Data Mining and Predictive Analytics in Public Safety and Security

Colleen McCue

Used for many years in the business community, data mining and predictive analytics are finding new roles in areas outside business. Also referred to as *knowledge discovery* or *sense making* tools these analytical processes can help analysts, managers, and operational personnel identify actionable patterns and trends in data. Briefly, data mining is "[a]n information extraction activity whose goal is to discover hidden facts contained in databases" (http://www.twocrows.com). In other words, data mining involves the systematic analysis of data using automated methods in an effort to identify meaningful or otherwise interesting patterns, trends, or relationships in the data.

Business can learn some lessons from law enforcement in translating data into actionable information.

Crime and criminal behavior, including the most aberrant or heinous crimes, frequently can be categorized and modeled—a characteristic used successfully in the apprehension of serial killers and child predators, as well as drug dealers, robbers, and thieves. So it's no surprise that data mining and predictive analytics are rapidly gaining acceptance and use in the applied public safety, security, and intelligence setting given their ability to process large data sets and identify actionable patterns and trends. In particular, the areas of force deployment and homeland security have emerged as early adopters of this technology.

MAKING THE LEAP FROM BUSINESS TO LAW ENFORCEMENT

Unfortunately, the translation from business to applied law enforcement and intelligence has been neither straightforward nor easy. Significant challenges associated with data quality and the need for operationally actionable output has limited widespread use. Although varied, most of the challenges fall into two categories: data and output.

Data challenges

The challenges associated with public safety and intelligence data transcend the oft-cited "stove pipes" that limit access and functional integration of data resources. Fundamental issues associated with the fact that most (if not all) public safety and intelligence data resources were collected for reasons other than analysis seriously limit the analyst's ability to extract meaningful output from these data. Moreover, it is difficult (if not impossible) to anticipate the array of data resources that analysts will encounter during the course of their work. Incident data, narrative reports, financial transactions, telephone records, and Internet activity represent only a few of the many and varied information resources used in crime and intelligence analysis. In addition, challenges associated with reliability and validity plague public safety and intelligence data. Criminals lie, witnesses forget, and victims are often too anxious to be reliable reporters of an incident. Taken together, these issues seriously compromise most data that crime and intelligence analysts encounter.

Output challenges

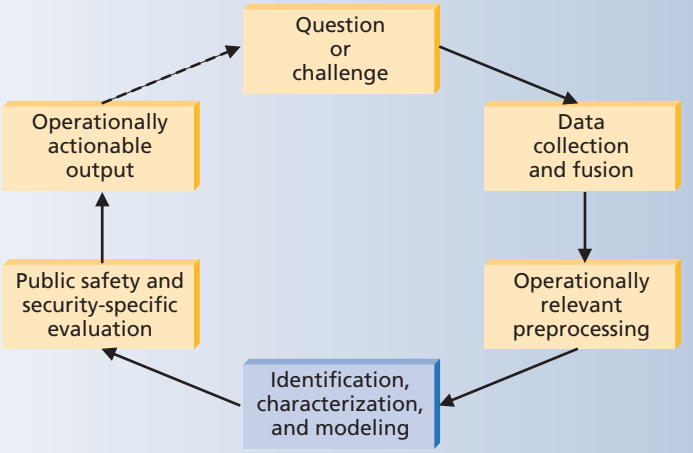
On the other hand, someone must be able to understand and use the analytical output. Given this, analysts must frequently balance accuracy and the ability to create operationally-actionable out-

Published by the IEEE Computer Society
1520-9202/06/$20.00 © 2006 IEEE

put when developing models. Compromises between the two are the rule rather than the exception. All of this requires a significant degree of subject matter or domain expertise to make judicious choices in this area. Author Tom Clancy, in a keynote speech at the Gartner IT Security Expo, advised security and intelligence professionals to seek out the "smart people," observing that, "The best guys are the ones who can cross disciplines … The smartest ones look at other fields and apply them to their own" ("Clancy Urges CIOs: Seek Out the 'Smart People'," D. Fisher, *eWeek*, June 2003; http://www.eweek.com/article2/0,1895,1657818,00.asp).

