Discovering Criminal Behavior Indicators in a Survey of Substance Use, Mental Illness, and Crime Data

Peter Sigur
Computer Science Graduate Student
Hood College
Frederick, MD, USA
ps18@hood.edu

***Abstract*—This paper details the use of decision tree classification to predict criminal behavior from an extensive set of attributes concerning the mental health, substance use, and criminal activity of the respondents. The goal of this data mining project is to determine the most important behavior or demographic indicators within the data that might classify potentially dangerous individuals, with the hope that such information could be used to provide targeted preventative treatment or counseling to at-risk groups.**

Keywords—data mining; decision tree classification; Hunt's algorithm; ID3; mental health; drug use; substance abuse; criminal activity; treatment for mental illness; addiction treatment

I. Introduction

Research and statistical analysis have already determined a link between mental illness (including drug abuse) and crime [1], however there is now enough data available to further explore the intricacies of that link within subsets of the data. By looking at specific kinds of mental illness, use of certain legal or illegal drugs, and treatment received for any of the above, we could potentially find patterns that point to particular at-risk populations. Providing preventative care or counseling to such groups would, we hope, have a noticeable effect in reducing crime.

Because of the vast amount of data available, data mining techniques offer a viable way to analyze the information and return potentially interesting patterns. These scalability and high dimensionality challenges are indeed some of the driving factors behind data mining [2]. As part of the larger knowledge discovery in databases (KDD) process, data mining would allow us to find unexpected patterns in the data, which could be passed along to domain experts for validation. While there is certainly the risk of finding meaningless or invalid patterns (at the level where data mining devolves into data dredging [3]), it is those unexpected patterns that could be the most valuable, as they diverge from prior beliefs [4].

The data gathered by the National Household Survey on Drug Use and Health (NSDUH) presents a nationally representative view of mental health, substance use, and criminal activity of respondents in the United States, ages twelve and older. This data set provides an ideal source from which we can build classification structures to further investigate.

II. Background

The 2004 National Household Survey on Drug Use and Health contains survey data from 55,602 respondents. This data is broken down in to 3,011 variables or attributes, 2,690 of which come directly from the survey, while the remaining 321 were created by the principal investigator, M. Thompson [1]. The data was collected from respondents of various ages from all fifty states. The attributes in this data set include demographic information, drug use and dependence indicators, indices of mental illnesses, history of criminal activity, and various types of treatment or rehabilitation related to the attributes above.

While this survey data was provided willingly by respondents, this type of information is usually protected (in the case of health information) or at least not readily provided (in the case of criminal activity, illegal drug use, or arrest history). Concerns about privacy and the ethical responsibility of data scientists while mining sensitive data have already appeared in literature and conference proceedings, and will undoubtedly continue to grow as data is collected in larger quantities and in more locations. To address these concerns, research has been done by Clifton, Jin, and Kantarcioğlu to find a way to objectively measure the acceptable tradeoff between privacy and the benefit of the data mining results [5]. In terms of ethical responsibility, efforts have been made by Seltzer and Pedreschi, Ruggieri, and Turini to provide guidelines for avoiding targeting vulnerable individuals or population subgroups [6] or reinforcing discriminatory prejudices [7] based on the findings of data mining techniques. While the data set used here is publicly available, caution must be exercised when attempting to apply any resulting classification techniques built on the data set against members of the public.

Even though mental illness, substance use, and crime are not necessarily outliers or fringe cases of the general population, they certainly do not make up the most frequent patterns. When looking at medical data to identify risk, Li et al. encountered a similar problem, namely that the interesting risk patterns they were looking for were not visible using usual data mining objectives of frequency and support [8]. This required the use of an epidemiological metric, relative risk, to measure interestingness of the patterns found while mining the data.

Decision tree classification is another method of data mining which has been used to classify criminal behavior by constructing a profile of potential criminals based on existing data [9][10], and in the mental health domain, in conjunction with biomedical data, to classify genetic and environmental factors that can cause mental illnesses [11]. This method has the added benefit of producing results that are easily understood without knowledge of the underlying method – following a classification tree from the root on down is a simple, visual task.

III. Approach and Methodology

To mine this data set, I will be using a decision tree classification approach, to search specifically for the key indicators of crime (as evidenced by prior arrests, or time spent in jail or a detention center) in the data. Because of the size of the data set, specifically its high dimensionality and the "curse of dimensionality" that that entails [2], only a limited subset of the available attributes was mined. When possible, attributes were combined or dropped to simplify the data set, however, with 3,011 attributes available, some interesting information was undoubtedly left buried.

Through the use of decision tree classification, I hoped to both gain new insight from this particular data set, as well as produce a simple, useful classification system for new or incomplete data. A modified version of the ID3 algorithm, based on Hunt's algorithm, was used to build the classification tree. Details about these algorithms can be found in the following section.

