Crime Mapping and Analysis Beyond GIS: An Introduction to Spatial Crime Data Mining

By Albert K. Yeung, PhD, O.L.S., O.L.I.P., Ontario Police College

DEFINITION OF DATA MINING

In technical literature, data mining is commonly called or used in conjunction with terms such as knowledge discovery in databases (KDD), business intelligence (BI), exploratory data analysis (EDA), predictive analytics (PA) and decision support systems (DSS), among many others. In practice, these terms usually carry different meanings, and are used in different contexts for the same or similar set of techniques and applications. For the purpose of this paper, the term data mining is used to embrace all these variant terms, and is defined generically as the exploration and analysis of huge amounts of data in order to detect meaningful rules, patterns, trends, insights, and relationships that are previously unknown and potentially useful for decision making.

The notion of "previously unknown" or "hidden" knowledge is central to the definition of data mining. This particular characteristic practically distinguishes data mining from conventional statistical analysis. Users of statistical analysis always have certain assumptions, called hypotheses, about the data to be analyzed, and the aim of the analysis is to validate or refute these assumptions. On the contrary, users of data mining invariably start the knowledge discovery venture without any *a priori* assumptions about the data or any ideas of what will come out as a result. In other words, whereas conventional statistical analysis is used to confirm or reject the assumptions about a particular data set, data mining aims explicitly to explore a data set for previously unknown or hidden patterns, trends, relationships and anomalies.

DATA MINING TECHNIQUES

In essence, data mining is not an independent and well-defined discipline in its own right, but is rather the confluence of many existing disciplines such as database management, information science, statistics, artificial intelligence, scientific visualization, pattern recognition, and optimization techniques, among many others. It is often perceived as one of the phases of the process that aims to turn raw data into useful information, knowledge and intelligence as shown in Figure 1. Besides, contrary to popular belief, data mining does not depend entirely on automated techniques. Depending on specific mining methods used, human knowledge and skills may play a crucial role during data mining, and are essential for the interpretation, evaluation and deployment of data mining results. In this regard, it is more appropriate to look at data mining as an interactive human-computer interface for knowledge discovery rather than as an exclusively automated information extraction methodology (Yeung and Hall, 2006).

Generally speaking, data mining techniques can be classified into the following five major categories:

1. *Entity extraction,* which identifies particular patterns and clusters of similar characteristics from large data sets.
2. *Association rule mining,* which detects the relationships and sequences among frequently occurring data items and establishes the relationships as rules, for example, if A occurs, there is a 90% chance that B will also occur.
3. *Predictive modeling,* also called *trend* or *regression analysis,* which aims to forecast what will happen in future on the basis of analyzing data of the past.
4. *Deviation detection* or *outlier analysis,* which uses specific measures to identify data items that differ exceptionally from the rest in the data set.
5. *Pattern visualization,* which uses animated graphics to describe dynamically the distributions, movements, connections and interactions of the investigated subjects over space and over time.

Figure 2 summarizes where the five categories of data mining techniques noted above are applicable with respect to major categories of crimes.

SPATIAL MINING OF CRIME DATA

The concepts of spatial data mining are distinctly different from those for conventional alphanumeric data mining. This is because spatial data are characterized by their high dimensionality, spatial autocorrelation, heterogeneity, complexity, ill structure and dependency on scale. In this context, high dimensionality means that spatial data have up to four dimensions (i.e., x, y, z, and time) of information space for the measurement and correlation of other dimensions (e.g., types of crime incidents, characteristics of offenders, the nature of victimization, environment of crime scene, etc.). Computation and representation of spatial crime information require an implied topological and geometric measurement framework that affects the patterns, trends and movement of criminal activities. The concept of spatial autocorrelation is founded on Tobler's first law of geography, which states, "everything is related to everything else, but nearby things are more related than distant things." Such a concept is in sharp contrast to the assumption of discrete and independent existence of individual data items in conventional statistical analysis.