Successful data miners in the applied public safety and security field are those that have a solid understanding of crime, criminals, operational requirements, and tactics. They also understand the math and computational science associated with data mining and predictive analytics. It is only through these combined skills and tacit knowledge that an analyst can effectively translate the output of data mining and predictive analytics to actionable information for applied law enforcement and intelligence.

Figure 1. Process model for Actionable Mining and Predictive Analysis for Public Safety and Security.



The modeling box is blue to indicate the possible use of standard COTS tools. Think of it almost like an "insert your favorite data mining software here" box. The steps represented by the yellow boxes are unique to actionable mining. The dashed line connecting the output to the question or challenge box indicates that the process can be iterative.

Other challenges

Data mining has been the subject of considerable controversy recently, particularly regarding privacy concerns and the US federal government's use of these tools. Data mining itself, however, poses no direct threat to privacy. Restricting access to sensitive data—rather than limiting or otherwise restricting the use of computational techniques—maintains privacy and protects rights. Unfortunately, this confusion between data access and analytical process threatens to limit one of the most powerful analytical techniques available to address threats to public safety and national security.

ACTIONABLE MINING AND PREDICTIVE ANALYSIS FOR PUBLIC SAFETY AND SECURITY

Approximately 80 percent of the data mining process is spent on the data preprocessing and preparation steps (*Data Mining with Confidence*, 2nd ed., C. Helberg, SPSS Inc., 2002)—that is, the decisions about what data to collect. Analytical strategies developed to specifically address the unique issues associated with data quality and availability, as well as the need for operationally-actionable analytical output hold the promise of greatly increasing the ability to fully use data mining tools in support of public safety and homeland security. In most situations, once analysts address the data preprocessing and output issues, they can use commercially available software packages for the actual modeling.

In an effort to address these challenges, I developed a process model for Actionable Mining and Predictive Analysis for Public Safety and Security ("actionable mining" for short), which includes the following steps:

- question or challenge;
- data collection and fusion;
- operationally relevant preprocessing (including recoding and variable selection);
- identification, characterization, and modeling;
- public safety and security-specific evaluation; and
- operationally actionable output.

The first three steps constitute the estimated 80 percent of the process, as I mentioned earlier. Figure 1 shows a basic diagram.

Question or challenge

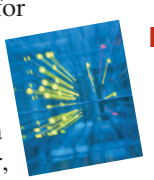
In the process' initial phase, analysts identify the general question or challenge, and convert it into a specific question that the data mining process can answer. Although this might appear simplistic and unnecessary, analysts will use this question to structure the analytical design plan, guide the process, and ultimately evaluate the answer's fitness and value. Therefore, this particular exercise is crucial.

Sometimes, the questions are specific: Are these crimes linked? When do most robberies occur? Do people continue to buy drugs in bad weather?

Other times, however, the task initially presents itself as a vague question or issue, which might require some preliminary analysis to reveal the underlying question or challenge. For example, an analyst might receive a list of financial transactions and be asked to determine whether there are any indications of fraud or other suspicious financial activity.

Data collection and fusion

As mentioned earlier, most if not all data analyzed in public safety and security were collected for some other purpose, which affects data form, content, and structure. For example, crime incident reporting forms generally are not constructed with data mining and analysis in mind. Moreover, some of the most important information in an incident report appears in the unstructured narrative portion of the report; this includes information relating to modus operandi (MO) and other important behavioral indicators. Unfortunately, it is this section of the report that also contains misspellings, typographical errors, jargon, slang, and incomplete and missing information. These as well as other irregularities limit the analysts' ability to effectively analyze the data and interpret results.



Dividing data into temporal and spatial sets helps identify potential threats.

Operationally relevant preprocessing

This phase of the data mining process includes data validation, cleaning, recoding, and variable selection. This task assumes even greater importance in public safety and security analysis given the limitations associated with public-safety-related data. To address these issues, the actionable mining model divides the preprocessing step into operationally relevant recoding and variable selection.