In order to prepare the data for classification, multiple pre-processing steps were required. The initial data set was received in a single tab-delimited file, which was converted to multiple comma-delimited files, allowing all the data to be read into an SQL database. After inspecting the data codebook (containing descriptions of the attributes, as well as their potential values, and statistics concerning response rates), a set of potential attributes was selected based on high-level representation of the respondents. For instance, the attribute BOOKED (indicating an arrest) was selected, rather than one of the attributes concerning arrest for a specific class of crime. This decision was made to ensure that the major sections of the data were represented, without running into the problems inherent in the curse of dimensionality. By loading the data into an SQL database, I was able to accelerate the process of filtering the selected attributes, combining similar response values (e.g. user-selected 'Yes' versus auto-assigned 'Yes', or the different responses combined in 'missing data'), merging attributes to use as the class (jailed and/or arrested), and removing those records that were missing class data.

The final data set used for this project is described in Table 1. After pre-processing, the number of records dropped to 55,391 after removing those missing class data, and the number of attributes was reduced from 3,011 to 11. While this was obviously a dramatic reduction from the initial data set, I believe it offers a good starting point for finding a high-level classifier that could be used to explore additional subsets of the data.

TABLE I. List of Attributes After Pre-Processing

Attribute	Possible Values	Value Definition
Age	1 2 3 4 5 6	12-17 years old 18-25 26-35 36-47 48-60 61+
Gender	0 1	Female Male
Total Family Income	1 2 3 4	<$20,000 $20,000-$49,999 $50,000-$74,999 >$75,000
Education Level	0 1 2 3 4	12-17 years old < High School High School Some College College Graduate
Any Tobacco Use	0 1 -1	No Yes Missing Data
Any Alcohol Use	0 1 -1	No Yes Missing Data
Any Illicit Drug Use	0 1 -1	No Yes Missing Data
Any Depressive Episodes	0 1 -1	No Yes Missing Data
Any Alcohol or Drug Treatment	0 1 -1	No Yes Missing Data
Any Mental Health Treatment	0 1 -1	No Yes Missing Data
Ever Arrested and/or Jailed	0 1	No Yes

This approach was not without issues. Primary among them was the size of the data set in terms of number of attributes. While it would have been ideal to be able to put all of the data through a classification method, the curse of dimensionality limits that possibility. In order to reduce the number of dimensions, they had to be understood, which involved time spent investigating the 1,500 page codebook. While this data set was already quite clean, especially with its use of numeric codes for all responses, it did require some additional pre-processing in the form of merging responses and attributes.

IV. Algorithms and Implementation

In order to build the decision tree, I used a modified version of the ID3 algorithm, based on Hunt's algorithm:

1) *If D_t is empty, t is a leaf labeled by the default class*

2) *If D_t contains only records of the same class, t is a leaf node labeled by that class*

3) If D_t contains records of multiple classes, use an attribute test to split the set into subsets, go back to step 1 for each

The attribute test mentioned in step three was completed with the use of information gain determined by entropy, as defined by the following:

$$GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right)$$

$$Entropy(t) = -\sum_{j} p(j \mid t) \log p(j \mid t)$$

In this way, the tree was built in a greedy fashion by always selecting the attribute that offered the greatest decrease in overall entropy. This kept the tree relatively short, and made it easier to determine which attributes are the stronger indicators for the class.

The Python programming language was used to write all of the code for this project. As a quick, easy to read, scripting-style language, Python is a popular option for data mining coding. While I was entirely inexperienced with Python prior to beginning this project, I do believe it was a good option for keeping the methods simple to read and debug.

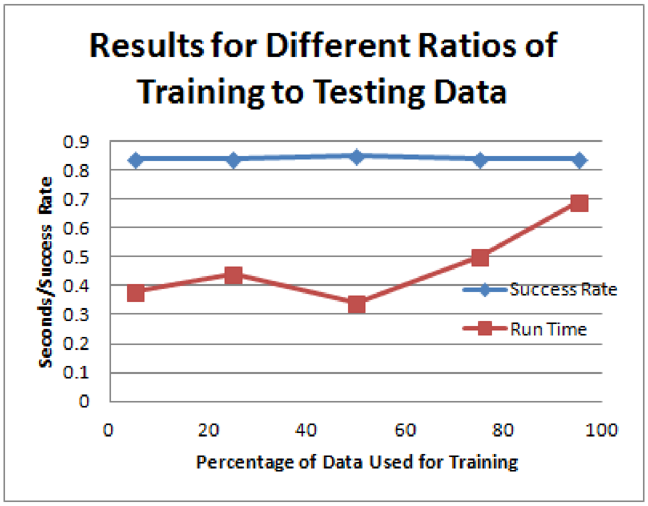
The final script used to build the classification tree and classify test data was approximately 180 lines long, including comments. Methods were developed to determine entropy and information gain, split the data against a specific attribute, and classify testing data not included in the training set.

V. Results

After running my decision tree classification script against randomized selections of the data set, the following results were found: the depressive episode attribute always offered the greatest initial information gain, and was therefore always the root of the decision tree, followed by the drug or alcohol treatment attribute. While drug use is commonly considered a strong indicator for criminal behavior (or is the criminal behavior leading to arrest or incarceration consequences itself), in this particular data set, the mental health attribute was a stronger differentiator. While all respondents with a criminal record may have had depressive episodes in this data set, this may also simply indicate the growing prevalence of depression, which, as predicted by the World Health Organization, will be the most common cause of disability world-wide by 2020 [11]. Correlation, even at the level found here, does not imply causation.

