Data Mining for Crooks

Mark Ingebretsen

Criminal investigations have always been about gathering and processing information. But in recent years, the tools have adapted significantly to the times. The dog-eared detective’s notebook and grimy loose-leaf folder full of mug shots have been replaced by intelligent search algorithms, machine learning, and statistical techniques designed to identify suspects and decipher their methods of operation.

“Any process that has some order to it—including the planning and committing of crimes—can lend itself to computer analysis,” explains John Bond, honorary fellow at the University of Leicester’s Forensic Research Centre. The AI techniques used by Bond and his colleagues throughout the world are aimed at uncovering that order.

Theirs is a mission that only AI can handle. That’s because today, as never before, criminals can move freely from state to state or nation to nation. Each new crime they perpetrate yields vast amounts of new data. Finding patterns within that data means sorting through seemingly unrelated bits of information scattered among dozens of law enforcement and private data sources. And those data sources can include everything from credit card and cell phone records to DNA traces, eye scans, and fingerprint data.

The need for better AI-based data mining techniques took on a new urgency following the 9/11 terrorist attacks. As a result, funding has flowed in from governments worldwide. “In the US, significant Department of Justice, Department of Homeland Security, and intelligence agencies’ ‘War on Terror’ funding in the past five years has helped public safety agencies,” says forensic data mining expert Hsinchun Chen, professor of management information systems at the University of Arizona and director of the school’s Artificial Intelligence Lab. That funding has been crucial in supporting much-needed information sharing among law enforcement agencies, says Chen, who is also associate editor in chief and news editor of this magazine. Additionally, he adds, those fresh funds have helped develop the intelligent analytical tools required to perform viable searches for patterns of criminal activity.

Heavy vs. Stocky

In a study conducted several years ago (www.springerlink.com/content/t240136u10298614), Bond and fellow researcher Richard Adderley, with the IT firm A-ESolutions, revealed the important role AI plays in forensic data mining. When Bond and Adderley used a neural network to look for errors and inconsistencies in crime reports compiled over the years, they found they were commonplace in each of the five UK police departments that formed part of their study.

Even the various words investigators use to describe a suspect’s common physical traits can foil standard search methods, Donald Brown, professor and chair of the University of Virginia’s Department of Systems and Information Engineering has noted. Normal information retrieval methods require exact matches between words, Brown

Also Featured

Intelligent Handicapping

Features Editor: Brian Brannon, bbrannon@computer.org
explained in a paper he cowrote, “Data Association Methods with Applications to Law Enforcement” (http://portal.acm.org/citation.cfm?id=640407). But if one police department lists a suspect as “heavy” and another one says he is “stocky,” a simple search might bypass the latter and overlook a valuable clue as a result.

As a workaround, Brown helped devise an algorithm that searches for matches by focusing on the physical and behavioral attributes of suspects. The system works by assigning a greater importance to readily observed characteristics that are also difficult for a person to change. For instance, if a suspect is heavy, the algorithm would give that attribute more importance than if his hair is red, since it is fairly easy to change the latter. Similarly, it would assign more weight to a weapon used by a suspect if it were a Japanese sword (an unusual choice) than to a common handgun.

**Needs within Many Haystacks**

Intelligent data mining techniques are particularly useful at focusing investigations in high-profile or especially heinous cases where vast amounts of data come together quickly and expand with each new day. Take the famous case in which two individuals randomly shot and killed people throughout the Washington, D.C., area during the fall of 2002. Thousands of tips flowed in through police hotlines, and each new crime scene required extensive analysis and cataloging of information. Investigators needed to sift through the data quickly, as new victims were gunned down.

One of the resources law enforcement used was Coplink (www.coplink.com). Developed by Chen and his colleagues at the University of Arizona and a recipient of War on Terror funds, Coplink is a secure Internet application that lets law enforcement agencies catalog and analyze data extracted from incompatible databases.

As Chen explains, Coplink’s search methods found a common element in witness reports, specifically the white van investigators once believed the DC snipers drove. The link was found “via association analysis based on related records in different jurisdictions,” Chen says. “The van was spotted in many gas stations during the DC sniper investigation. Association rule mining allowed that specific van to rise to the top of crime scene associations.”

**Crime Forecasting**

Coplink is reactive in nature. That is, law enforcement uses it to find perpetrators after a crime has taken place. Hundreds of law enforcement agencies have accepted it as an investigative tool.

AI-based data mining techniques quickly turn controversial when they become proactive—that is, when they perform routine searches to ferret out potential criminal activity. Such data mining activity is already occurring in Europe. As Bond explains: “The UK is participating currently in a European-based project to use AI to look for patterns of behavior that might indicate imminent or future terrorist activity. This ranges from the actions of someone about to plant or detonate a bomb to financial or lifestyle action that might indicate future terrorism sympathies.”

Such data mining techniques remain controversial in Europe and elsewhere. Shortly after 9/11, for example, the Bush administration sought to launch a mass surveillance program aimed at identifying terrorists called Total Information Awareness (see the logo on page 6), administered by DARPA.

According to media accounts, TIA was an umbrella program. Among its components was something called evidence extraction and link discovery, which was charged with creating technologies that would identify and find links among fragments of evidence found within both classified
and unclassified databases. Another TIA component called scalable social network analysis was supposed to employ common social-network analysis techniques to create a model of terrorist group traits.

Public outcry caused the TIA program to be quickly abandoned. American organizations ranging from the liberal American Civil Liberties Union to the conservative Cato Institute decried the fact that the program would gather cell phone records, Web sites visited, and library books checked out by ordinary Americans to identify them as potential terrorists.

