Fuzzy Association Rule Mining for Community Crime Pattern Discovery

Anna L. Buczak
Johns Hopkins University Applied Physics Laboratory
11100 Johns Hopkins Rd, Laurel, MD 20723 USA
Anna.Buczak@jhuapl.edu

Christopher M. Gifford
Johns Hopkins University Applied Physics Laboratory
11100 Johns Hopkins Rd, Laurel, MD 20723 USA
Christopher.Gifford@jhuapl.edu

ABSTRACT
Current manual inspection of crime data by analysts is limited, primarily due to the amount of data that can be processed concurrently and in a reasonable time frame. Further, complex relationships between various crime attributes can be overlooked by human analysts. Providing automated knowledge discovery tools becomes attractive to accelerate the efforts of local law enforcement. In this paper, we study the application of fuzzy association rule mining for community crime pattern discovery. Discovered rules are presented and discussed at regional and national levels. Rules found to hold in all states, be consistent across all regions, and subsets of regions are also discussed. A relative support metric was defined to extract rare, novel rules from thousands of discovered rules. Such an approach relieves the need of law enforcement personnel to sift through uninteresting, obvious rules in order to find interesting and meaningful crime patterns of importance to their community.

Categories and Subject Descriptors
I.2.3 [**Artificial Intelligence**]: *Deduction and Theorem Proving – deduction, uncertain, "fuzzy," and probabilistic reasoning.*

General Terms
Algorithms, Performance, Experimentation, Theory.

Keywords
Crime data mining, fuzzy association rules, rule pruning, community-based crime.

1. INTRODUCTION
Crime data mining is receiving increased attention to discover underlying patterns in crime data. The need to act quickly to suppress crime activity and discover links between various data sources persists. State law enforcement are continuing to call upon modern geographic information systems and data mining technologies to enhance crime analytics and better protect their communities and assets. Real-time solutions can save significant resources and push the capability of law enforcement closer to the pulse of criminal activity.

Modern computing systems provide a unique opportunity to study this vast amount of data in ways that were previously not feasible. The volume of data being digitally recorded about crimes, suspicious activities, and suspect records is at an all-time high. These data are typically multi-dimensional and too large to manually examine to discover salient patterns which offer significant leads. The nature and sensitivity of the data present important issues that need to be addressed, such as data storage, warehousing, and privacy.

Current manual inspection of crime data by analysts and investigators is limited, primarily due to the amount of data that can be processed concurrently and in an acceptable time frame. Further, complex relationships between various crime attributes can be overlooked or misinterpreted by human analysts. Providing automated knowledge discovery tools becomes attractive to enhance and accelerate the efforts of local law enforcement.

Local, regional, national, and international crime play important roles in allocating law enforcement resources and influencing investigative priorities across jurisdictions. For example, an aggravated assault is a local jurisdiction matter, whereas drug trafficking and terrorism exhibit regional and global implications. Local crime patterns may differ from surrounding communities, creating localized trends of criminal activity which are unique to a community. Similarly, certain crimes may be more probable in locations with higher populations and dense housing. Regional crime patterns can be discovered which enable law enforcement personnel and criminal investigators to address large-scale trends.

Crime is typically temporally, thematically, and geospatially correlated, exhibiting complexities which make the analyst's task very challenging. Moreover, evidence can be loosely coupled while being geospatially sparse, forcing a more widespread analysis effort. Leveraging data mining techniques provides the ability to better analyze, predict, prepare for, and respond to criminal acts and potential security risks.

Community-based data can be integrated to study associations between socio-economic characteristics and local law enforcement information. Examples of such national crime data sources are the U.S. Census, U.S. FBI Uniform Crime Report, U.S. Law Enforcement Management and Administrative Statistics survey, National Criminal Record Database, and National Archive of Criminal Justice Data.

In this paper, we study the application of fuzzy association rule mining for community crime pattern discovery. The following sections discuss available crime data sources, previous criminal act analysis efforts, and techniques for crime data mining. Fuzzy association rule mining is introduced as a novel means for knowledge discovery in the crime domain, supported by experimental results on the open-source Communities and Crime data set [2]. This paper concludes with a discussion on directions for further research.

ISI-KDD 2010, July 25, 2010, Washington, D.C., USA

2. CRIMINAL ACT ANALYSIS

Analysts differ in their methods of scouring data sources to discover patterns. Rules and heuristics which govern the determination of what represents important crime information also vary from analyst to analyst. These aspects construct a challenging problem for creating an automated criminal act analysis system.

There exist several applications of crime data analysis which have been studied by the research community. Most efforts focus on crime nature, severity, location, duration, and frequency. Following are several current trends in crime analysis:

- Geospatial, map-based visualization.
- Geographical clustering of crime activity, such as identifying hot spots.
- Serial criminal behavioral pattern profiling and criminal career analysis.
- Gang criminal network analysis.
- Data stream anomaly, novelty, or outlier detection.
- Temporal analysis of crime patterns, such as crime sprees (temporal association of crime from an individual or group).
- Linking threats to risk of critical infrastructure based on vulnerability assessments.

Data mining plays a key role in each of these crime data exploitation applications. Knowledge discovery is one accepted tool for identifying underlying novel patterns in large volumes of crime data. For example, association rule mining has become a prominent method for performing knowledge discovery on large transactional databases. Other popular data mining techniques employed for similar purposes are summarized below:

- Semantic analysis and text mining for entity extraction from free-text narratives, police reports, and FBI bulletins [6][7][11].
- Rule-based, expert systems established through knowledge engineering. The utility of this technique is limited due to the dynamic nature of crime. It is also difficult to quantify and adequately capture the knowledge of field experts with significant experience.
- Clustering and graph representations [16], both for identifying similar crimes and for visualization purposes. Cluster size, shape, and distribution can aid in inferring details about related crimes. Clustering is also utilized to group classes of criminals.
- Machine learning and classification for crime pattern recognition.
- Case-based reasoning for identifying closed cases exhibiting characteristics similar to open cases.

Research efforts for mining crime patterns have recently increased, focusing on a variety of approaches. For example, knowledge discovery was performed by mining association rules, training a classifier for prediction, and utilizing clustering methods in [3] using a US State database. The research in [14] takes a geospatial approach at discovering crime patterns by clustering and displaying regional crime on a map. The k-means clustering algorithm was coupled with an attribute importance-weighting algorithm to cluster crimes by type.

Various commercial and research systems have experienced success when applied to portions of the overarching crime analysis problem. COPLINK [10] is one such system, aimed at improving criminal-intelligence analysis through the use of a co-occurrence concept space built from detailed case reports and terms of interest. The concept space utilizes five primary categories to study link analysis: Person, Organization, Location, Crime, and Vehicle. The COPLINK system was found to effectively increase operating efficiency, while improving case closure and solvability ratings.