The concept of spatial data mining is further complicated by the heterogeneous nature of spatial data, which means that all spatial features and phenomena tend to vary by location and over time, and that it is practically impossible to describe using a universal estimate of parameters. Spatial data are said to be complex because the spatial features and phenomena that they represent vary in size, shape, boundary property and direction, which in turn affect their topological relationships with one another. Spatial data are invariably ill structured because, unlike alphanumeric data stored in forms and tables, they have no obvious order or organization that will facilitate search, interpretation and presentation. Finally, there is the issue of scale, which affects the level at which spatial data are aggregated and generalized. An experiment that was conducted using identical data mining techniques but with data at different scales, and hence at different levels of aggregation or generalization, showed that this can sometimes lead to contradictory results.

Spatial mining of crime data, therefore, requires special techniques. Conventionally, there are two distinct approaches to spatial data mining. The first one is to design and develop new spatially enabled algorithms that are particularly adapted to explore data with location attributes. A relatively rich set of spatially enabled algorithms is now available for use in spatial data mining. Examples include spatial regression, spatial association rules (SAR), Markov Random Fields (MRF), Bayesian classifier, co-location rules, spatial clustering and spatial outlier detection methods. The second approach is to explicitly model spatial properties and relationships in a preprocessing phase before using classical data mining algorithms. Santos and Amaral (2004), for example, described spatial dependencies (including direction, distance, and topological relations) through inferred rules, and then applied the classical data mining methods called decision tree and Self-Organized Map (SOM) to the resulting spatial model. A fair number of research articles on similar approaches to spatial data mining can be found in learned journals in geographic information science, computer science, artificial intelligence, and decision science.

An emerging trend in spatial crime data mining is to use a visual approach. Visual spatial data mining is part of the growing discipline of geovisualization (GeoViz). This is a special branch of information visualization that integrates the concepts and methods of exploratory spatial statistics, image analysis and pattern recognition, cartography, and geographic information science, among many others, to provide the theories, methods and software tools for visual data exploration, analysis, synthesis and presentation, and for the extraction of spatial knowledge and intelligence.

Moffatt (2006) demonstrated that GeoViz techniques could be effectively incorporated into current philosophies, strategies and practices of policing and law enforcement. The visual approach does not only allow crime data to be analyzed spatially with respect to where crime incidents have taken place and will likely to take place, but also temporally with respect to when crime incidents have occurred and how their patterns have changed and will possibly change over time. What is more important is the ability to integrate the analysis of crime data with land use and socio-economic data which, according to the theories of the ecology of crime and environmental criminology, are closely associated with certain types of crimes. Visual data mining techniques, therefore, provide crime analysts and criminal investigators a set of useful tools to identify, interpret and compare criminal activities and their underlying causes over space and over time.

DISCUSSION AND SUMMARY

The construction of knowledge from huge amounts of crime data is a highly complex and computationally intensive process. A variety of computer programs have been or are being developed that aim to help crime analysts and criminal investigators locate crime hot spots, spatially relate potential suspects to actual crime incidents, profile serial crime geographically, and understand the interrelationships between criminal activities and the socio-demographic characteristics of the community. Advances in spatial data mining techniques appear to be particularly promising, in view of their analytical power over conventional statistical methods and as evidenced by their successful deployment in various fields of business, science and technology.

Do sophisticated spatial data mining techniques really work for crime data? Apparently, the answer is not a simple "yes" or "no." Supporters of these techniques can easily cite tens of success stories to demonstrate how spatial data mining has been used to provide insight into crime and criminals not available before. On the other hand, critics remain skeptical, contending that existing spatial data mining methods have not yet been adequately tested to prove their superiority over the intuition of a detective or a crime analyst. Some people also argue that computer modeling of human behaviours is not a precise science because these behaviours are impossible to quantify and simulate. No mathematical models, no matter how elegant, are capable of representing and accounting for the complex and intricate interplay of all the variables and factors pertaining to human decisions and actions in criminal activities.

