Open Access Library Journal
2023, Volume 10, e9685
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Using Data-Mining to Solve Crimial Cases

Tongxing Li, Yongfeng Zhang, Jiaojiao Zhao, Hui Zhang

School of Mathematics and Statistics, Taishan University, Tai'an, China
Email: tsultx@126.com

How to cite this paper: Li, T.X., Zhang, Y.F., Zhao, J.J. and Zhang, H. (2023) Using Data-Mining to Solve Crimial Cases. *Open Access Library Journal*, **10**: e9685.
https://doi.org/10.4236/oalib.1109685

Received: December 14, 2022
Accepted: February 6, 2023
Published: February 9, 2023

Abstract

Crime is a serious social problem, which affects the improvement of the quality of life and the development of social economy. Under the background of the information age, how to effectively use the collected large amount of crime data to analyze and predict crime is particularly important. Data-mining can extract implicit, unknown and potentially valuable information from a large number of original data. This paper studies how to use data-mining to analyze the intelligence value behind massive data, which has important reference value for the construction of crime detection and control model.

Subject Areas

Artificial Intelligence, Big Data Search and Mining, Criminology

Keywords

Data-Mining, Link Analysis, Machine Learning, Artificial Intelligence

1. Introduction

The well-known methods and tools used in data-mining include link analysis, such as looking for gangs and other forms of links between criminals or terrorists; software agents are small and independent computer program fragments that can monitor, collect, analyze and act on information; machine learning is algorithms that can infer the contour characteristics of crime and the distribution map of crime; neural network is a special kind of computer program, which can predict the probability of crime or terrorist attack. Geometric clustering is a special form of link analysis [1]. In the era of big data, a major difficulty in predicting crime is how to accurately and effectively analyze a large number of crime data. In addition to the basic information mastered by the police, it also involves the information of relevant industries such as network, communication, finance and transportation, as well as the relevant social information such as e-commerce,

logistics and transportation, social services, and so on. How to integrate massive information and find valuable clues is very important.

In the following, based on big data mining technology, we give four effective ways to analyze criminal activities. The innovation of the paper is the general theory of "Link analysis", "Geometric clustering", "Software agent" and "Software agent" were introduced and the relative theory of the methods used in the data-mining to solve criminal cases is introduced. However, due to the difficulty in obtaining judicial data, the analysis of some cases using data-mining will appear in future articles.

2. Link Analysis

Link analysis, also known as "point connection", is one of many branches under the heading of data-mining. Data-mining is to obtain useful information from a large amount of public data that can be provided by modern society. Link analysis is mainly the process of tracking the relationship between people, places and organizations. These links may be business relationships, criminal partnerships, family ties, direct meetings, financial transactions, e-mail exchanges, and so on. Link analysis plays an important role in the fight against terrorism, organized and purposeful crime, money laundering and telephone fraud, especially.

Link analysis is a human expert driven process. Mathematics and technology provide human experts with flexible and powerful computing tools, which makes it easier to reveal, track and study possible connections. Those programs generally allow analysts to form connected data into a network that can be displayed on the computer screen for research. Nodes on the network represent interested individuals, places and organizations, and connections between nodes represent relationships or transactions. This tool also allows analysts to investigate and record the details of each connection, and find new nodes associated with existing nodes or new connections between existing nodes.

In the investigation of a suspected criminal group, investigators can link up the phone calls that the suspects have played or receive, analyze the number of calls, phone records, the duration and duration of each call, or the next dialing number. Then investigators can decide to follow up the phone network to see who the phone is calling from and who they are from, to see who had previously talked to the original suspect. Through this investigation, investigators can pay attention to those who did not pay attention before. Some of them are likely to be proven innocent, but others may be proven to be accomplices or accomplices of criminals. Another investigation path is to track the cash flow between domestic and foreign bank accounts of suspected criminal groups.

Another path is to analyze the network composed of people and places visited by suspects. Such as the records of purchasing air tickets, train tickets and entry-exit ports in and out of a specific country, credit card shopping records, car rental records, records of accessing websites and such data.