Based on experience gained from the COPLINK project, a framework for crime data mining is presented in [8]. The authors categorize and discuss levels of implication for various types of crime established via consult from an experienced local detective. General techniques for analyzing crime data are summarized, such as entity extraction, clustering, association rule mining, and sequential pattern analysis. Similarly, The Regional Crime Analysis Program (RECAP) [5] was created to assist Virginia law enforcement. RECAP incorporated aspects of data fusion, data mining, and geospatial clustering of crime. Crime analysis tools are now available for a wide variety of GIS software packages.

The authors of [15] developed an incremental mining algorithm, called ITAR, for crime pattern discoveries via temporal association rules. As databases grow large, incremental approaches are necessary to avoid expensive database rescanning operations. Mining temporal association rules can yield new insights into crime trends for various time frames and how they change. The ITAR algorithm was applied to crime data for a district of Hong Kong, organized by offence and modus operandi (MO) with various categorizations of seriousness. For a general discussion of data mining to the crime domain, the reader is referred to [19].

3. FUZZY ASSOCIATION RULE MINING

3.1 Association Rules

The goal of data mining is to discover inherent and previously unknown information from data. When the knowledge discovered is in the form of association rules, the methodology is called association rule mining. An association rule describes a relationship among different attributes. Association rule mining was introduced by Agrawal et al. [1] as a way to discover interesting co-occurrences in supermarket data (the market basket analysis problem). It finds frequent sets of items (i.e., combinations of items that are purchased together in at least *n* transactions in the database), and generates from the frequent itemsets such as {X, Y} association rules of the form X → Y and/or Y → X.

More formally, let D = $\{t_1, t_2, ..., t_n\}$ be the transaction database and let t_i represent the i^{th} transaction in D. Let I=$\{i_1, i_2, \ldots i_m\}$ be the universe of items. A set $X \sqsubseteq I$ of items is called an itemset. When X has k elements, it is called a k-itemset. An association rule is an implication of the form X → Y, where $X \subset I, Y \subset I$ and $X \cap Y = \emptyset$.

The support of an itemset X is defined as:

$$Support(X) = \frac{number\ records\ with\ X}{number\ records\ in\ D}$$

The support of a rule (X → Y) is defined as:

$$Support(X \rightarrow Y) = \frac{number\ records\ with\ X\ and\ Y}{number\ records\ in\ D}$$

The confidence of a rule (X → Y) is defined as:

$$Confidence(X \rightarrow Y) = \frac{number\ records\ with\ X\ and\ Y}{number\ records\ with\ X}$$

Confidence can be treated as the conditional probability (P(Y|X)) of a transaction containing X and also containing Y.

A high confidence value suggests a strong association rule. However, this can be deceptive. For example, if the antecedent or consequent have a high support, they could have a high confidence even if they were independent. This is why *lift* was suggested as a useful metric.

The lift of a rule (X → Y) measures the deviation from independence of X and Y. A lift greater than 1.0 indicates that transactions containing the antecedent (X) tend to contain the consequent (Y) more often than transactions that do not contain the antecedent (X). The higher the lift, the more likely that the existence of X and Y together is not just a random occurrence, but rather due to the relationship between them.

$$Lift(X \rightarrow Y) = \frac{Confidence(X \rightarrow Y)}{Support(X \rightarrow Y)}$$

Apriori [1] is the most widely used algorithm for finding frequent k-itemsets and association rules. It exploits the downward closure property which states that if any k-itemset is frequent, all of its subsets must be frequent as well. The Apriori algorithm proceeds as follows:

1. Calculate the support of all 1-itemsets and prune any that fall below the minimum support, specified by the user.

Loop:

2. Form candidate k-itemsets by taking each pair (p,q) of (k-1) itemsets where all but one item match. Form each new k-itemset by adding the last item of q onto the items of p.
3. Prune the candidate k-itemsets by eliminating any itemset that contains a subset not in the frequent (k-1)-itemsets.
4. Calculate the supports of the remaining candidate k-itemsets and eliminate any that fall below the specified minimum support. The result is the frequent k-itemsets.

3.2 Fuzzy Association Rules

A limitation of association rule mining is that it only works on binary transaction data (i.e., an item was either purchased in a transaction (1) or not (0)). In many real-world applications, data is either categorical (e.g., blue, red, green) or quantitative (e.g., number of murders). For numerical and categorical attributes, Boolean rules are unsatisfactory. Extensions have been proposed to operate on these data, such as quantitative association rule mining [18] and fuzzy association rule mining [12].

Fuzzy association rules are of the form:

(X is A) → (Y is B)

where X and Y are attributes, and A and B are fuzzy sets that characterize X and Y respectively. An example fuzzy association rule is the following:

(temperature, hot) and (humidity, high) → (energy-usage, high)

Fuzzy association rules are easily understood by humans because of the linguistic terms that they employ (e.g., hot, high). Fuzzy logic assigns degree of membership between 0 and 1 (e.g., 0.4) to each element of a set, allowing for a smooth transition between membership and non-membership of a set. The measures of support, confidence and lift have been fuzzified for the purpose of fuzzy association rules.

The fuzzy support is defined as:

$$Fuzzy\ Support(< X, A >) = \frac{sum\ of\ votes\ satisfying < X, A >}{number\ of\ records\ in\ D}$$

Let D = $\{t_1, t_2, ..., t_n\}$ be the transaction database and let t_i represent the i^{th} transaction in D. Let's define the itemset-fuzzy set pair <X, A> where X is the set of attributes x_j and A is the set of fuzzy sets a_j. A transaction satisfies <X, A> means that the vote of the transaction is greater than zero. The vote of a transaction is calculated by the membership grade of each x_j in that transaction. The *membership_grade* for attribute a_j in transaction t_i is defined as:

$$membership\_grade_{a_j}(t_i[x_j]) = \begin{cases} membership\_function_{a_j}(t_i[x_j]), & if \geq threshold \\ 0, otherwise \end{cases}$$

The fuzzy support of a 1-itemset is defined as:

$$Fuzzy\ Support(< X, A >) = \frac{\sum_{t_i \in T}(membership\_grade_{a_j}(t_i[x_j]))}{number\ of\ records\ in\ D}$$

The fuzzy support of k-itemsets is defined as:

$$Fuzzy\ Support(< X, A >) = \frac{\sum_{t_i \in T} TNorm_{x_i \in X}(membership\_grade_{a_j}(t_i[x_j]))}{number\ of\ records\ in\ D}$$

where *TNorm* can be any of T-Norm operators [21]: product, minimum, etc.

When a frequent itemset <X, A> is obtained, fuzzy association rules of the form "If X is A then Y is B" are generated, where $X \subset Z$, $Z = X \cup Y$, $A \subset C$ and $C = A \cup B$.

