1-1-2007

Predicting Crime Reporting with Decision Trees and the National Crime Victimization Survey

Gondy Leroy
Claremont Graduate University

Juliette Gutierrez '13
Claremont Graduate University

12-31-2007

Predicting Crime Reporting with Decision Trees and the National Crime Victimization Survey

Juliette Gutierrez
Claremont Graduate University

Gondy Leroy
Claremont Graduate University

Predicting Crime Reporting with Decision Trees and the National Crime Victimization Survey

Juliette Gutierrez
Claremont Graduate University
juliette.gutierrez@cgu.edu

Gondy Leroy
Claremont Graduate University
gondy.leroy@cgu.edu

ABSTRACT

Crime reports are used by law enforcement to find criminals, prevent further violations, identify problems causing crimes and allocate government resources. Unfortunately, many crimes go unreported. This may lead to an incorrect crime picture and suboptimal responses to the existing situation. Our goal is to use a data mining approach to increase understanding of when crime is reported or not. An increased understanding could lead to new, more effective programs to fight crime or changes to existing programs. We use the National Crime Victimization Survey (NCVS) which comprises data collected from 45,000 households about incidents, victims, suspects and if the incident was reported or not. We use decision trees to predict when incidents are reported or not. We compare decision trees that are built based on domain knowledge with those automatically created. For the automatically created trees, we compare three variable selection methods: two filters, Chi-squared and Cramer's V Coefficient, and a forward selection wrapper. We found that the decision trees that are automatically constructed are as accurate as those based on domain knowledge while they show a different picture. We conclude that decision trees lead to several new hypotheses for criminologists while they are automatically constructed and easy to understand which makes them practical and useful.

KEYWORDS

Data Mining, Decision Trees, Filters, Wrappers, Crime Reporting, Law Enforcement, National Crime Victimization Survey

INTRODUCTION

The financial loss due to violent and personal crimes in 2004 was $15.85 billion (Sedgwick 2006) and about 57.5% of these crimes were not reported to the police (BJS 2005). Other costs of unreported crimes include counseling costs, alarms, electronic surveillance equipment and indirect costs such as insurance and taxes (Sedgwick 2006). The National Crime Victimization Survey (NCVS) is used to gather data on injury, theft, damage, the amount of lost work and other characteristics of the incident, victim and suspect. Each year, 45,000 households are interviewed about past incidents where they were the victim. The NCVS has been used to collect data on personal and household victimization since 1973 and it is the main source of data on the characteristics of criminal victimizations (NACJD 2006). It also describes the types of crime not reported to law enforcement and the characteristics of violent offenders (NACJD 2006). The survey classifies each incident as a personal or property crime. Personal crimes include rape, sexual attack, robbery, assault and purse snatching. Property crimes include burglary, theft and vandalism. One of the goals of the NCVS is to understand the quantity and crime types that are not reported to the police (BJS 2005). For example in 2005, 51.3% of personal crimes and 59.3% of property crimes were not reported (BJS 2006a). Table 1 shows the number of personal crimes in 2005 and whether or not they were reported. This research focuses on personal crimes.

Table 1. Number of victimizations, by crime type and whether or not reported (BJS 2005)

Crime Type	Number of Victimizations	Percentage Reported		
		Yes	No	Unknown
Completed Violence	1,658,660	62	37	1
Attempted/Threatened Violence	3,515,060	41	57	2
Rape/Sexual Assault	191,670	38	62	0
Crimes of Violence	5,365,390	47	51	2
Completed robbery	415,320	61	39	1
Attempted robbery	209,530	36	64	0
Robbery	624,850	52	47	1
Aggravated	1,052,260	62	37	1
Simple	3,304,930	42	55	2
Assault	4,357,190	47	51	2
Completed purse snatching	43,550	51	49	0
Attempted purse snatching	3,260	0	100	0
Pocket picking	180,260	32	67	2
Purse snatching/Pocket picking	227,070	35	64	1

According to statistics from the Bureau of Justice Statistics (BJS), the criminal justice system does not act in response to many crime incidents because so many crimes are not discovered or reported to the police (BJS 1967). Our goal is to define new techniques that can help law enforcement evaluate unreported versus reported crime data. Previous research done using the NCVS and descriptive statistics is limited to few variables which show only a limited view of the problem. In contrast, data mining allows for the use of more variables. Moreover, this existing work uses descriptive statistics, such as logistic regression or binomial regression, which require a good understanding of these underlying techniques to interpret the outcome. Decision trees, in contrast, reveal which variables are most important and provide an easy to understand overview for users without a data mining background.

