Data Mining of Crime Research Information Statistics Portal: The Experience and Lessons Learned

Christopher Dondanville, The University of Memphis, Memphis, TN 38152; E-mail: cdndnvll@memphis.edu
Xihui Zhang, The University of Memphis, Memphis, TN 38152; E-mail: xihui.zhang@memphis.edu
Ted E. Lee, The University of Memphis, Memphis, TN 38152; E-mail: elee@memphis.edu

ABSTRACT

In an effort to find interesting information in a volume of crime data provided by the Secure Police Department (a fictitious name), we applied data mining techniques and algorithms to this data. We discovered that data mining is an effort that takes a lot of planning and preparation, requires domain expertise, and is extremely resource intensive. The results we got are lackluster, for instance, correlations and patterns from the results are readily apparent, they are very rarely of the "interesting" variety. However, we think the experience and lessons learned have helped us and will help others be prepared for any future similar endeavors. The lessons can be summarized as: 1) Planning and preparation is a key to successful data mining, 2) Domain expertise is a necessity to perform successful data mining, 3) Data mining applications are extremely processor and main memory resource intensive and the appropriate hardware and software are necessary for successful data mining.

INTRODUCTION AND BACKGROUND

In the United States today crime is a major concern. The CIA, FBI, and other federal agencies are concerned with national security; city and county law enforcement agencies keep a constant vigil on criminal activities in their own jurisdictions; and individuals are worried about their safety within the community. In the meantime, the budget for local law enforcement and intelligence agencies takes a hit due to mediocre economic growth, which leads to downsizing of both personnel and programs. To better make decisions and optimally allocate limited resources, local law enforcement and intelligence agencies need better quality information in a timely manner.

It has been observed that "A major challenge facing all law-enforcement and intelligence-gathering organizations is accurately and efficiently analyzing the growing volume of crime data" (Chen et al., 2004). As the volume of this crime data becomes enormously large, new techniques have to be used to turn this data into usable information and knowledge; and thus, appropriate actions can be taken accordingly. Data mining holds the promise of making it easy, convenient, and practical to explore very large databases for law enforcement and intelligence agencies. Data mining is a powerful tool that uses complex algorithms to look for patterns in very large sets of data. Criminal investigators who may lack extensive training as data analysis can easily explore large databases quickly and efficiently through the implementation of data mining techniques in the form of commercial and other applications (Chen et al., 2004; Fayyad and Uthurusamy, 2002).

The Shared Urban Data System (SUDS) is a web based secure portal that has been created as part of the Community Information Project by the Center for Community Criminology and Research at the *Academic University* (a fictitious name) located in the Mid-south area. SUDS provides a host of information-based services for the local community including neighborhood planning and development, public safety, and public health, etc. While SUDS is a portal to all information and services housed in a large number of databases, the CRISP (Crime Research Information Statistics Portal) system is a more specific subset of information services tied to a single database. The CRISP database is of crime incident information provided by the *Secure Police Department* as part of the Community Information Project. The database is used for queries and reports centered on offense types, geographic locations (through a GIS mapping interface), dates and time ranges, and suspect and victim information.

This paper describes a project that applied a variety of data mining techniques to the CRISP database using a commercial data-mining tool. The tool selection was based on the tool availability, usage characteristics, and architecture. The tools evaluated were IBM Intelligent Miner, SPSS Clementine, and Megaputer PolyAnalyst. Our eventual decision to use Megaputer PolyAnalyst 5.0 was due to the availability to the student for a research product, the ability to use the tool directly, via ODBC, in connection with an SQL Server 2000 server without an intermediate tier, and the incorporation of a large number of data mining algorithms available within the base program.

LITERATURE REVIEW

Our summary of existing research literature on data mining of crime data is organized as follows. We first talk about the imperatives of applying data mining techniques to crime data and we then go into details on different data mining techniques that have been successfully applied to crime data.

To us, the marriage between data mining techniques and crime data happened naturally. A retrospective review tells that there are roughly three major reasons why this happened. First, the volume of crime data has become unwieldy. When investigating a crime, law enforcement agencies usually record as much information as possible about the crime. After the attack on WTC in New York City on September 11, 2001, concern about national and local security has increased dramatically. At a national level, the FBI, CIA and Homeland Security are all acting to collect and analyze information to prevent further terror attacks. These efforts have had an affect on local authorities, motivating them to monitor criminal activities in their own jurisdictions using similar intelligence techniques (Chen et al., 2004). All above activities result an exponential increase in the volume of crime data. Manual processing of this data is not feasible anymore; new techniques to accurately and efficiently process crime data had to be found.