The fuzzy confidence value can be computed as follows:

$$Fuzzy\ Confidence(< X, A >, < Y, B >) = \frac{Fuzzy\ Support\ of < XUY, AUB >}{Fuzzy\ Support\ of < X, A >}$$

$$Fuzzy\ Confidence(< X, A >, < Y, B >) = \frac{\sum_{t_i \in T} TNorm_{z_i \in Z}(membership\_grade_{a_j}(t_i[z_j]))}{\sum_{t_i \in T} TNorm_{x_i \in X}(membership\_grade_{a_j}(t_i[x_j]))}$$

where $Z = X \cup Y$ and $C = A \cup B$.

Finally, the fuzzy lift measure is defined as:

$$Fuzzy\ Lift\ (X \rightarrow Y) = \frac{Fuzzy\ Confidence\ (X \rightarrow Y)}{Fuzzy\ Support(X \rightarrow Y)}$$

For the purpose of mining fuzzy association rules, Apriori was extended to Fuzzy Apriori. The difference between the two algorithms is that Fuzzy Apriori uses definitions of fuzzy support and fuzzy confidence instead of their crisp counterparts used in Apriori.

3.3 Rare Association Rules

Most association rule mining algorithms, such as Apriori, concentrate on finding frequent itemsets (i.e., itemsets with at

least a specified minimum support (*minsup*)). The rules are called strong association rules when they meet or exceed a minimum confidence (*minconf*).

Occasionally the rules of interest have high confidence but a low support. Such rules are called rare association rules. If one wanted to determine in a set of supermarket transactions if there was a relationship between buying a food processor and a cooking pan, this would be difficult due to the fact that each of these items is rarely purchased. Thus, even though the two items are almost always purchased together, this association is usually not found since its support is too low [13]. When dealing with rare diseases, violent crimes, machinery failure, etc., one is interested in finding rare association rules.

Rare association rule mining is a newer and less well understood discipline than frequent association rule mining. One of the approaches to rare association rule mining is to use the same algorithms as for frequent item mining (such as Apriori) while selecting a low minimum support. Setting *minsup* very low causes a combinatorial explosion in the number of generated itemsets and rules. This necessitates the use of rule post-pruning methods, which facilitates extraction of interesting rules from a large set.

One of the methods we utilize for rule pruning is the consequent-constraint rule pruning [4], in which an item constraint is used that requires the consequents of the rules to satisfy a given constraint. This method requires prior knowledge of which consequents should be interesting.

Rules are additionally pruned based on their support, confidence, and lift. Confidence- and lift-based pruning methods are the same for frequent and rare rule mining. Support-based pruning must be different, since in frequent rule mining it is usually trivial to find the minimum support that is adequate for the entire data set. In contrast, rare rule mining requires setting the *minsup* to low, causing a combinatorial explosion of the number of rules.

Yun et al. [20] proposed the Relative Support Apriori (RSAA) algorithm to generate rules in which significant rare itemsets take part. This technique uses relative support, defined as:

$$RSup\ (X \rightarrow Y) = \frac{Support(X \rightarrow Y)}{min(Support(x_1), \dots, Support(x_k), Support(y_1), \dots, Support(y_l))}$$

This algorithm decreases the support threshold for items which have low frequency, and increases the support threshold for items that have high frequency.

4. COMMUNITIES AND CRIME DATA

4.1 Data Set

The Communities and Crime Data Set [2], available from the UCI Machine Learning Repository, was utilized for this study. This data set focuses on United States communities, and combines socio-economic data from the 1990 US Census, law enforcement data from the 1990 Law Enforcement Management and Administrative Statistics (LEMAS) survey, and crime data from the 1995 US FBI Uniform Crime Report. The data set was submitted to the UCI Machine Learning Repository in July 2009. The reader is referred to [17] for work involving a related data set.

The data set is comprised of 2215 total instances and 128 attributes for communities from each state. Some communities were omitted based on occurrence of significant missing or known incorrect crime statistics, the majority of which were from the Midwest. Certain attributes contain a significant number of missing values for which the data was unavailable or not recorded for particular communities (e.g., *Police Officers Per 100K Population*, *Police Request Per Officer*, *Officers Assigned to Drug Units*, *Police Operating Budget*).

Attributes include information across a variety of crime-related facets, ranging from the percent of officers assigned to drug units, to population density and percent considered urban, to median household income. Also included are measures of crimes considered violent, which are murder, rape, robbery, and assault. For more detail on the attributes and their statistics, the reader is referred to [2].

4.2 Experimental Setup

The following mechanisms were used to select the final set of attributes:

- All attributes with a large number of missing values were removed.
- Odds ratios between each remaining attribute and *Violent Crimes*, *Murders*, *Robberies*, and *Assaults* were computed. Attributes exhibiting small odds ratios were removed.
- Similar attributes were omitted (e.g., from the attributes *Divorced (%)*, *Male Divorced (%)*, and *Female Divorced (%)*, only *Divorced (%)* was kept).

This methodology yielded 40 attributes, enumerated below, that were utilized for this study. Those attributes whose value represents a percentage are indicated with (%).

1. *Mean People Per Household*
2. *Race: African American (%)*
3. *Race: Caucasian (%)*
4. *Race: Asian (%)*
5. *Race: Hispanic (%)*
6. *Age: 12-21 (%)*
7. *Age: 12-29 (%)*
8. *Age: 16-24 (%)*
9. *Age: 65+ (%)*
10. *People in Urban Area (%)*
11. *Median Household Income*
12. *Houses with Salary Income (%)*
13. *Houses with Social Security Income (%)*
14. *Houses with Public Assistance Income (%)*
15. *Houses with Retirement Income (%)*
16. *Per Capita Income*
17. *People Under Poverty Level (%)*
18. *Education: Less than 9th Grade (%)*
19. *Education: No High School Diploma (%)*
20. *Education: Bachelor's or Higher (%)*
21. *Unemployed (%)*
22. *Employed (%)*
23. *Divorced (%)*
24. *Houses with Kids Living with Two Parents (%)*
25. *Kids Born to Never Married (%)*
26. *People Speaking English Only (%)*
27. *People Speaking No English (%)*
28. *People in Dense Housing (%)*
29. *People in Owner Occupied Households (%)*
30. *Occupied Housing Units Without Phone (%)*
31. *Median Gross Rent*

32. *People in Homeless Shelters*
33. *Homeless People Counted in Street*
34. *Foreign Born (%)*
35. *Population Density (Persons Per Square Mile)*
36. *People Commute Using Public Transit (%)*
37. *Violent Crimes Per 100K Population*
38. *Murders*
39. *Robberies*
40. *Assaults*

4.3 Data Preprocessing

The majority of the 40 selected attributes are percentages, or are computed per 100K population. However, certain attributes (e.g., *People in Homeless Shelters*, *Homeless People Counted in Street*, *Murders*, *Robberies*, *Assaults*) are absolute values. Our first attempt to generate rules with consequents such as *Murders*, *Robberies*, and *Assaults* gave us rules that simply pinpointed big cities (e.g., San Francisco, Los Angeles, Chicago). Five variables were redefined by computing their values per 100k population: *People in Homeless Shelters*, *Homeless People Counted in Street*, *Murders*, *Robberies* and *Assaults*. The transformed data was used in the study. In the remainder of this paper, when referring to the variables computed per 100k population, we will use their proper names. For example, we will refer to *Violent Crimes* instead of *Violent Crime Per 100K Population.*

4.4 Data Fuzzification

A set of membership functions were defined for each of the 40 attributes. Membership functions were defined using subject matter expert knowledge and statistical measures such as mean and standard deviation.