Given the large number of variables available in the survey, selective processing using filters and wrappers is necessary to eliminate useless variables. In comparison to the limited number of variables used by descriptive statistics, we believe that many more variables influence the decision to report or not.

REPORTING CRIME

A BJS Special Report states that on average between 1992 and 2000, only about 31% of rapes and sexual assault were reported to the police (BJS 2003). However, more robberies and assaults are reported: 57% of robberies and 55% of aggravated assaults were reported. A review of current research shows that many variables influence if a crime will be reported, but such individual projects show a partial picture based on only a few variables.

Many used regression analysis and frequency distributions with the NCVS data to look at the different variables related to reporting crime. Based on this research, three aspects of an incident seem most influential in the decision to report. The first is the relationship of the offender to the victim. Based on data available from the BJS, violent crime committed by a stranger is more likely to be reported than violent crime committed by non-strangers (BJS 2003). Research by Felson, Messner, Hoskin and Deane (Felson et al. 2002) used frequency distributions and logistic regression and found that people are just as likely to report domestic assaults, which is one type of violent crime, as they are to report assaults by other people they know. Earlier research by Felson (Felson et al. 1999) included third party reporting data and used the NCVS data with frequency distributions and regression analyses. They found that the relationship between the offender and the victim affects third party but not victim reporting. These conclusions are consistent in that it shows no effect on the relationship to victim reporting, but it goes beyond the earlier work by looking at third party reporting. Third parties are people other than the victim such as members of the household or other witnesses. There were two reasons for this effect: 1) third parties are hesitant to report minor assaults and 2) third parties are not likely to witness an assault between people in ongoing personal relationships.

The age of the victim is the second decisive factor in reporting behavior. Over the years, the BJS reported that older victims, age 65 and older, are more likely than younger victims to report violence and personal theft to the police (BJS 2000; BJS 2003). Others have found similar results: most crimes against juveniles are not reported (Finkelhor et al. 2001; Finkelhor et al. 2003). This underreporting is serious since we know that persons age 12 to 19 are the most frequent victims of crime in

the United States (BJS 2006b). Watkins showed with binomial and regression analyses that this inequality in reporting between juveniles and adults could not be accounted for by individual, incident or situational variables (Watkins 2005).

The third factor in reporting crimes is experience with past victimizations. Xie, Pogarsky, et al. (Xie et al. 2006) used NCVS data from 1998 – 2000 and concluded that an individual was more likely to report a subsequent victimization when prior incidents had been investigated by the police. However, this was only true when the victim reported the incident, not when it was reported by someone else. They also found that whether there was an arrest as a result of past incidents did not effect whether the individual reported a subsequent victimization to the police and also that whether an arrest was made as a result of past incidents with a member of the victim's household did not effect whether the individual reported a subsequent victimization.

DECISION TREES

Overview

The goal of data mining is to find patterns as a tool for helping to explain the data and make predictions about it. Decision trees are one approach for such predictions. A decision tree is a classifier in the form of a tree where each node in the tree is either a leaf node or a decision node. In Figure 1, we show a partial decision tree. Leaf nodes (rectangles) provide a final classification (REPORT_NO or REPORT_YES) to all data instances that arrive at that node. A decision node (oval) is where a particular variable is tested. In this example, when the value of VICTIM_AGE is greater than 23 the classification will be REPORT_YES meaning that the crime was reported. When the value of VICTIM_AGE is less than or equal to 23 the classification will be REPORT_NO.