Second, the budget for local law enforcement and intelligence agencies continues to see reductions due to mediocre economic growth (as of this writing in 2006). These economic factors lead to downsizing of both law enforcement personnel and programs. To better make decisions and optimally allocate limited resources, leadership in law enforcement needs quality information in a timely manner.

Third, its ability to reveal patterns in extremely large data sets makes data mining a tool that fits this job perfectly (Brown, 1998). Compared to manual processing, which is time and labor intensive, data mining holds the promise of making it easy, convenient, and practical to explore very large stores of crime data for law enforcement organizations and users.

There are many data mining techniques can be applied to crime data. Some examples of the specific applications and techniques of this work are given below as derived from current scholarly literature on the data mining of crime data.

Lin and Brown (2006) present an outlier-based data association method for linking criminal incidents. In this technique, according to them, "an outlier score function is defined to measure the extremeness of an observation, and a data association method is developed based upon the outlier score function." They applied this

method to the robbery data in Richmond, Virginia, and compared the result with a similarity-based association method. Their results show that the outlier-based data association method is promising.

Estivill-Castro and Lee (2001) incorporate two knowledge discovery techniques, clustering and association-rule mining, into a fruitful exploratory tool for the discovery of spatial-temporal patterns. They present two methods for this exploratory analysis and the detail algorithms to effectively explore geo-referenced data. They illustrate the algorithms with real crime data. They demonstrate their approach to a new type of analysis of the spatial-temporal dimensions of records of criminal events.

Brown (1998) describes a software framework for building and applying data mining algorithms to crime analysis problems. This framework provides specific focus on spatial data mining. The author provides several reasons to justify this focus: 1) spatial queries are more time consuming, 2) spatial analysis is harder to do than analyses based on attribute matching, 3) spatial data mining has the potential to yield important immediate benefits for crime analysis as crimes have an inherently spatial component (Brantingham and Brantingham, 1984), and 4) spatial analysis is a key to law enforcement resource allocation.

Chen et al. (2004) present a general framework that shows the relationship between data mining techniques applied in criminal and intelligence analysis and the crime types. They identify and arrange eight crime types (traffic violations, sex crime, theft, fraud, arson, gang/drug offences, violent crime, and cyber crime) in increasing order of public harm on the horizontal axis. On the vertical axis, they arrange the techniques in increasing order of analysis capability. They identified four major categories of crime data mining techniques: entity extraction, association, prediction, and pattern visualization. Each category represents a set of techniques for use in certain types of crime analysis. They then identified the intersection of the techniques with the crime types denoting where each technique could be effectively used for each crime type, completing the framework.

THE ARCHITECTURE USED IN THIS STUDY

PolyAnalyst is a product of Megaputer, and Version 5.0 is the latest version of PolyAnalyst. It is a data-mining tool that incorporates a large number of data mining algorithms in a single package. It can be configured to work as a traditional client server application or can be set up in a multi-tier architecture with a dedicated PolyAnalyst server. The client server implementation can be configured to read the data from the database importing all of the data into the project file created by PolyAnalyst or, alternatively, a subset of the algorithms can be used, via OLEDB to mine the data in place. We took the conventional means of using a traditional client server approach and allowed the application to import the data. Our connection to the database was done over a wide area link via a secure VPN connection to the network of *Academic University* to the SQL Server 2000 server where the CRISP data was located.

Some challenges encountered were the ability of the program to deal with the large amount of data and the physical limitations of the client server infrastructure. With the size of our datasets being much larger than 500,000 records, the PolyAnalyst program seemed to have some difficulty reading and dealing with a dataset this size given the simple client server configuration. In talking with the representatives of Megaputer, many of these shortcomings would be addressed by moving to a multi-ties architecture, using more advanced, server-based, architecture. Running the client under Windows XP Professional or Windows 2000 workstation would often reach the 4 GB limit on the operating systems address space.

To get around some of the limitations, we used a strategy that allowed us to work with smaller subsets of the data. Since we had to select our data via a join across six tables to get the information that we need, we included conditions that would select one year at a time in our data selection statements. Once we had the data imported into PolyAnalyst in a project file, we further subdivided the data into more manageable sections for testing different algorithms prior to starting large mining runs.

DATA PREPARATION AND ANALYSIS

Discussions with criminologists to narrow area of interest and type of discovery led us to believe that weapons crimes, specifically gun related crimes, were of particular interest to the SUDS team. There is a special program within the *Secure Police Department* at this time to crack down on gun crime. Anything that could be found to assist in the pursuit of this endeavor would be seen as a positive production of information from a data mining perspective.

