and T. Wang, “Learning to Detect Patterns of Crime”. http://web.mit.edu/rudin/www/WangRuWaSeECML13.pdf

[6] City of Chicago Data Portal, “Crimes – 2001 to present,” 2011. https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2

[7] Fournier-Viger, P., Gomariz, Gueniche, T., A., Soltani, A., Wu., C., Tseng, V. S. (2014). SPMF: a Java Open-Source Pattern Mining Library. Journal of Machine Learning Research (JMLR), 15: 3389-3393.

[8] H. Verlinde, M. De Cock, and R. Boute, “Fuzzy Versus Quantitative Association Rules: A Fair Data-Driven Comparison,” In *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 36, No. 3, June 2006.

[9] Hipp, Jochen, Ulrich Güntzer, and Gholamreza Nakhaeizadeh. "Algorithms for association rule mining—a general survey and comparison." *ACM sigkdd explorations newsletter* 2.1 (2000): 58-64.

[10] Kazienko, Przemysław. "Mining indirect association rules for web recommendation." *International Journal of Applied Mathematics and Computer Science* 19.1 (2009): 165-186.

[11] Merceron, Agathe, and Kalina Yacef. "Interestingness measures for association rules in educational data." *Educational Data Mining 2008*. 2008.

[12] Pang-Ning, Tan, Michael Steinbach, and Vipin Kumar. "Introduction to data mining." *Library of Congress*. 2006.

[13] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," *In Proceedings of the International Conference on Management of Data*, Washington, D.C., pp. 207-216, May 1993.

[14] R. Srikant and R. Arawal, "Mining Quantitative Association Rules in Large Relational Tables," In *Proceedings of the International Conference on Management of Data*, Montreal, Quebec, Canada, pp. 1-12, 1996.

[15] Tan, Pang-Ning, Vipin Kumar, and Jaideep Srivastava. "Indirect association: Mining higher order dependencies in data." *Springer Berlin Heidelberg.* 2000.

[16] Y. Sucahyo, R. Gopalan. "CT-ITL: Efficient Frequent Item Set Mining Using a Compressed Prefix Tree with Pattern Growth." In *14th Australasian Database Conference (ADC2003)*, Vol 17, Adelaide, Australia, 2003.

Appendix

The data sets, source code, and test results can be provided upon request.

Vita

RILEY ENGLIN

EDUCATION

Gonzaga University
Bachelor of Science, Computer Science — May 2013

Eastern Washington University
Master of Science, Computer Science — In-Progress

HONORS AND AWARDS

Graduate Assistantship
Computer Science Department, Eastern Washington University — 2013 – 2015

PROFESSIONAL EXPERIENCE

Graduate Assistant, Instructor
Eastern Washington University — 2013 – 2015

WORK AND INTERNSHIPS

Application Engineer
Pacific Northwest National Lab, Richland, WA — March 2015 – Current

Database Development Intern
Career Path Services, Spokane, WA — May 2014 – March 2015

Development Intern
Zipline Interactive, Spokane, WA — January 2014 – March 2014