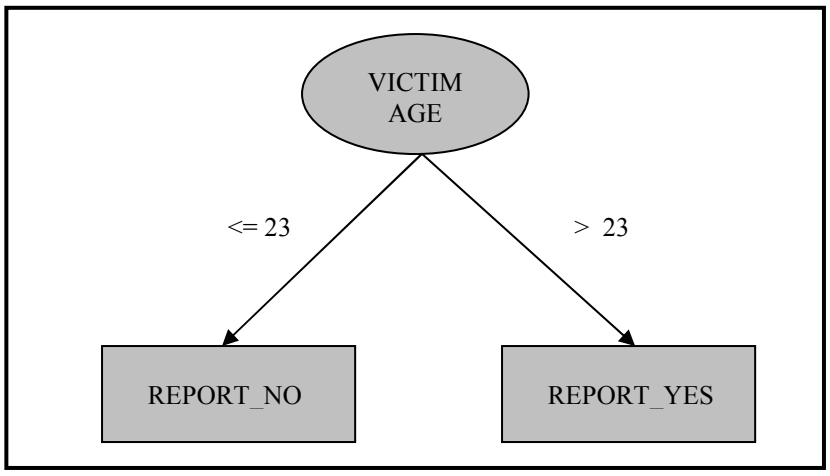

Figure 1. Partial Decision Tree

Justification for using Decision Trees

Decision trees have the advantage of being easy to interpret which makes them practical and useful. Quinlan (Quinlan 1986) states that despite the fact that a decision tree is a simple representation of knowledge, it can still be useful for generating practical solutions to complicated problems. Decision trees make no assumptions about the data which means that the induction algorithms do not rely on any other information other than that which exists in the data. They can handle both categorical and continuous variables. The NCVS data contains both categorical and continuous variables. Furthermore, decision trees are unaffected by worthless variables since the algorithms are developed to learn which variables are the best to use through the concept of information gain or entropy. As a consequence they indirectly reveal importance of variables: the most important variables are selected first when building a tree.

However, decision trees do have disadvantages. Their main weakness is that they can be unstable which means that variations in the data can cause the induction algorithms to create different decision trees on seemingly similar data. Decision trees also do not work well on small datasets since the induction algorithms learn from the data. In general, a larger dataset helps the algorithms to make better decisions about variables. Larger test samples also allow for more accurate error estimates (Witten et al. 2005). Finally, decision trees optimize locally at each level of the tree and so are not good tools to show interactions.

The ID3 algorithm, developed by Quinlan (Quinlan 1986), uses a divide-and-conquer approach based on entropy to create decision trees. Entropy is a measure of the amount of uncertainty. The algorithms for creating decision trees calculate the information gain for each variable – and so reduce uncertainty – and select the next decision node as that variable that leads to the most information gain or entropy reduction. Several improvements to ID3 which deal with noisy and incomplete data (Quinlan 1986) produced a practical and influential algorithm for creating decision trees called C4.5 (Witten et al. 2005). This is helpful because real-world data, such as that from the NCVS, is unlikely to be entirely accurate and is therefore noisy.

Variable Elimination

The primary goal of this phase of research was to pre-process the data to eliminate variables. Many datasets contain too many variables for end users to evaluate. According to Witten and Frank, the best way to select relevant variables is manually (Witten et al. 2005). Manual variable selection or exclusion should be based on an understanding of the data, the learning problem and what the variables really mean (Witten et al. 2005). Two additional fundamentally different approaches to reducing the number of variables are filters and wrappers (Witten et al. 2005).

Filters. Filters process the data independently of the learning algorithm. Variable ranking is a filter process that scores each variable according to a method such as Chi-squared or Cramer's V Coefficient, ranks the output and selects the best, i.e. highest rank , variables (Wang et al. 2005).