As an example, consider the attribute *Houses with Public Assistance Income (%)* (Figure 1). Three membership functions have been defined (Figure 2): *Low*, *Medium*, and *High*.

For the attribute *Violent Crimes* (Figure 3), four membership functions have been defined (Figure 4): *Blank* (missing data), *Low*, *Medium*, and *High*. Four similar membership functions have been defined for *Robberies* and *Assaults*. *Murders* has the following membership functions defined: *None*, *Low*, *Medium*, and *High*.

4.5 US Regions

Community data were grouped into five regions: Northeast, Southeast, Midwest, Southwest, and West [9]. Northeast comprises the following states: CT, DE, ME, MD, MA, NH, NJ, NY, PA, RI, and VT. This subset covers 632 communities. Southeast encompasses the states: AL, AR, FL, GA, KY, LA, MS, NC, SC, TN, VA, and WV. This subset covers 420 communities. Midwest is composed of the following states: IL, IN, IA, KS, MI, MN, MO, NE (no data), ND, OH, SD, and WI. It covers a total of 513 communities. Southwest encompasses the states: AZ, NM, OK, and TX. This subset covers 228 communities. Lastly, West is composed of the following states: CA, CO, ID, MT (no data), NV, OR, UT, WA, and WY. This represents 418 communities.

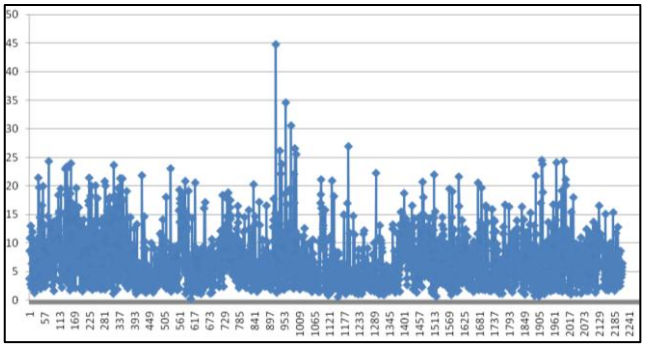

Figure 1. Houses with Public Assistance Income attribute.

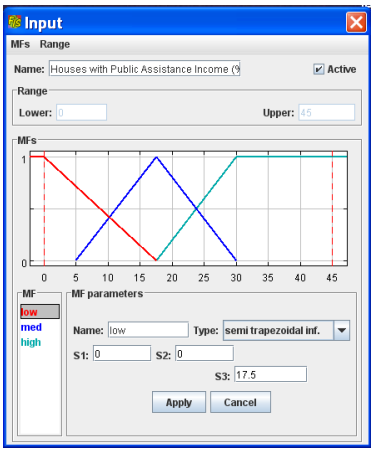

Figure 2. Houses with Public Assistance Income memberships.

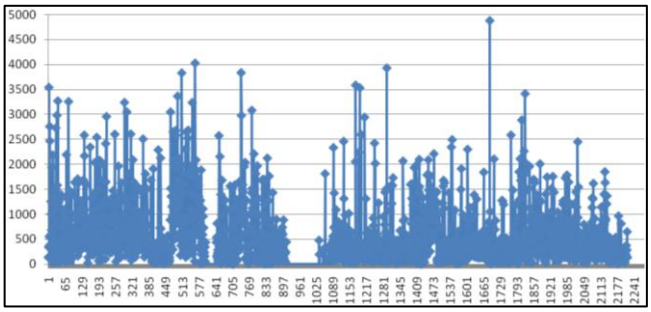

Figure 3. Violent Crimes attribute.

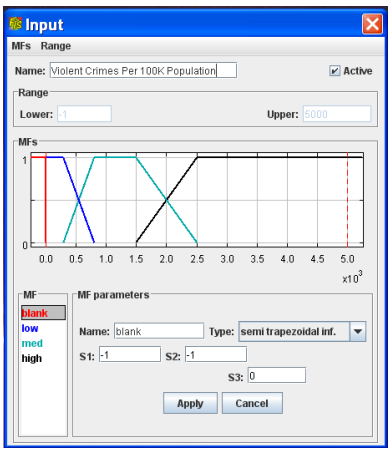

Figure 4. Violent Crimes memberships.

5. METHODOLOGY

The developed methodology has the following primary steps:

1. Variable fuzzification. This includes defining membership functions for each of the variables, and computing the membership values for each data item.
2. Running the Fuzzy Apriori algorithm on the data set. This includes initial pruning of the generated rules.
3. Rule post-pruning.

The initial pruning of the rules includes the consequent-constraint rule pruning method [17] mentioned in Section 3.3. We have developed a similar method that we call antecedent-constraint rule pruning, in which an item constraint is used that requires the antecedents of the rules to satisfy a given constraint. This is the second technique used in our work. This technique requires prior knowledge of which items are of interest in the antecedent. In an application such as the crime domain, the user usually knows very well which attributes are of interest as antecedents or consequents.

Rule post-pruning is concerned with pruning rules after they have been generated by an algorithm, such as Fuzzy Apriori. We are post-pruning rules based on 60% fuzzy confidence.

We have also developed a new Relative Fuzzy Support measure:

$$Relative\ Fuzzy\ Support\ (X \rightarrow Y) = \frac{Fuzzy\ Support\ (X \rightarrow Y)}{(Fuzzy\ Support(Y))^2}$$

The above definition of Relative Fuzzy Support allows reduction of the support threshold for consequents that have low frequency and increasing the support threshold for consequents that have high frequency. The reduction or increase of support is significant because of the square in the denominator.

The Relative Fuzzy Support differs from RSup (defined in Section 3.3) in two ways:

- The denominator is squared so the reduction or increase of the support is more significant compared to RSup.
- The denominator involves only the support of the consequent. In RSup, the minimum support of all the antecedents and consequents are used. RSup increases the support of a rule if any of its antecedents or consequents is rare. In contrast, Relative Fuzzy Support increases the support only if the consequent is rare.