In today's society, it is almost difficult for anything to leave electronic traces. The challenge of link analysis is usually not insufficient data, but how to select

what effective information from millions of data for further analysis. Link analysis can play an important role when supported by other types of information, such as relevant information from potential suspects' neighbors or useful information provided by police informants. The advantage of link analysis is that once the initial link analysis has identified a possible criminal or terrorist network, it can determine the key people suspected of crime by studying who these people have the most contact with in the network.

3. Geometric Clustering

Under the condition of limited resources, law enforcement departments usually devote most of their energy to solving major criminal cases, but some small illegal cases are easy to be ignored. However, if a criminal gang or individual regularly creates similar cases for many times and accumulates to a certain number, it will become a major criminal activity, which will attract the special attention of the police. How to find out which are the serial crimes of a group or individual from the large number of minor violations that occur every day is extremely significant.

In the case of distraction theft, generally, one person appears around a house owner, pretends to be some kind of staff to communicate with the house owner, and another person quickly sneaks into the apartment or room to steal. Such victims usually call the police, and the police officers in charge of peripheral investigation will go to the victim's house to listen to the statement. Because one of the perpetrators has communicated for a long time to attract the attention of the homeowner, the victim's statement often contains more details, including gender, body shape, height, approximate age, face, accent, special accessories, number and gender of partners, etc. This valuable information makes the criminal cases of this nature very effective for data-mining judgment. It can determine that this group of cases is related to a criminal gang, and it also plays a key role in the analysis of using geometric clustering technology.

When we really use data-mining to practice, we need to face more complex situations. First of all, most of the content of the description of the offender is recorded by the police officer in charge of the investigation in the form of narrative statements when listening to the statement of the perpetrator. It is necessary to use text-mining technology to transform such description into an organized form. In practice, there are many limitations in the available text-mining software, which often requires manual input to process a considerable number of records. After some initial analysis, researchers usually focus the main information on eight variables: height, body shape, age, race, hair length, hair color, accent and number of associates. Once the data is processed into an organized format, then geometric clustering is used to divide the description of criminals into several sets, which may point to the same person.

Specifically, the above eight variables are numerically coded in turn. Height may give an approximate height (meters) or a range, or words such as "medium",

"high" and "short", which requires some strategy to convert its corresponding into a single number. Age is often estimated and can be recorded as a number or a range. Gender is male or female, usually coded as 1 or 0. Similarly, some schemes need to be designed to express the remaining variables in digital form. After the numerical coding is completed, an eight dimensional vector is used to describe each perpetrator. At this time, the coordinates of a point in an eight dimensional Euclidean geometric space are the characteristics of a perpetrator. Using the description of the distance between two points in Euclidean space, in the sense of this measurement, the next point corresponds to the approximate common characteristics in the description of the perpetrator; The closer the distance is, the closer the point is, the more common features are described. At this time, the distance between Euclidean two points is given by:

$$d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_8 - y_8)^2} \quad (1)$$

The key point is how to identify the clustering of adjacent points. If there are only two variables, mark all points on a plan with only x and y coordinates, and the possible clustering can be seen by visual inspection. Unfortunately, it is quite difficult to find clusters in an eight-dimensional space. The effective method is to transform the array of points in the eight-dimensional space into a two-dimensional matrix, that is, arrange all data points into a two-dimensional grid. The arrangement rules are as follows:

1) Place a pair of adjacent points in the eight-dimensional space into the same grid;

2) Any pair of adjacent points in the grid are also adjacent in the eight-dimensional space;

3) Points that are far away in the grid are also far away in eight-dimensional space.

In practice, we can use Kohonen self-organizing map in neural network to arrange the data according to the above rules. After the data is input into the grid, the law enforcement personnel analyze the grid box. These data may come from a criminal gang responsible for this series of cases. At the same time, it can simply identify the clusters on the grid, which is likely to represent the activities of criminal gangs. Therefore, in both cases, the police can analyze the table value of the corresponding case statement and dig out the cases that are actually committed by a gang.