Chi-squared can be used as a filter-based approach to eliminating variables. It is used widely in research because of the large amount of data which is collected at a nominal level of measurement (Sproull 1995). Chi-squared tests the null hypothesis that the variables are independent of each other and have no association (Sproull 1995). It can be useful to find the variables that have a significant association in regard to the target variable (Wang et al. 2005). Chi-squared can be calculated by (Becher et al. 2000)

$$X^2(X,Y) = N_{jq}(P_{jq}-P_jP_q)^2 / (P_jP_q)$$

Where X is the source variable with values j ranging from 1 to J and Y is the target variable with values q ranging from 1 to Q. P represents the distribution. N is the total number of data instances. The outcome indicates whether there is a relationship between the variables.

Cramer's V Coefficient is a measure of the strength of the relationship. It is a well-known normalization of Chi-squared (Becher et al. 2000). It is similar to Chi-squared in that it measures the level of association between two variables. The values of the Cramer's V Coefficient range from 0 to 1. The higher the V Coefficient of a variable indicates more relevancy to the target variable (Wang et al. 2005). The Cramer's V Coefficient can be calculated by (Becher et al. 2000)

$$V(X,Y) = X^2(X,Y) / (N \min(Q,J) - 1)$$

Wrappers. Wrappers process the data using the evaluation function dependent on the learning algorithm (Wang et al. 2005). In order to implement a wrapper it is necessary to know how to search all possible variable subsets, how to assess performance, when to halt the wrapper process and which learning algorithm to use (Guyon et al. 2003).

Forward Selection is a wrapper-based approach to eliminating variables. It starts without any variables but adds variables, one at a time into a subset based on evaluation criteria from the learning algorithm, then retests the subset of the variables with the learning algorithm (Witten et al. 2005). We used a forward selection process implemented using Weka, an open source Java data mining application (http://www.cs.waikato.ac.nz/ml/weka/). The J48 tree classifier is the C4.5 implementation available in Weka.

METHODOLOGY: PREDICTING CRIME REPORTING WITH DECISION TREES

We compare decision trees based on variables collected using 4 approaches: All relevant variables (baseline), filtering (Chi-squared and Cramer's V Coefficient), wrappers and variables selected based on domain knowledge.

Dataset

The NCVS has a large number of variables on the frequency and characteristics of criminal victimizations (NACJD 2006). In its present state, there are 828 variables. We included all 28,069 rows which includes personal crimes from 1992 to 2004 for the training dataset in order to eliminate any bias based on the selection of training and testing data. The resulting models will be tested with 2005 data once it becomes available. The data and the codebook are available to download from http://www.icpsr.umich.edu/NACJD/NCVS/. We used dataset DS3: 1992-2004 Incident-Level Concatenated File and used all values for the selected variables. No translations were done on any of the variables except for the classification variable (V4399) which was translated to 1 = YES, 2 = NO and all remaining values were translated to MAYBE.

Variable Selection Methods

Baseline Variables. For the baseline dataset, we include all variables that may contribute to the decision to report crime or not. We excluded three sets of variables whose information cannot contribute to the victim's decision. The first set consists of linkage and identification variables that are part of the management of the survey data. The second set of variables that are excluded are those that have a direct correlation to V4399 (Reported to Police?) or directly depend on this outcome. The third set of variables were those without a relation to the incident but informative about the NCVS interview. This process was subjective and we will collaborate with a detective in the near future to refine this process.

Filters. We created the Chi-squared filter using an Oracle database with a PL/SQL procedure to calculate the Chi-squared value for each variable in the baseline dataset. The Chi-squared values were written to a table then sorted by the Chi-squared value. The variables with the highest Chi-squared values were selected. We also test the Cramer's V Coefficient filter for comparison to Chi-squared. We wrote an additional PL/SQL procedure to calculate the Cramer's V Coefficient value for each variable in the baseline dataset. The Cramer's V Coefficient values were also written to a table then sorted by the Cramer's V Coefficient. The variables with the highest Cramer's V Coefficient values were selected.