Prediction and prevention are other major goals of the *Secure Police Department* and something that they would like to be able to do better. If a model can successfully predict some criminal activities, prevention mechanism can be deployed to effectively prevent some crimes from being committed. This would aid in crime deterrence and more effective utilization of police resources. Any information that data mining could provide to assist in the mission of prediction and prevention would be seen as positive generation of information.

Geospatial Modeling using Geographic Information Systems (GIS) is a very powerful tool that can be used to visually present crime data based on geospatial features. These kinds of presentation can be used to spot crime patterns. This type of information and modeling is being done on a small scale through the SUDS project at this time. A larger and more automated effort than the current one is desirable. There is the potential to create such special game maps from the data used from the CRISP database.

Data preparation for our task was significant and mostly a trial-and-error experience. The data had been pre-cleaned for the SUDS project so most problems had been dealt with. Other than having to tune our selection SQL to return meaningful, yet manageable dataset, we had to convert string values to categorical and Boolean fields that PolyAnalyst could interpret and also to handle some missing values that were in the data. In the creation of a commercially viable data-mining tool, Megaputer had to create generic algorithms that could be used under a wide variety of circumstances. In doing so they had to place some restrictions and criteria on the input of data to the algorithms. Such restriction took the form of selective data types, for example. When using a directed algorithm, such as a Decision Tree, a target attribute could only be of certain data types (say, categorical). A previous understanding of the direction of the inquiries and the organization of the data will help greatly in setting up the data sets for analysis.

Part of the data preparation was to study the metadata that we had been provided by the criminologists and system administrators of the SUDS database for context and use of the attributes. While a lot of the variables appeared straightforward, some variables needed explanation as to their particular use or importance, relevant to our tasks of data exploration. Redundancies and other dimensionality reductions were done at this point. There was often overlap from one table to another that was not apparent due to lack of naming conventions and other factors.

The analysis that we performed on the dataset was largely explorative in nature. We looked at many algorithms to see if we could apply some of them to our project. Below are a list of the techniques we considered and a brief description of the algorithms used taken from the PolyAnalyst User Manual.

- Nearest Neighbor - The Nearest Neighbor exploration engine uses a memory-based classification system: assigning values to data points based on their "proximity" to other data points.
- Market Basket - The explored dataset consists of some number of records which are referenced below as "transactions" and a number of attributes or "products." The Basket Analysis engine finds sets of products that are present together in a significant part of all transactions. Such typical sets are called baskets of products. Certain additional limitations can be imposed when searching for these baskets, such as specifying the minimum portion, in percent, of all transactions containing a discovered basket.
- Cluster - Clustering is one of typical problems solved by data mining methods. This is a process of grouping cases or database records into subsets, such that the degree of similarity between cases in one group is significantly higher than between members of different groups. An exact definition of the similarity between cases as well as other details varies for different clustering methods.
- Decision Tree – This is the name given to large family of machine learning algorithms for the automated construction of tree-like classification rules for categorizing structured data. The process of creating a Decision Tree can be represented as splitting the analyzed dataset into diminishing parts consisting of increasingly homogeneous records: in terms of the percentage of different values of the target field.
- Text Analysis - The Text Analysis exploration engine performs morphological and semantic analysis of unstructured textual notes in a database format. Text Analysis extracts and counts the most important words and word combinations from textual notes, and stores terms-rules for tokenizing database records with patterns encountered terms.

Once we had created subsets of our main dataset it was easy to test many algorithms for their appropriateness. The subsets of the datasets were done with tools in PolyAnalyst that would let you take random, top or bottom samples to create data subsets. They could be done on a percentage of total records or a total count basis. You could also decide which of the attributes to include in the subsets of data. Once the subsets were created it was a simple matter to choose that subset and apply a technique to it. Similarly, you could use a subset to train a directed algorithm and then apply the rule set generated by the training to the remaining data, or the World data as it was referred to in PolyAnalyst.

The analyses that we completed included a Link Analysis, a Decision Tree, and a Text Analysis. We directed our decision tree and link analysis algorithms to target the attribute that contained values for "Weapon Type" and our text search was directed to search for weapons in the text phrases of the narratives given with each of the cases. PolyAnalyst contained built-in semantic libraries through a connection to external rule base for guiding the text search as well as common files of stop words. The results of the classification and association routines provided a large number of links, most of which were meaningless. Some of the more meaningful links, that had stronger correlation and coverage were also, unfortunately, quite intuitive. The text analysis looked primarily at occurrences and the results there seemed good, but not perfect. For example the semantic database picked up on 35 occurrences of the term "cutlass" in the narratives and determined that it was a weapon. Although this is possible, we found it more probable that these occurrences were in reference to an Oldsmobile and not to a sword fight.