Legal issues might not weigh as heavily in the War on Terror, where the object is not so much to convict someone as to prevent a violent act from taking place. “What we have developed and what is being developed are solely for what we could call ‘intelligence use,’” Bond explains. “That is, it is not used as evidence but as a means to focus police activity to find the evidence to present to a court.”

### Circumstantial Evidence

Yet the same analytical techniques could also identify someone as a suspect in a criminal case, or profile them as a potential gang member or drug dealer. As Chen explains, legal barriers stand in the way of that occurring, among them, “information sharing across different local, state and federal agencies; each has its own rules and regulations. The biggest issue,” Chen adds, “is about combining citizens’ data—such as registration records, flight records, car rentals, water bills, etc.—with crime records. This raises significant civil liberty issues. Most police agencies do not do this.”

A skillful defense attorney might even succeed at casting doubt on the data mining techniques used to pinpoint a suspect. Any methodology used to evaluate evidence must follow a well-established procedure before it is acceptable in court. “In order to become useful in court, a technique has to become generally accepted by the scientific community,” explains David Baldwin, who heads up the Midwest Forensics Resource Center at the US Department of Energy’s Ames Laboratory. “One measure of that is the peer-reviewed publications dealing with it.” Typically, the path to acceptance by the scientific community Baldwin describes refers to the analysis of DNA, trace chemicals, or other physical evidence found at a crime scene. In such instances, other labs can duplicate the testing results, thus assuring jurors that the results are correct. Testing a profiling technique might prove more difficult, and it could turn into a legal gray area with few precedents to serve as guidelines.

For now, “Crime data mining has not raised problems in court,” says Chen, “especially when all the results are validated by law enforcement before actions are taken.” But that might change, as the mountains of data law enforcement agencies must contend with continues to rise and as the analytical techniques used by forensic data miners become ever-more sophisticated.
promoted the Seattle Supersonics to substitute back-up center Frank Brickowski during the 1996 NBA Finals against the Chicago Bulls. With the Sonics down 3-0 in the series, it was a risky choice. But the data revealed that the Sonics consistently outscored the Bulls when Brickowski was on the court. And, in fact, when he entered the starting lineup, the Sonics went on to win the next two games before ultimately losing the series.

Since then, of course, player salaries have continued to rise along with sponsorships and broadcast revenues. At the same time, the Web has spawned legions of non-US sports betting sites (they’re illegal in the US). All that has no doubt kept developers and researchers up late at night honing a growing number of sports data mining applications.

Likewise, the number of sports these applications can analyze has grown to include everything from swimming to cricket. “Every sport has tons of variables. It is simply a matter of finding the ones that explain the most data,” said Robert Schumaker, an information systems expert at Iona College’s Hagan School of Business and author of a paper on using machine learning to pick winners at greyhound racing. “Once that has been performed, predictions can be made.”

**Team Think Tanks**

That begs the question, of course, which sports might be the best candidates for AI tools?

“AI techniques depend upon past patterns to predict the future and assume that the past is an indicator of the future. So given enough data, I think all of the sports ... would be equally easy,” says Ramesh Sharda, regents professor of Management Science and Information Systems and director of the Institute for Research in Information Systems at Oklahoma State University. Sharda adds one caveat: “The attributes of a game involving humans can be identified more easily because we have a reasonably good understanding of how humans think.”

Sharda recently cowrote a paper (doi:10.1016/j.eswa.2008.06.088) on using neural networks to select teams for cricket matches, specifically players who deserved a spot on 2007 World Cup teams from various countries, especially India. He and his colleagues analyzed prospective players’ performance by looking at data going back to 1985. Thus trained on the data, the neural network model made its selections, which in fact were validated by the actual team members picked for the event.

The truth is, AI techniques can evaluate team members in ways that simple statistics cannot. Schumaker explains, using a baseball example: Two players on different teams each have the same batting average. Players on Team A tend to get base hits more often than players on Team B. “Although both players perform the same actions day after day, the first player will have more runs batted in,” he says. The Team A player can thank his teammates for that fact. In reality, both players might be equally good, even though the Team A player had better stats than his competitor.

A finding like that could save a Major League Baseball team millions when pondering which players to acquire or release. And that’s just one of the ways AI could add to a sports team’s bottom line. Sharda sees another way, as a decision-support tool for agents or colleges looking for talent. His neural nets could pour through reams of stats to identify “stars before they become stars.”

That said, could AI tools geared to sports generate the kind of revenues hedge fund managers have been known to make? Perhaps. And that’s because there’s at least one huge difference between sports and finance when it comes to the AI tools they employ. Financial analysis is so competitive that any tool that appears to give someone an advantage will be quickly copied or nullified so that the advantage is fleeting. With comparatively fewer sports prediction tools being used today, there’s liable to be a longer-term advantage for anyone developing a new tool. Moreover, sports prediction tools “are fairly domain insensitive,” Schumaker explains. Meaning, if your model can pick winners in hockey, there’s a decent chance it will work for NFL games, too.

With hundreds of Web sites advertising programs to pick sports contests, the gambler in all of us might well wonder, do these tools really work? Schumaker’s peer-reviewed paper suggests they can work quite well. As Schumaker recalls, “I spent several evenings at Tucson Greyhound Park tweaking the wagering schema to use through ‘empirical testing.’ On the third night, I got the validation of results I was looking for and ended up pocketing a 24 percent return.” Twenty-four percent—that’s good even by hedge-fund standards. So did Schumaker quit his day job and move to Vegas? It seems gambling fever was no match for academic curiosity: “Once the discovery was made and the problem solved, I published a paper on it and moved on to new problems,” he says. And he hasn’t been to the track since.