The Relative Fuzzy Support is well suited for applications in which the user knows the consequents of interest. This is the case in this crime application, as the user is most interested in *Violent Crimes*, *Murders*, *Robberies* and *Assaults* being *High*.

6. EXPERIMENTAL RESULTS

A set of fuzzy membership functions was defined for each of the 40 attributes and each attribute was fuzzified. The fuzzified data constituted the input to Fuzzy Apriori. Fuzzy Apriori was run with confidence 60% and with different supports depending on the data subset. All membership functions for attributes 1-36 were selected as antecedents, and the following membership functions for attributes 37-40 were selected as consequents: *Violent Crimes (Low, Medium, High), Murders (No, Low, Medium, High), Assaults (Low, Medium, High), Robberies (Low, Medium, High).*

Experiments were run on each region data set, producing rules for each region. The data sets for each region were also combined into a single data set comprising all US states. The minimum supports used for each region are as follows: Northeast (0.135%), Southeast (0.714%), Midwest (0.585%), Southwest (1.316%), and West (0.718%).

As the number of communities within each region differs, minimum supports were calculated for each region based on the support of a rule occurring the same number of times within that region. This unifies support across regions for reliable comparison of rule measures, and facilitates discovery of consistent rules across all or subsets of the five US regions.

6.1 All State Rules

All data was used in this experiment, which included 2215 communities located throughout the US. Fuzzy Apriori was run with confidence 60% and support 0.135%. A total of 13,657 rules were generated. Post-pruning using a Relative Fuzzy Support of 1.0 reduced this number to a much more manageable number of 657 rules of interest. This represents a 95.2% reduction in the number of rules.

Figure 5 shows the number of all rules, and the number of rules remaining after pruning, separately for each consequent. The largest numbers of rules were generated for *Assaults (Low)*, *Robberies (Low)* and *Violent Crimes (Medium)*, and post-pruning reduced the number of those rules most significantly.

Figure 6 depicts the average support of all rules, and of rules remaining after pruning, separately for each consequent. The average support of rules remaining after pruning with membership functions *Low* and *No* increased most considerably. This was the goal of performing pruning: to automatically remove the rules of no interest. When the support of a given consequent is high, the rules with low support for that consequent need to be removed, as they are less interesting than rules with higher support.

Figure 7 shows the average relative support of all rules, and of rules remaining after pruning, separately for each consequent. The rules with consequents *Murders (High)* and *Robberies (High)* have the highest relative support. Pruning does not increase the average relative support of this class of rules. The reason being is that it keeps all of those rules (which are rules of interest). The average relative support of rules with membership functions *No*, *Low*, and *Medium* remaining after pruning increased. Pruning does not greatly affect the average confidence of remaining rules.

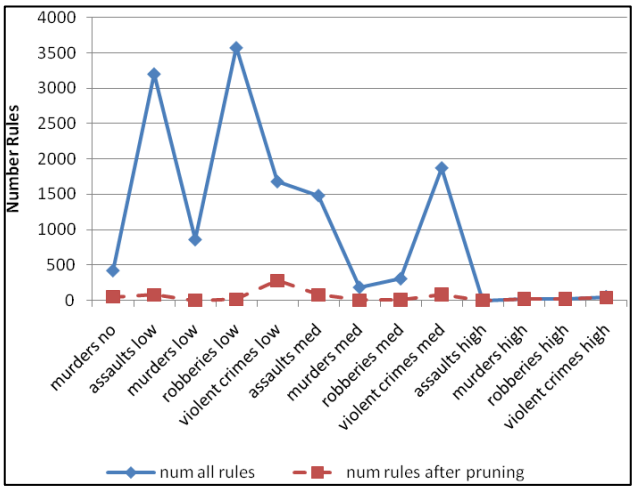

Figure 5. Number of total and pruned rules.

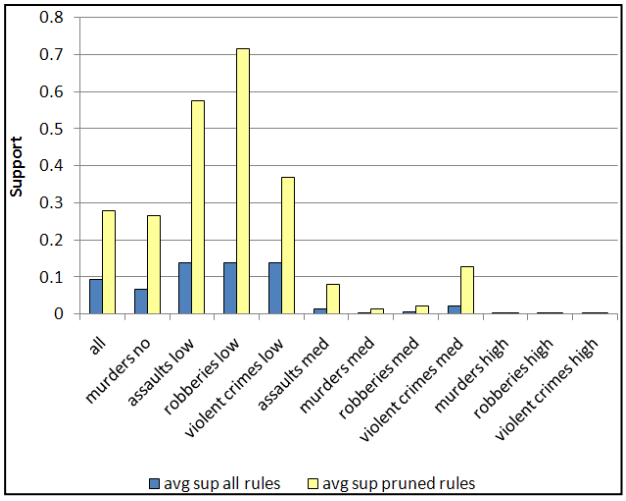

Figure 6. Average support for all and pruned rules.

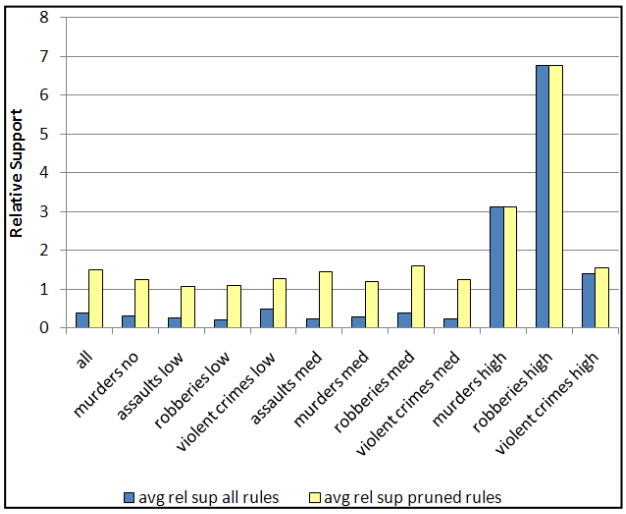

Figure 7. Average relative support for all and pruned rules.

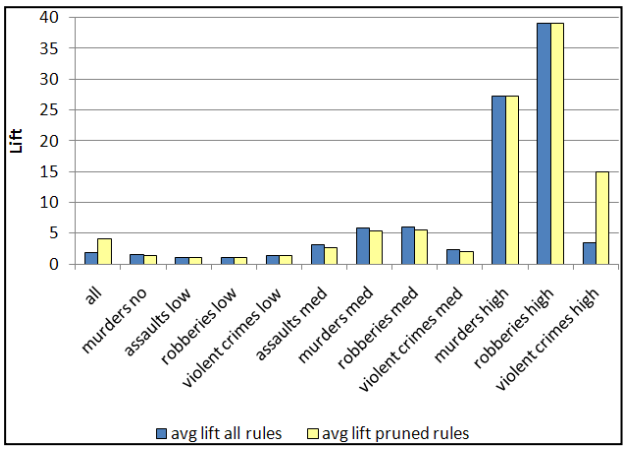

Figure 8. Average lift for all and pruned rules.