Recoding. The recoding phase of data preparation includes transformation and cleaning, as well as the assessment of data quality. Data recoding also occurs during this step. Most problems or challenges in public safety and security are reducible to an analysis of time, space, and the nature of the incident or threat.

The advent of handheld GPS (global positioning system) tools and collection devices has greatly facilitated the collection of precise data, which can support the accurate and reliable identification of crime incident locations. For example, GPS technology can be used for the analysis of border-related crime in that GPS can link specific incidents to unique locations in the absence of traditional location indicators such as street address.

However this is not always the best strategy for data mining and predictive analysis. Rather than treating time and space as continuous variables, my experience supports the creation and use of temporal and spatial *sets* that facilitate analysis and output. This recoding strategy results in cate-

gorical data better suited for many of the mining and modeling steps, and also provides a better match for reality.

For example, although you could characterize a series of crimes as having occurred at 4:58 p.m. ±32 minutes, this representation is an extremely difficult number to use operationally. Rather, creating 4-hour time blocks serves to aggregate the data, which can be critical with low-frequency events, and results in output that is easier to translate into existing schedules. On the other hand, it is unlikely that a criminal selected the time 4:58 PM ±32 minutes. Rather, the incident time might likely correlate with some other time or activity related to the criminal's preference or the potential victim's availability.

Similarly, criminals rarely target the same exact location multiple times. Crimes typically occur in clusters at similar locations or in close proximity. Therefore, reliance on specific location information might obscure the identification of relationships based on qualitative spatial attributes or general location.

Finally, recoding the nature of the incident or threat is useful in supporting specifically targeted tactics and strategy. For example, you could recode armed robberies of a convenience store, a beauty salon, and a dry cleaning store as "commercial robberies" because they involve robbery of a commercial entity rather than a person. Because the original incident information comes from various sources, it might not use the same terms for similar incidents. To make the most of the data, you want to recode the incident descriptions to use the same descriptors for similar incidents.

Similar to a treatment-matching approach in medicine, identification and characterization of the potential threat reduces the likelihood that operational personnel will bring the proverbial knife to a gunfight.

Variable selection. Not all variables available for analysis will have meaning or value in the applied setting. For example, my colleagues and I found that the use of a sawed-off shotgun was associated with an increased risk for assault during an armed robbery. Unfortunately, it was not possible to proactively deploy resources or create any sort of operational plan based on this finding.

Similarly, academic research is frequently based on solved crimes: The ability to use an array of variables, including those related to offender characteristics and attributes, can generate models that are enviable in their accuracy and predictive value. These findings, although interesting, often hold limited value for the applied setting in that they can use data and variables that are not available during an investigation. Therefore, in applied data mining, analysts must be aware of the potential operational value of variables as well as their availability.

Identification, characterization, and modeling

Analysts select and apply specific statistical algorithms to the data during the identification, characterization, and modeling phase of analysis in an effort to identify, characterize, and model natural trends, patterns, and relationships in the data. The statistical algorithms used in data mining generally fall into two groups:

- unsupervised learning or clustering techniques and
- rule induction models or decision trees.

Selection of a particular algorithm is based on the analysis' overall goal and the data's specific numeric attributes. Unsupervised learning or clustering techniques group the data into naturally occurring groups or sets based on similar attributes or features. These techniques can also help detect data that are anomalous or significantly different from the rest of the sample.

Rule induction models, on the other hand, exploit the fact that criminal behavior can be relatively predictable or homogeneous. Analysts can characterize and model specific attributes or behavioral patterns using rule induction models, which they can then apply to new data or incidents in an effort to quickly categorize them based on common features or attributes. This use is similar to the scoring algorithms used to predict risk in the financial industry. These models can be based on empirically determined clusters identified using unsupervised learning techniques or those predetermined by the analyst.

Public safety and security-specific evaluation

In this phase, analysts should evaluate three goals. First, do the analysis results answer the question(s) posed in the beginning? This is a relatively standard measure of success that is not unique to the public safety and security arena.