Wrappers. A forward wrapper was implemented to compare against the filter methods. The wrapper was implemented by running each variable in the baseline dataset with V4399 (Reported to Police?) as the classification variable to be predicted. A decision tree was created for each variable and the correctly classified percentage was maintained for each. The variable which created the decision tree with the highest accuracy was selected into a subset of variables to use for the next iteration. The subset of chosen variables was run with all remaining variables in the baseline dataset with V4399 (Reported to Police?) as the variable used for prediction. A decision tree was created with the subset of chosen variables and each remaining variable and the correctly classified percentage was maintained for each variable. The variable with the highest accuracy or correct classifications was selected into the subset of variables to use for the next iteration. This process was iterated until the accuracy of the decision tree started to decline.

The Weka input file was created with data generated from SQL queries that were run against the Oracle database and the required header portion of the file in the ARFF format. Weka (http://www.cs.waikato.ac.nz/ml/weka/) was used to create decision trees and the results were maintained both in Oracle and in Excel. Every classifier run done with Weka used a 10-fold cross validation. The data was then sorted by the accuracy of the resulting decision trees and the variables with the highest accuracy were selected.

RESULTS

In 2005, 57.4% of all crimes were not reported to the police, 41.3% of all crimes were reported and the status of 1.3% of all crimes was not known and not available (BJS 2006a).

Baseline Variables

We excluded 232 variables from the dataset; 32 linkage variables, 77 highly correlated variables and 123 non-incident related variables. The result of this selection process created the baseline dataset for the remainder of the research.

Decision Trees with Chi-squared Variable Selection (Filter 1)

The variables in Table 2 were selected using Chi-squared on the baseline dataset. These variables had the highest ranked Chi-squared values. We used the top 30 variables with the Weka J48 classifier. Thirty variables were used because of memory and processing limits. The resulting decision tree contained twenty variables showing that including more variables would not have changed the outcome since these already had a lower Chi-squared value.

V4130, Medical Care: Home, Neighbor's, Friends, is the root node selected with this subset of variables. This means it has the highest information gain, so it is most useful at this stage to predict if crime will be reported or not. V4205, Number of Other's Harmed or Robbed, and V4481, Is the Business Incorporated?, are on the lower levels of the tree. The resulting decision tree has an accuracy of 66.0257% with 1,075 leaves. The total size of the tree is 2,149 nodes.

Table 2. Chi-squared

Description	Variable	Chi-squared
Help From Victim's Agencies?	V4467	8372.760
Was Victim Agency Government or Private?	V4468	6768.317
Incident Occur at Work Site?	V4484	3655.600
Job Located in City/Suburb/Rural Area?	V4483	2914.522
Is the Business Incorporated?	V4481	2243.496
Type of Industry at Time of Incident?	V4482	2032.310
Usually Work Days or Nights?	V4485	1879.695
Type of Crime Code	V4529	1802.063
Where Did Incident Happen?	V4024	1639.409
Total Amount of Medical Expenses?	V4140	1605.138
Any Others Harmed or Robbed?	V4203	1445.340
Number of Others Harmed or Robbed?	V4205	1435.278
Activity at Time of Incident?	V4478	1348.205
Medical Care: Emergency Room	V4133	1265.782
Is the Business Incorporated?	V4481A	1236.723
One or More Than One Offender?	V4234	1189.496
Medical Care: Home, Neighbor's, Friends	V4130	1149.601
Current Job?	V4485A	1139.623
Attempt/Threat: Weapon Present?	V4084	1136.713
How Offender Threatened or Tried to Attack	V4077	1136.544

Decision Trees with Cramer's V Variable Selection (Filter 2)

The variables in Table 3 were selected using Cramer's V Coefficient on the baseline dataset. These variables had the highest ranked Cramer's V values. As with the previous filter and so we can compare decision trees, we selected the top 30 variables. The resulting decision tree also contained 20 variables.

V4136, Residual: Medical Care, is the root node selected with this subset of variables which means it was the most useful at this level to predict if crime would be reported or not. V4478, How Offender Tried to Attack, and V4024, Where Did Incident Happen?, are on the lower levels of the tree since they have the smallest information gain. The resulting decision tree has an accuracy of 65.9424% with 1,003 leaves. The total size of the tree is 2,005 nodes.