Figure 8 presents the average lift of all rules, and of rules remaining after pruning, separately for each consequent. *Murders (High)* and *Robberies (High)* have the highest lift, exceeding several times the average lift of the other consequents. The average lift of rules with consequent *Violent Crimes (High)* increased more than three times, in comparison to all rules for that consequent. The average lift of rules with membership functions *No*, *Low*, and *Medium* remaining after pruning is unchanged. Examples of rules which produce the highest values for each measure follow.

Support:
[*People Speaking No English (Low)*] & [*People in Dense Housing (Low)*] → [*Robberies (Low)*], conf=85.0, lift=1.0, rel sup=1.1, sup=75.3
[*Kids Born to Never Married (Low)*] & [*People in Dense Housing (Low)*] → [*Robberies (Low)*], conf=88.0, lift=1.1, rel sup=1.1, sup=73.9

Relative Support:
[*People in Urban Area (High)*] & [*Kids Born to Never Married (High)*]→ [*Robberies (High)*], conf=63.0, lift=34.7, rel sup=11.9, sup=0.4
[*Race Caucasian (Minority)*] & [*Kids Born to Never Married (High)*] → [*Robberies (High)*], conf=61.0, lift=33.3, rel sup=10.9, sup=0.4

Confidence:
[*Kids Born to Never Married (High)*] & [*People Commute Using Public Transit (High)*] → [*Robberies (High)*], conf=96.0, lift=52.9, rel sup=4.5, sup=0.1
[*Race African American (Minority)*] & [*People Speaking No English (Low)*] → [*Robberies (Low)*], conf=91.0, lift=1.1, rel sup=1.0, sup=65.9

Lift:
[*Kids Born to Never Married (High)*] & [*People Commute Using Public Transit (High)*] → [*Robberies (High)*], conf=96.0, lift=52.9, rel sup=4.5, sup=0.1
[*Houses with Kids Living with Two Parents (Low)*] & [*People Commute Using Public Transit (High)*] → [*Robberies (High)*], conf=86.0, lift=47.4, rel sup=5.6, sup=0.2

It is interesting to note that, in the above sets of rules, the attributes *Kids Born to Never Married* and *Houses with Kids Living with Two Parents* show very prominently. While this is not surprising, it is unexpected that these two attributes show in 6 out of 8 rules with the highest metric values. These discovered rules represent patterns that are of interest to law enforcement officials.

6.2 Rules Consistent Throughout US

Rules generated from each region's data set were analyzed to measure consistency of discovered frequent patterns across all regions and region subsets. Rules represented by multiple regions are considered more general, with generality increasing with the number of matched regions. Therefore, a consistent rule can range from occurring across only two regions to occurring across all regions. Additionally, rules unique to each region were identified to provide more fine-grained regional trends.

Statistics were gathered for each rule appearing in multiple regions, including the minimum, maximum, and average for each primary measure of rule support, confidence, and lift. The region which each overlapping rule's minimum and maximum value correspond to is stored to track the rule's variation across regions. Rules consistent across all regions are those which offer the coarsest level of granularity. Such rules tend to exhibit high support and confidence, as they are prominent in all regional data sets. These rules also can be considered more state-independent, and describe general crime at a nationwide level.

A total of 3188 rules with at least 60% confidence were present in all regions. All rule consequents contain *Assaults (Low)*, *Assaults (Medium)*, *Robberies (Low)*, *Murders (Low)*, or *Violent Crimes (Low)*. None of the rule consequents contain *High* for any of the major crime categories. This observation reinforces that patterns indicative of *High* major crime differ by region, which is a function of the state and community demographics within them.

Table 1 shows the variation of rule measure values for those rules consistent across all US regions. Minimum, maximum, and average values for each rule measure across all consistent rules are reported. On average, rules that are present in each of the five regions from this study exhibit 21% support, 81% confidence, and 1.24 lift. Rules range from rare/novel to highly supported, and exhibit lift values approaching 9.6 in some instances.

Table 1. Rule metrics for rules consistent throughout the US.

Value	Support (%)	Confidence (%)	Lift
Minimum	0.48	60.0	0.581
Maximum	91.843	100.0	9.571
Average	21.291	81.11	1.2447

Examples of rules which produce the highest average values for each measure follow. These specific rules offer insight into crime patterns that are most frequent, probable, and meaningful at a national level. Moreover, these rules provide examples of characteristics which create safe neighborhoods.

Support:
[*People in Dense Housing (Low)*] → [*Robberies (Low)*]
[*People Speaking No English (Low)*] → [*Robberies (Low)*]

The rules with highest average support indicate that robberies are not a significant risk in communities where housing is not dense, English is widely spoken, and public transit systems are not used for daily commutes. This conversely suggests that communities with dense housing (e.g., apartment complexes), a large number of non-English speakers, and heavy use of public transit (e.g., subway) experience higher volumes of robberies.

Confidence:
[*Houses with Retirement Income (High)*] & [*People in Homeless Shelters (None)*] → [*Robberies (Low)*]
[*Race Caucasian (Majority)*] & [*Age 12-29 (High)*] → [*Assaults (Low)*]

The rules with highest average confidence further indicate that robberies occur less in communities with a high number of retired individuals, no homeless, and a high number of traditional family living arrangements. Retirement and family communities are typically low in crime due to better neighborhoods. Assaults are also low in predominantly Caucasian, under-30 communities.

Lift:
[*Race African American (Middle)*] & [*Houses with Public Assistance Income (Medium)*] → [*Assaults (Medium)*]
[*Age 16-24 (Medium)*] & [*Homeless People Counted in Street (Low)*] → [*Murders (Low)*]

The rules with the highest average lift illustrate that murders are lower in communities with a low number of homeless and medium number of people aged 16-24 (e.g., collegiate communities). Violent crimes are also low in under-30 communities which don't receive public assistance income.

6.3 Rules Consistent in Multiple Regions

Rules were also observed to appear in a subset of the regions (i.e., not consistent across all five regions). Those consistent across a subset of any four regions are more general, whereas those consistent across only two regions are considered more regional and state-dependent crime patterns. These rules describe crime at a multi-region level, such as the Northern US and Southern US. Analysis is divided into rule sets occurring in four, three, and two regions. These rule sets are mutually exclusive, such that rules existing in only two regions are not included in the rule sets for three and four regions, and vice versa.