There are good reasons to be optimistic and skeptical about the use of new technologies for crime data analysis. However, the future of the use of technologies is not dependent on what people view from their own vantage points, but on how they can work together to make technologies work. In this context, there are two basic issues that need to be addressed. One is the accuracy and completeness of the data obtained by police departments. The results of any spatial analysis techniques, as we all understand, can only be as good as the input data. In crime analysis and criminal investigations, knowledge obtained from erroneous and incomplete data is often worse than having no knowledge at all. The other issue is that not all crime analysts and criminal investigators are necessarily familiar or experienced enough with sophisticated computer programs developed to help them solve problems. The use of spatial data mining techniques is never a simple plug-and-play task. It requires education and training to understand the concepts and master the techniques of using the software and interpreting the results.

In summary, police departments contemplating the implementation of spatial data mining technology have to realize that it is not simply a matter of acquiring hardware and software, but it also necessitates significant investment in data and people as well. The use of technology alone will not solve crime analysis problems, but the combined power of time-tested technologies, accurate and complete data, and adequately trained personnel will.

REFERENCES

1. Chen, H., Chung, W., Xu, J.J., Wang, G., Qin, Y. and Chau, M. (2004) "Crime Data Mining: A General Framework and Some Examples," *IEEE Computer,* Vol. 37, No. 4, pp. 50-56.

2. Moffatt, F. (2006) "Geovisualization of opportunity – Making the invisible, visible," a panel presentation delivered at the Crime Mapping and Analysis Seminar, April 19th and 20th, 2006, Ontario Police College, Aylmer, ON.

3. Santos, M.Y. and Amaral, L.A. (2004) "Mining geo-referenced data with qualitative spatial reasoning strategies," *Computers and Graphics,* No. 28, pp. 371-379.

4. Yeung, A.K.W. and Hall, G.B. (2006) "Chapter 11: Spatial data mining and decision support systems," *Spatial Database Systems: Design, Implementation and Project Management,* Dordrecht, the Netherlands: Springer Science+Business Media.

Acknowledgement and disclaimer: The author wishes to thank the Director of the Ontario Police College for permission to publish this paper, which is based on the author's presentations at the Crime Mapping and Analysis Seminar, held in April 2006 at the Ontario Police College. The opinions expressed are the personal views of the author only, and do not in any way reflect or represent the views or policies of the Ontario Police College, the Ministry of Community Safety and Correctional Services, or the Government of Ontario.

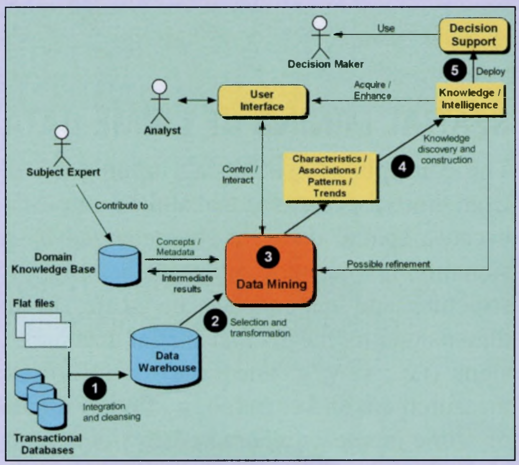
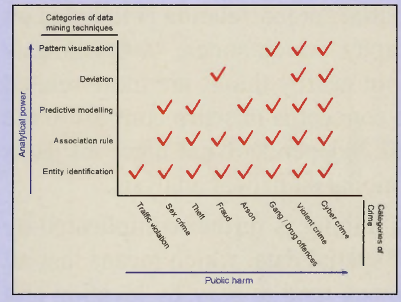


Figure 1: Data mining as one of the phases in the process of extracting knowledge and intelligence from databases (Source: Yeung and Hall, 2006)



Categories of data mining techniques	Traffic violation	Sex crime	Theft	Fraud	Arson	Gang / Drug offences	Violent crime	Cyber crime
Pattern visualization						✓	✓	✓
Deviation				✓			✓	✓
Predictive modelling		✓	✓		✓	✓	✓	✓
Association rule		✓	✓	✓	✓	✓	✓	✓
Entity identification	✓	✓	✓	✓	✓	✓	✓	✓

Figure 2: Application of data mining techniques in the analysis of different categories of crime (Source: Chen et al., 2004)