Table 3. Cramer's V

Description	Variable	Cramer's V
Help From Victim's Agencies?	V4467	0.3179
Was Victim Agency Government or Private?	V4468	0.2858
Incident Occur at Work Site?	V4484	0.2101
Job Located in City/Suburb/Rural Area?	V4483	0.1876
Is the Business Incorporated?	V4481	0.1646
Type of Industry at Time of Incident?	V4482	0.1566
Anything Damaged?	V4387	0.1530
Usually Work Days or Nights?	V4485	0.1506
Type of Crime Code	V4529	0.1475
Where Did Incident Happen?	V4024	0.1407
Total Amount of Medical Expenses?	V4140	0.1392
How Other's Action Helped?	V4186	0.1367
How Attacked?	V4093	0.1335

Residue: How Other's Helped	V4193	0.1332
Any Others Harmed or Robbed?	V4203	0.1321
Number of Others Hurt or Robbed?	V4205	0.1316
Residual: Medical Care	V4136	0.1311
How Other's Actions Worsened Situation	V4195	0.1302
Activity at Time of Incident	V4478	0.1276
How Other's Action Hurt	V4202	0.1272

Decision Trees with Forward Selection of Variables

All variables in the baseline dataset were processed with the Weka J48 classifier. The 200 variables that made up the decision trees with the highest accuracy were selected to continue with the Forward Selection process due to the manual nature of this processing. In the future, we will automate this process.

The variables in Table 4 were selected using this Forward Selection wrapper. These variables had the highest ranked decision tree accuracy values. V4498, Total Number Days Lost?, is the root node selected with this subset of variables which means that it was most useful in predicting crime reporting at this level. V4478, How Offender Tried to Attack, and V3014, Age (Allocated), are on the lower decision nodes of the tree since they have the smallest information gain. The resulting decision tree has an accuracy of 68.8466 % with 678 leaves. The total size of the tree is 1,355 nodes.

Table 4. Forward Selection

Description	Variable
Type of Crime Code	V4529
Activity at Time of Incident	V4478
Single Offender: How Did Respondent Know Offender?	V4245
Check B: Attack, Threat, Theft	V3052
Which Best Describes Your Job?	V3074
How Many Times Incident Occurred?	V4016
Total Number Days Lost?	V4498
Help From Victims Agencies?	V4467
Age (Allocated)	V3014
Anything Damaged?	V4387
Covered By Medical Insurance	V4139
Number of Household Members Harmed/Robbed	V4207
Stolen, Attack, Threat: Offender Known	V3044
Thought Crime But Didn't Call Police	V3054
Something Taken? (Allocated)	V4288
Number of Others Harmed or Robbed	V4205

Decision Trees with Domain Knowledge Variable Selection

The BJS research discussed the relationship of the offender to the victim and violent crime (BJS 2003). The research done by Felson and other authors also discuss violent crime and report of assaults by people they know (Felson 2002; Felson et al. 1999). The age of the victim was discussed by BJS (BJS 2000; BJS 2003), Finkelhor (Finkelhor et al. 2001; Finkelhor et al. 2003) and Watkins (Watkins 2005). The research done by Xie, Pogarsky, et. al. (Xie et al. 2006) discussed how victims were more likely to report subsequent victimizations.

To compare our approach with the existing work, we selected variables considered relevant in previous literature. This was a subjective process as some research did not list the actual variables from the NCVS survey. We used 5 variables to build the decision tree. After building the tree, variable V4529, Type of Crime Code, became the root node selected. V3014, Age (Allocated), and V4245, Single Offender: How Did Respondent Know Offender?, are on the lower decision nodes of the tree since they have the smallest information gain. The resulting decision tree has an accuracy of 64.4867% with 126 leaves. The total size of the tree is 251 nodes. According to these decision trees the crime type is the predictor of whether or not victims will report a criminal victimization.