A total of 799 rules exceeding 60% confidence were found to be present in four out of five regions, 1821 rules present in three regions, and 2939 rules present in two regions. Rule consequents from the four-region and three-region sets do not contain *High* for any of the major crime categories. Four-region rules contain up to *Medium* for the *Assaults*, *Robberies*, *Murders*, and *Violent Crimes* consequent variables. Three-region rules contain up to *Medium* for all major crime variables, including *Murders*. The rule set unique to only two regions contains consequents of *High* for *Robberies* and *Violent Crimes*, and up to *Medium* for all other major crime variables. These results suggest that rules found to be consistent across fewer regions offer utility with respect to *High* occurrences of crime. For the two-region set, the regions of Midwest and Northeast contribute the three rules which contain *High* in the consequent.

Table 3 shows the variation of rule measure values for those rules consistent across four, three, and two US regions. As before, minimum, maximum, and average values for the average of each rule measure across all consistent rules are reported.

Table 3. Minimum, maximum, and average rule metrics for rules consistent across subsets of regions.

Four Regions			
Measure	**Support (%)**	**Confidence (%)**	**Lift**
Minimum	0.4777	60.0	0.59
Maximum	77.431	100.0	13.775
Average	11.295	70.837	1.71797
Three Regions			
Measure	**Support (%)**	**Confidence (%)**	**Lift**
Minimum	0.4753	60.0	0.576
Maximum	74.297	100.0	54.696
Average	11.407	66.47	2.0857
Two Regions			
Measure	**Support (%)**	**Confidence (%)**	**Lift**
Minimum	0.4755	60.0	0.586
Maximum	70.789	100.0	69.164
Average	11.111	67.41	2.0558

On average, rules that are present in any four of the five regions from this study exhibit 11% support, 71% confidence, and 1.72 lift. Rules in this set exhibit decreased average support and confidence, as well as increased lift values (exceeding 13.75 in some cases), compared to those consistent across all regions. Three- and two-region rules are similar across all three measures, with increased lift values.

Examples of rules exhibiting the highest average measure values within each overlapping region subset size are presented in the following sections. As before, rules corresponding to the highest average lift value involve higher consequent variable values. The

fewer number of regions in the subset, the more meaningful the rules become at the state level.

Demographic data such as poverty level, income, population density, employment, and living situation for children are directly linked to the occurrence of violent crime. These rules also support the conjecture that high income, family-based communities of educated individuals are less at risk to major crime. Higher income typically translates directly to better security and fewer individuals in the community which would commit such crimes.

6.3.1 Four Regions

Support:

[*Population Density (Low)*] → [*Robberies (Low)*] & [*Assaults (Low)*]

[*People in Owner Occupied Households (Medium)*] → [*Assaults (Low)*]

Confidence:

[*Education Bachelor's or Higher (High)*] & [*People in Homeless Shelters (None)*] → [*Assaults (Low)*]

[*Education Bachelor's or Higher (High)*] & [*Divorced (Low)*] → [*Assaults (Low)*]

Lift:

[*Race African American (Middle)*] & [*Homeless People Counted in Street (Low)*] → [*Robberies (Medium)*]

[*Race Caucasian (Minority)*] & [*Unemployed (Medium)*] → [*Assaults (Medium)*]

6.3.2 Three Regions

Support:

[*People Speaking No English (Low)*] & [*Occupied Housing Units Without Phone (Low)*] → [*Violent Crimes (Low)*]

[*Education Less than 9th Grade (Low)*] & [*People in Dense Housing (Low)*] → [*Violent Crimes (Low)*]

Confidence:

[*Age 65+ (Medium)*] & [*Education Bachelor's or Higher (High)*] → [*Assaults (Low)*]

[*Education Bachelor's or Higher (High)*] & [*Median Gross Rent (High)*] → [*Assaults (Low)*]

Lift:

[*Race African American (Middle)*] & [*Houses with Kids Living with Two Parents (Low)*] → [*Robberies (Medium)*]

[*Divorced (Low)*] & [*Houses with Kids Living with Two Parents (Low)*] → [*Robberies (Medium)*]

6.3.3 Two Regions

Support:

[*People in Dense Housing (Low)*] & [*Homeless People Counted in Street (None)*] → [*Murders (None)*]

[*Race African American (Minority)*] & [*Race Hispanic (Minority)*] → [*Murders (None)*]

Confidence:

[*Per Capita Income (High)*] & [*Homeless People Counted in Street (None)*] → [*Violent Crimes (Low)*]

[*Median Household Income (High)*] → [*Violent Crimes (Low)*]

Lift:

[*Race Caucasian (Minority)*] & [*People in Homeless Shelters (Low)*] → [*Robberies (High)*]

[*Houses with Kids Living with Two Parents (Low)*] & [*Homeless People Counted in Street (Low)*] → [*Violent Crimes (High)*]

6.4 Surprising Rules

In order to find novel rules in the final set of all state rules, a subject matter expert (retired police officer) evaluated the post-pruned set of rules. Several rules were identified as surprising, two of which are discussed below.

[*Employed (High)*] & [*Kids Born to Never Married (High)*] → [*Violent Crimes (High)*], conf=0.67, lift=15.2, rel sup=1.0, sup=0.002

This rule was surprising to the subject matter expert because, according to his experience, when a *High* percentage of people are employed, violent crimes should be *Low*. We identified the following high-support rule that agrees with the expert's intuition:

[*Employed (High*)] → [*Violent Crimes (Low*)], conf=0.67, lift=1.2, rel sup=1.0, sup=0.297

There also exists a rule that specifies the relationship between *Kids Born to Never Married (High)* and *Violent Crimes (High)*:

[*Kids Born to Never Married (High*)] → [*Violent Crimes (High*)], conf=0.58, lift=13, rel sup=2.1, sup=0.004

The rule identified as surprising (containing both antecedents) is a rare rule which shows that, for areas with *High* employment and a *High* percentage of kids born to never married, violent crimes are *High*. This rule has a very high lift (15.2), indicating that the rule premise has a positive effect on the occurrence of the rule consequent. It was very surprising to the expert that the effect of *Kids Born to Never Married* was so pronounced.

The following rule was also identified as surprising by the expert:

[*Age 16-24 (Low)*] & [*Kids Born to Never Married (High)*] → [*Murders (High)*], conf=0.6, lift=13.5, rel sup= 2.1, sup=0.004

This rule was surprising since people between the ages of 16 and 24 are the primary age group for the gang population. Therefore, when the percentage of people from this age group is *Low*, the murders should be *Low* as well. We identified in the rule set the following high-support rule that agrees with the expert's intuition:

[*Age 16-24 (Low*)] → [*Murders (Low*)], conf=0.6, lift=1.1, rel sup=1.3, sup=0.365

There also exists a rule that specifies the relationship between *Kids Born to Never Married (High)* and *Murders (High)*:

[*Kids Born to Never Married (High*)] → [*Murders (High*)], conf=0.52, lift=20.5, rel sup=5.9, sup=0.004

The surprising rule is again a rare rule with a very high lift (13.5), indicating a strong relationship between the antecedents and the consequent. The effect of *Kids Born to Never Married* was yet again very surprising to the subject matter expert.