Second, does the model have an acceptable level of accuracy, and do analysts understand the nature of errors? Predicting low-frequency events like crime can be particularly challenging, and the overall accuracy of the models created can be somewhat deceptive with these low-frequency events. For example, a model would be correct 98 percent of the time if it always predicted "no" for an event with an expected frequency of 2 percent. So although the overall accuracy is impressive, it would be an unacceptable measure for the predictive value of this type of model. In

Resources

The Two Crows Web site (http://www.twocrows.com) is an excellent source of accurate, yet easy to understand information on data mining and predictive analytics. The Web sites at http://www.dataminginglab.com and http://www.kdnuggets.com are also helpful.

Figure 2. Sample output of crime data used to support force deployment decisions.

Saturday-Sunday

Area	Time					
	1200-1559	1600-1959	2000-2359	0000-0359	0400-0759	0800-1159
1						
2						
3						
4						
5						

these situations, the nature and direction of the errors can provide a better estimate of the model's overall value.

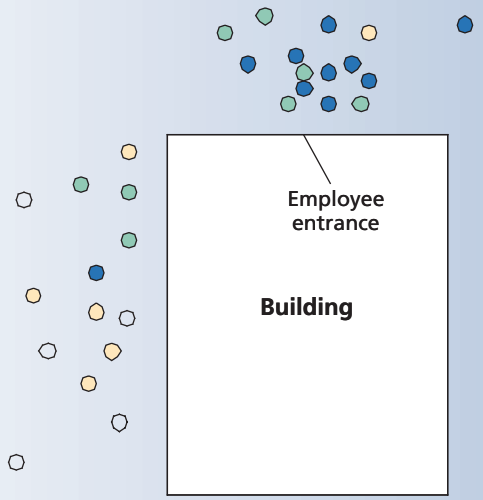
Solid subject matter expertise is essential to creating models that effectively balance accuracy with other operational considerations. In some situations, such as predicting weather-related crime, almost anything that provides accuracy better than just flipping a coin represents an improvement over what is currently available. In other situations, however, analysis or interpretation errors can cause public safety agencies to misallocate resources, waste time, or even lose lives. Therefore, it is essential that analysts work closely with operational users to ensure that the models are valuable and actionable in the applied setting, and that any necessary compromises in accuracy are acceptable.

Third, are the analytical output and models readily understood in and translated to the applied setting? Even the most elegant model has limited value in the applied public safety and security world if the people who must use the information cannot translate it into the operational setting, which foreshadows the final step in the actionable mining process.

Operationally actionable output

The ability to translate complex analytical output into a format usable in the operational setting is essential to the effective exploitation of data mining technology in applied public safety and security. Sophisticated analytical tools have been available for several years, and complex analytical strategies are commonplace in academic criminal-justice research. Only recently, however, have public safety and security personnel started to use these tools and approaches because the analytical output generated by sophisticated algorithms and tools had little direct relevance to the applied setting.

Figure 3. Sample output of possible hostile-surveillance data.



CASE STUDIES

The following case studies exemplify the use of data mining in applied public safety. They illustrate the methods used to convey data mining results in an operationally-actionable format.

Deployment

Reduced to its simplest form, force deployment basically represents a resource allocation challenge; specifically, the allocation of personnel across time and space. Analysts have used data mining and predictive analytics to support this process by creating models of criminal activity in a locality. Using these models, law enforcement organizations can deploy resources proactively in an effort to deter or prevent crime or to respond more rapidly when a crime occurs.

Figure 2 shows sample output from these modeling efforts. The different cells in the table represent the relative likelihood of crime in five predefined policing areas across 4-hour time blocks that span Saturday to Sunday. As the figure shows, the greatest activity is in area 3 from midnight until approximately 4 a.m. Using these relatively

simple techniques, it was possible to illustrate the relative risk of activity or need for police services in a format that law enforcement could easily interpret and translate directly into deployment strategy and action.