In addition to using different random selections of the data as my training set, I tried altering the amount of records used for training versus testing to see if there was a noticeable difference in performance or accuracy. Across all tests, this algorithm achieved an average 84.34% accuracy in predicting the Arrested and/or Jailed class for respondents, and completed the tree and classifications in an average of 0.47 seconds. As seen in Figure 1, changing the ratio of training to testing data did impact run time, but did not have any strong impact on the accuracy of the decision tree. Considering that using only 5% of the records to build the tree resulted in approximately the same accuracy as 95%, we can conclude that either the data set contained many of the same kinds of records (which could be confirmed by a clustering technique), or that the class was one-sided enough that simply predicting 0 (or 'no') for the result would be a good prediction.

Fig. 1. Comparison of different ratios of training-to-testing data. Points represent: 5% used for training/95% used for testing, 25%/75%, 50%/50%, 75%/25%, and 95%/5%.



VI. Future Work

There are certainly many additional avenues to explore within this data set, and the domain as a whole. In terms of extending the work I have detailed here, additional improvements could be incorporated into the existing code in terms of pruning the decision trees it builds (which would become even more important if additional attributes were included in the data set), as well as optimizing the code itself to ensure adding larger data sets would not be a crippling experience.

In terms of expanding the scope of the work here, the remaining 3,000 attributes found in the initial data set undoubtedly contain their own information to be mined and analyzed. They could be approach either en masse, with a decision tree working against a wider data set, or as subsets with a narrower focus – looking particularly at tobacco usage for instance, or abuse of prescription or over-the-counter medications, or age-specific patterns.

Entirely different methods of data mining are also worth exploring. Clustering has been used in the criminal data domain [10], and association analysis used to generate a list of risk factors has been successful in the medical data side [8]. Considering the infrequency of some of the important attributes found in this kind of data (for instance, arrests for violent crime) anomaly detection methods could also be implemented to find the more unique individuals.

Given access, I would also like to incorporate additional data providing more domain knowledge, especially in the sphere of potential treatments for some of the issues found within this data set. With that kind of information, suggested treatment options could be recommended as well.

VII. Conclusions

Mental health, substance use and abuse, and criminal behavior and history, and all of their specific treatments are all complex domains individually. In combination, they present a complex record of an individual. Data mining techniques like the decision tree classification presented in this paper are one promising way of making sense of these kinds of complex data sets. While privacy and ethical issues must always be a part of the conversation, predictive techniques based on real data can be an important tool for domain experts to assist in exploring treatment and prevention options for the good of their patients.

References

[1] M. Thompson, *Gender, Mental Illness, and Crime*, tech. report, Dept. of Sociology, Portland State Univ., 2008.

[2] V. Kumar, M. Steinbach, and P. Tan, *Introduction to Data Mining*, Addison Wesley, 2006.

[3] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From Data Mining to Knowledge Discovery in Databases,” *AI Magazine*, vol. 17, no. 3, Fall 1996.

[4] T. De Bie, “An Information Theoretic Framework for Data Mining,” *Proc. 17th ACM SIGKDD Int’l Conf. on Knowledge Discovery and Data Mining* (KDD ’11), 2011, pp. 564-572.

[5] C. Clifton, J. Jin, and M. Kantarcioğlu, “When do Data Mining Results Violate Privacy?,” *Proc. 10th ACM SIGKDD Int’l Conf. on Knowledge Discovery and Data Mining* (KDD ’04), 2004, pp. 599-604.

[6] W. Seltzer, “The Promise and Pitfalls of Data Mining: Ethical Issues,” *2005 Proc. American Statistical Association, Section on Gov’t Statistics*, 2005, pp. 1441-1445.

[7] D. Pedreschi, S. Ruggieri, and F. Turini, “Discrimination-aware Data Mining,” *Proc. 14th ACM SIGKDD Int’l Conf. on Knowledge Discovery and Data Mining* (KDD’08), 2008, pp. 560-568.

[8] J. Li et al., “Mining Risk Patterns in Medical Data,” *Proc. 11th ACM SIGKDD Int’l Conf. on Knowledge Discovery in Data Mining* (KDD ’05), 2005, pp. 770-775.

[9] S. Srisuk and P. Thongtae, “An Analysis of Data Mining Applications in Crime Domain,” *IEEE 8th Int’l Conf. on Computer and Information Technology Workshops*, 2008, pp. 122-126.

[10] T. K. Cocx, J. S. de Bruin, J. N. Kok, W. A. Kosters, and J. F. J. Laros, “Data Mining Approaches to Criminal Career Analysis,” *Proc. 6th Int’l Conf. on Data Mining* (ICDM ’06), 2006, pp. 171-177.

[11] T. Dillon, F. Hadzic, and M. Hadzic, “Tree Mining in Mental Health Domain,” *Proc. 41st Hawaii Int’l Conf. on System Sciences*, 2008, pp. 230.