Many of the other surprising rules contained *Kids Born to Never Married (High)* as an antecedent, and were accompanied with variables which would ordinarily lead to *Violent Crimes* (*Murders* or *Robberies*) *Low*. With the addition of *Kids Born to Never Married (High)*, the consequent instead contained *Violent Crimes* (*Murders* or *Robberies*) *High*. The expert recommended that law enforcement personnel and analysts should further analyze this set of surprising rules and the corresponding underlying data in an attempt to better understand crime patterns and develop more effective approaches to combat crime.

7. CONCLUSIONS

Fuzzy association rule mining has proven useful for this crime application, and has utility for other crime-related data sets. To the knowledge of the authors, this is the first experimental study of applying fuzzy association rule mining to a crime data set. Results presented in this paper suggest that further analysis is required to gain a closer understanding of crime at both the community and national levels.

Crime patterns were discovered which are consistent across all regions, subsets of regions, and all states. The attributes of interest were computed to measure their occurrence per 100K population, so as to remove the element of community and state size during the rule generation process. Rules discovered as part of this study therefore offer utility for use from the national level down to the state and community level.

A novel relative support measure was proposed to prune the set of rules and to extract rare rules from the larger original set. The use of relative support achieves a 95.2% reduction in the final number of rules. These resulting 675 final rules represent a much more manageable number of rules for a crime analyst to investigate. This enables law enforcement personnel to more easily understand the discovered rules by removing the need to sift through uninteresting, obvious rules in order to find interesting and meaningful patterns. In the future, feedback from crime analysts will be utilized to determine if this is a satisfactory number of final rules, or whether additional pruning methods need to be developed to further reduce the number of rules.

The data set used in this study has resolution down to the community (town) level. The generated rules are therefore general to that resolution. Several attributes of the data set contained a significant number of missing values (e.g., *Police Officers Per 100K Population*, *Police Requests Per Officer*, *Officers Assigned to Drug Units*, *Police Operating Budget*). Acquiring accurate data for these attributes will enable the process to produce rules which relate directly to the police force. Rules which relate to the size, budget, and jurisdiction of the police force can then be leveraged. Utilizing data that contains precise locations of crimes or block-level demographics would help produce more meaningful rules for local law enforcement jurisdictions. These higher resolution data could then result in rules that are applicable to certain areas of a city, especially those with widespread crime of various types.

8. ACKNOWLEDGMENTS

The authors wish to thank Dr. Michael Redmond from La Salle University for providing the data set and patiently answering questions about it. The authors also wish to thank Mark Gabriele of Johns Hopkins University Applied Physics Laboratory for his time evaluating rules discovered as part of this study.

9. REFERENCES

[1] 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.

[2] A. Asuncion and D.J. Newman, "UCI Machine Learning Repository," School of Information and Computer Science, University of California, Irvine, CA, 2007, URL: archive.ics.uci.edu/ml/datasets/Communities+and+Crime.

[3] S. Bagui, "An Approach to Mining Crime Patterns," *International Journal of Data Warehousing and Mining*, 2(1), pp. 50-80, March 2006.

[4] R.J. Bayardo, R. Agrawal, and D. Gunopulos, "Constraint-Based Rule Mining in Large, Dense Databases," *Data Mining and Knowledge Discovery*, 4(2/3), pp. 217-240, 2000.

[5] D. Brown, "The Regional Crime Analysis Program (RECAP): A Framework for Mining Data to Catch Criminals," In *Proceedings of the International Conference on Systems, Man, and Cybernetics*, pp. 2848-2853, 1998.

[6] J. de Bruin, T. Cocx, W. Kosters, J. Laros, and J. Kok, "Data Mining Approaches to Criminal Career Analysis," In *Proceedings of the International Conference on Data Mining*, pp. 171-177, 2006, Washington, D.C., IEEE Computer Society Press.

[7] M. Chau, J. Xu, and H. Chen, "Extracting Meaningful Entities from Police Narrative Reports," In *Proceedings of the National Conference on Digital Government Research*, pp. 1-5, 2002.

[8] H. Chen, W. Chung, J. Xu, G. Wang, Y. Qin, and M. Chau, "Crime Data Mining: A General Framework and Some Examples," *Computer*, 37(4), pp. 50-56, April 2004, Los Alamitos, CA, IEEE Computer Society Press.

[9] J. Dembsky, "United States Regions," Online. Available (October 2006): http://www.dembsky.net/regions/.

[10] R. Hauck, H. Atabakhsh, P. Ongvasith, H. Gupta, and H. Chen, "Using COPLINK to Analyze Criminal-Justice Data," *Computer*, 35(3), pp. 30-37, March 2002, Los Alamitos, CA.

[11] C. Ku, A. Iriberri, and G. Leroy, "Crime Information Extraction from Police and Witness Narrative Reports," In *Proceedings of the IEEE International Conference on Technologies for Homeland Security*, pp. 193-198, May 2008, Boston, MA.

[12] 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.

[13] B. Liu, W. Hsu, and Y. Ma, "Mining association rules with multiple minimum supports," In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, pp. 337-341, New York, NY, 1999.

[14] S. Nath, "Crime Pattern Detection Using Data Mining," In *Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology*, pp. 41-44, 2006, Washington, D.C., IEEE Computer Society Press.

[15] V. Ng, S. Chan, D. Lau, and C. Ying, "Incremental Mining for Temporal Association Rules for Crime Pattern Discoveries," In *Proceedings of the Australasian Database Conference*, pp. 123-132, Ballarat, Victoria, Australia, February 2007.

[16] P. Phillips and I. Lee, "Mining Top-*k* and Bottom-*k* Correlative Crime Patterns through Graph Representations," In *Proceedings of the IEEE International Conference on Intelligence and Security Informatics*, pp. 25-30, June 2009, Dallas, TX.

[17] M.A. Redmond, and A. Baveja, "A Data-Driven Software Tool for Enabling Cooperative Information Sharing Among Police Departments," *European Journal of Operational Research*, 141, pp. 660-678, 2002.

[18] R. Srikant and R. Agrawal, "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.

[19] P. Thongtae and S. Srisuk, "An Analysis of Data Mining Applications in Crime Domain," In *Proceedings of the IEEE International Conference on Computer and Information Technology Workshops*, pp. 122-126, 2006, IEEE Computer Society Press.

[20] H. Yun, D. Ha, B. Hwang, and K.H. Ryu, "Mining Association Rules on Significant Rare Data using Relative Support," *Journal of Systems and Software*, 67(3), pp. 181-191, 2003.

[21] L.A. Zadeh, "Fuzzy Sets," *Information and Control*, 8(3), pp. 338–353, 1965.