Table 5. Domain Knowledge Variable Selection

Description	Variable
Type of Crime Code	V4529
Age (Allocated)	V3014
Single Offender: How Did Respondent Know Offender?	V4245
How Many Times Incident Occurred?	V4016

Summary of Variable Selection Methods

Table 6 summarizes the results. Each of these methods excludes variables that are used to manage the NCVS dataset. The Forward Selection wrapper process yields the smallest and most accurate decision tree. This tree has about 32% fewer leaves and is smaller overall while being the most accurate. While the Forward Selection process is slower due to the multiple iterations, it does produce a better tree. The tree based on domain knowledge is much smaller due to the smaller number of variables that were used as input. It also has a slightly lower accuracy. The trees created with the Chi-squared and Cramer's V Coefficient filters are the most similar trees with their number of leaves and the size and accuracy of the trees being very comparable.

If we were to assign the outcome "not reported" to all incidents, our prediction would be 57.4% accurate (baseline). Our decision trees could predict reporting of crime with about 10% more accuracy than the baseline.

Table 6. Summary of Variable Selection Methods

Method	Leaves	Size	Overall Accuracy	Root Node	Description	Accuracy YES	Accuracy NO	Accuracy MAYBE
Chi-squared	1,075	2,149	66%	V4130	Medical Care: Home, Neighbor's, Friends	7,748/12,838 = 66%	10,473/14,427 = 73%	12/350 = 3%
Cramer's V	1,003	2,005	66%	V4136	Residue: Medical Care Site	7,801/12,838 = 61%	10,400/14,427 = 72%	9/350 = 3%
Forward Selection	678	1,355	69%	V4498	Total Number Days Lost	8,242/12,838 = 64%	10,755/14,427 = 75%	15/350 = 4%
Domain Knowledge	126	251	64%	V4529	Type of Crime Code	7,948/12,838 = 62%	9,860/14,427 = 68%	0/350 = 0%

DISCUSSION

The variables that we discovered are different from previous research endeavors. The Forward Selection wrapper method selected V4498, Total Number of Days Lost?, as the root. The two filter methods chose, V4130, Medical Care: Home, Neighbor's, Friends, and V4136, Residual: Medical Care. Both variables are possible answers to survey question V4128, Where was Medical Care received. Since both filter methods are related, the similarity of the two resulting decision trees is not a surprise. However, the wrapper approach provides a different, but equally accurate view of the data compared to the filter approach and compared to the domain knowledge approach. These results are especially interesting as none of the deciding variables were, to our knowledge, used in any prior research to predict crime reporting. Data mining has raised these variables as areas for further research. The area of medical care needs to be carefully researched further since some types of violent crimes require mandatory reporting to the police by healthcare professionals. All of these newly found predictor variables require further research.

Although this first step shows the strength of using decision trees, a shortcoming of our approach is the complexity of these decision trees due to the increased number of variables. With many variables, the size and complexity of the decision tree can easily grow to become unusable and it becomes more difficult for a human expert to interpret.

CONCLUSIONS AND FURTHER DIRECTIONS

We used data mining techniques, specifically decision trees, instead of traditional statistics to predict when crimes are reported. We compared decision trees that were created with variables selected with two filters, Chi-squared and Cramer's V

Coefficient, with a forward selection wrapper and with a decision tree based on current domain knowledge. We concluded that using decision trees lead to the discovery of several new variables to research further.

Traditional approaches to analyzing the NCVS data have been by descriptive statistics. The traditional approach is limited to a few independent and dependent variables. Our research differs from the traditional approach by using decision trees to analyze the NCVS data. Decision trees allow multiple variables to be brought into the analysis. By definition, the critical variables are towards the root of the tree and can be easily ascertained. By definition, the induction algorithms discover the variables that add the most information gain.