This example also illustrates the *force multiplier* capacity made possible by using data mining techniques in association with fluid deployment strategies. Again, the model predicts that the greatest likelihood for activity is in area 3 during the first 4 hours of Sunday morning, followed by somewhat diminished likelihood for activity in area 2 from 4:00 a.m. to 8:00 a.m. Depending on the activity and the relative proximity of these two locations, it might be possible to use the same personnel to address both areas.

Richmond, Virginia, used a similar approach to address increased citizen complaints of random gunfire on New Year's Eve. By identifying the times and locations historically associated with random gunfire complaints, police developed a targeted deployment strategy. The goals were to, at a minimum, increase the arrest rate by deploying resources when and where they were likely to be needed, thereby reducing response times. Law enforcement officials also hoped that an increased police presence in these areas would deter the illegal use of firearms.

The results were encouraging. These efforts increased the recovery of illegal weapons by 246 percent over the previous year, supporting the rapid-response value associated with the use of data mining to support deployment decisions. Even better, however, was the fact that citizen complaints for random gunfire decreased by 47 percent over the previous year. In addition to these public-safety benefits, the initiative required 50 fewer police officers than originally anticipated, resulting in a savings of $15,000 during the initiative ("Doing More with Less: Data Mining in Police Deployment Decisions," C. McCue and colleagues, *Violent Crime Newsletter*, Spring 2004, pp. 1, 4-5).

Surveillance detection

Data mining technology also has tremendous potential in homeland security to support early identification and prevention of possible terrorist attacks. The next example involves the use of data mining and predictive analytics to detect, characterize, and model a potential threat in support of infrastructure protection ("Data Mining and Predictive Analytics: Battlespace Awareness for the War on Terrorism," C. McCue, *Defense Intelligence J.*, vol. 13, pp. 47-63). In this example, the analysis focused on reports of suspicious behavior suggesting the hostile surveillance of a facility.

Figure 3 illustrates a sample output, plotting some of the collected and analyzed data. Using relatively simple visualization techniques, it was possible to depict key features of the analysis into an easy to interpret and use format. The plot shows the location of reported incidents of hostile surveillance. Each dot represents one incident report; the

darker the dot, the more recent the report. This plot visually highlights the apparent focus of the hostile surveillance activity on a particular aspect of the building. In this case, it was the employee entrance. Analysis of the time and day revealed that 25 percent of the incidents occurred on Wednesdays, most of them in the afternoon. Based on this information, managers focused more resources in this location to protect the facility. Additional analysis included a clustering technique, which identified two groups of hostile-surveillance behaviors and provided additional evidence in support of escalating activity associated with the facility.

Like the previous example, these findings can support deployment decisions, as well as guiding security enhancements and response planning. Highlighting the times and locations associated with possible hostile surveillance can reveal threats and help managers allocate resources.

The ability to successfully translate predictive analytics and their promise to applied public safety and security has required some accommodation of the unique challenges associated with these data, as well as the need for easily interpreted, operationally relevant output. The innovative use of visualization techniques helps translate the results of the analytical process directly to the applied setting. Using these techniques allows operational users to incorporate their subject matter expertise, intuition, and tacit knowledge in the interpretation of the results, adding significant value to the outcome.

The benefits of using predictive analytics in applied public safety and security are twofold. First, the early identification and characterization of a potential threat presents more options for prevention and deterrence. Second, predictive analytics supports information-based response planning. As highlighted by the New Year's Eve example, targeted prevention strategies also offer a greater return on the public safety investment by supporting the efficient allocation of resources. Moreover, prevention is almost always less expensive than response and recovery, particularly when measured in human terms. ■

Colleen McCue *is a senior research scientist at RTI International, an independent research organization dedicated to solving critical social and scientific problems. Previously, she was the program manager for the Crime Analysis Unit at the Richmond Police Department, during which time she also held adjunct appointments in the Departments of Surgery, Emergency Medicine, and Pediatrics at the Medical College of Virginia, Virginia Commonwealth University. She is the author of* Data Mining and Predictive Analysis: Intelligence Gathering and Crime Analysis *(Butterworth-Heinemann, expected release in September 2006). Contact her at cmccue@rti.org.*