RESULTS AND DISCUSSION

A result that became readily apparent in all of our analyses was that correlations are easily found and meaningless correlations are even easier to find. For example, the use of a weapon in crimes of type "Assault" had a large correlation and coverage from several of the algorithms giving the rule:

Assaults ⇔ Weapon Use

Similarly the absence of the use of a weapon in crimes of type "Home Burglary" had a large correlation and coverage from several of the algorithms giving the rule:

Home Burglary ⇔ No Weapon

Even more simplistically the correlations where 2 geographic location identifications were used in the same algorithm they were both correlated as can be seen by the following rule:

Zip code ⇔ Precinct and Ward

It was the "interesting" correlations that proved to be elusive to our exploration.

CONCLUSIONS, LESSONS LEARNED, AND FUTURE DIRECTION

Although the things that were learned here are not ground breaking or revolutionary, they were situations that we encountered in the actual attempt to perform a data mining exercise from scratch. While there are many studies and publications that talk about data mining and crime data analysis, these lessons "from the field" are presented here as practical information for both researchers and practitioners.

A major revelation in our data mining trials was that the preparation and planning is central to data mining; whereas the execution of algorithm and analysis can be an easy part. We spent much more time up front researching the importation, mapping, cleaning, and retyping of data than we did on the exploration phase of the project. Similarly we found that setting up a tool to mine existing data can be a challenging and time consuming process. The data preparation could have been done better to accommodate the tool and made the data mining process a lot easier. The data, which was primarily in string fields for use in a web portal environment, was not conducive to data mining. For example, it would have been much easier to manipulate those fields with a value of "True" or "False" if the field values were 1 or 0, respectively, in a data-mining context.

Another finding from this study was that this type of analysis is highly domain exclusive and a domain expert is an important resource. Without the assistance of criminology resources and even resources that had domain knowledge specific to our dataset, the task of exploration would have been much more difficult.

We also developed a new appreciation for how resource intensive data mining is. The ability to saturate a 4 GB address space in a matter of minutes of processing was impressive and let us know just how important it is to use well tuned and efficient algorithms and code in the data mining task. It was also apparent that these analyses were very processor intensive as well as main memory intensive. Some of the algorithms took several hours to run on datasets that were 80-90 thousand records in size.

In the future, we plan to develop a better understanding of tool and data handling inside the tool. This would make it much easier and less time consuming for us to choose the algorithm and prepare the data for that algorithm. This would also let us have more precision in selecting and assigning values to parameters of the algorithms. We would also like to be able to determine more interesting questions for exploration and analysis. This would involve a great deal more time with the criminologist and possibly with the police departments to find out what is of interest and helpful to them as we choose datasets, attributes and algorithms. Another improvement that could be done to advance this line of study would be to improve the infrastructure to handle larger datasets. The addition of another tier with a dedicated PolyAnalyst server would facilitate the use of larger datasets and give quicker processing times.

Applying data mining techniques on crime data is not an option, but a necessity. Data mining, given quality data and the appropriate techniques, can bring forth accurate information in a timely fashion and this information can enhance decision-making and analysis for all law enforcement agencies. Past successes in crime data mining suggest the future of it is very promising. We think our experience and lessons learned during this exploratory study have helped us and will help others be prepared in any future similar endeavors.

REFERENCES

Brantingham, P. and Brantingham, P. (1984). *Patterns in Crime.* Macmillan Publishing Company, New York.

Brown, D.E. (1998). The Regional Crime Analysis Program (ReCAP): A Framework for Mining Data to Catch Criminals. *Proceedings of 1998 IEEE International Conference on Systems, Man, and Cybernetics,* 3: 2848-2853.

Chen, H., Chung, W., Xu, J.J., Wang, G., Qin, Y., and Chau, M. (2004). Crime Data Mining: A General Framework and Some Examples. *Computer,* 37(4), 50-56.

Estivill-Castro, V. and Lee, I. (2001). Data Mining Techniques for Autonomous Exploration of Large Volume of Geo-Referenced Crime Data. *Proceedings of the 6th International Conference on Geocomputation.*

Fayyad, U.M. and Uthurusamy, R. (2002). Evolving Data Mining into Solutions for Insights. *Communications of ACM,* 45(8), 28-31.

Lin, S. and Brown, D.E. (2006). An Outlier-based Data Association Method for Linking Criminal Incidents. *Decision Support Systems,* 41(3), 604-615.