In the future, we will investigate data transformation methods to increase accuracy and reduce the complexity and size of the tree. Preliminary research shows that the addition of data transformations does decrease the complexity and size of the tree and increase the accuracy of the tree significantly. We will also look at other filters and wrappers. Backward elimination is another wrapper method where all variables are brought into the algorithm and then eliminated one at a time. Finally, we will integrate such decision trees as tools into a decision support system for law enforcement.

ACKNOWLEDGEMENTS

The authors would like to acknowledge LAPD Detective Dave McGowan for feedback on initial ideas.

The authors also acknowledge that this research was made possible with support from a Fletcher Jones Grant.

REFERENCES

Becher, J.D., Berkhin, P., and Freeman, E. "Automating Exploratory Data Analysis for Efficient Data Mining," Knowledge Discovery and Data Mining, Boston, MA USA, 2000.

BJS "What is the sequence of events in the criminal justice system?," U.S. Department of Justice, Washington, DC, 1967.

BJS "Crimes against Persons Age 65 or Older, 1992 - 1997," U.S. Department of Justice, Washington, DC, 2000.

BJS "Reporting Crime to the Police, 1992 -2000," BJS (ed.), U.S. Department of Justice, Washington, DC, 2003.

BJS "Percent distribution of victimization, by type of crime and whether of not reported to the police," Bureau of Justice Statistics, Washington D.C., 2005, pp. Personal and property crimes, 2005.

BJS "Percent distribution of victimizations, by type of crime and whether or not reported to the police," U.S. Department of Justice, Washington D.C., 2006a.

BJS "Teens and young adults experience the highest rates of violent crime," U.S. Department of Justice, Washington, DC, 2006b.

Felson, R., Messner, S.F., Hoskin, A.W., and Deane, G. "Reasons for Reporting and Not Reporting Domestic Violence to the Police," *Criminology* (40:3), August 2002, pp 617-650.

Felson, R.B. *Violence and Gender Re-examined* The American Psychological Association, Washington D.C., 2002.

Felson, R.B., Messner, S.F., and Hoskin, A. "The Victim-Offender Relationship And Calling The Police in Assaults," *Criminology* (37:4) 1999, pp 931-947.

Finkelhor, D., and Ormrod, R.K. "Factors in the Underreporting of Crimes Against Juveniles," *Child Maltreatment* (6:3) 2001, pp 219-229.

Finkelhor, D., and Wolak, J. "Reporting Assaults Against Juveniles to the Police: Barriers and Catalysts," *Journal of Interpersonal Violence* (18:2), February 2003, pp 103-128.

Guyon, I., and Elisseeff, A. "An Introduction to Variable and Feature Selection," *Journal of Machine Learning* (3) 2003, pp 1157-1182.

NACJD "National Crime Victimization Survey Resource Guide," NACJD, 2006.

Quinlan, J.R. "Induction of Decision Trees," *Machine Learning* (1:1) 1986, pp 81 - 106.

Sedgwick, J. "The Cost of Crime: Understanding the Financial and Human Impact of Criminal Activity," Bureau of Justice Statistics, Washington D.C., 2006.

Sproull, N.L. *Handbook of Research Methods A Guide for Practitioners and Students in the Social Sciences*, (Second ed.) The Scarecrow Press, Inc., Lanham, MD, 1995, p. 430.

Wang, H., Parish, A., Smith, R.K., and Vrbsky, S. "Variable Selection and Ranking for Analyzing Automobile Traffic Accident Data," 2005 ACM Symposium on Applied Computing, ACM, 2005.

Watkins, A.M. "Examining the Disparity Between Juvenile and Adult Victims in Notifying the Police: A Study of Mediating Variables," *Journal of Research in Crime and Delinquency* (42:3), August, 2005, pp 333 - 353.

Witten, I.H., and Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques*, (Second ed.) Morgan Kaufmann Publishers, San Francisco, CA, 2005.

Xie, M., Pogarsky, G., Lynch, J.P., and McDowall, D. "Prior Police Contact and Subsequent Victim Reporting: Results from the NCVS," *Justice Quarterly* (23:4), December, 2006, pp 481 - 501.