The disadvantage of geometric clustering is that the initial digital coding of case table value may not be standard. Therefore, when using the distance of eight-dimensional vector in Euclidean space to cluster table values, the size of one variable may play a major role, while the size of other variables has little effect. The scaling of each variable is an improvement in the process of normalization. Another problem is how to deal with the missing data, that is, how to cluster if there are missing (blank) entries. Missing data is one of the biggest obstacles in data-mining. Usually, if there are only a few such cases, the secondary entries can be directly ignored.

4. Software Agent

In essence, software agent is a specific computer program designed to achieve a set goal. When operating environment changes, the program will respond independently. Software agent can make various operations according to different input instructions in a certain range. It is one of the specific applications of artificial intelligence. For special types of criminal cases, it is impossible for the police to collect a large amount of data and analyze the results, so as to detect the sudden change of the situation and respond as soon as possible. Therefore, it must be assisted by software. Usually, the countermeasure used in practice is to develop a coordination system of multiple agent software, in which each agent software communicates with each other, and each agent software is set to complete a specific subtask. The coordination system mainly includes the following commonly used agent software:

1) Agent software that extracts and modifies data from different data.

2) Agent software that collects potentially relevant data from different databases.

3) Agent software that analyzes the data and find abnormal patterns for specific events.

4) Agent software for classification and identification of abnormal conditions.

5) Agent software that provides alerts to law enforcement personnel in an emergency.

5. Machine Learning

Machine learning is another important application of artificial intelligence. It is the most effective in data-mining technology when it is used for the contour analysis of criminals. The effectiveness of algorithms in machine learning is that they can automatically find and identify the key features in the sea volume data. Specially trained staff can also do fine identification and classification, but they can only process a small amount of data at a time, while machine learning can process a large amount of data, so as to save a lot of manpower and material resources.

Computer can make scientific decisions without human intervention through machine learning. It has been successfully applied in the fields of speech recognition, auto-driving, network search, and so on. Using machine learning, crime can be predicted based on historical reference data, so as to make up for the shortcomings of the traditional crime governance model, and open up a new concept of crime governance. Using machine learning to predict crime has become a research hotspot abroad, but domestic research in this field has just begun. The prediction process of crime by machine learning generally includes data collection, data classification, pattern recognition, prediction process and data visualization. Machine learning can use unstructured data and structured data for pattern recognition. Structured data is mainly used for association analysis, classification and prediction, cluster analysis and outlier analysis. Common ma-

chine learning algorithms include reinforcement learning, semi supervised learning, unsupervised learning and supervised learning. Classification is an important application of machine learning algorithm in crime prediction. Its process includes two stages: using training set to train classifier and prediction results. It is very effective to predict crime hotspots and crime types through classification.

6. Summary

Data-mining technology is usually used to efficiently explore the hidden patterns in a large number of crime data. By continuously improving the efficiency of crime data-mining, the accuracy of crime prediction can be improved accordingly. In [2] [3] [4] [5], in order to get an ideal crime prediction model and give a satisfactory conclusion of crime data analysis, so as to predict crime more accurately, we need enough historical data to train and optimize the model. In [6], it is pointed out that unfortunately, in the process of practice, the occurrence of criminal cases is often affected by a variety of factors, and each crime prediction model has certain disadvantages. A large number of updated crime data cannot be summarized and integrated in real time, and the prediction object is uncontrollable. Researchers need to put forward more improved and optimized algorithms and crime prediction models to improve the accuracy of crime prediction [7] [8] [9] [10].

Acknowledgements

The authors would like to thank the associate editor and the reviewers for their constructive comments and suggestions which improved the quality of the paper. This work was supported by the Support Plan on Science and Technology for Youth Innovation of Universities in Shandong Province (2021KJ086).

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Devlin, K. and Lorden, G. (2007) The Numbers behind Numbers: Solving Crime with Mathematics. Penguin, London.

[2] Hosseinkhani, J., Koochakzaei, M., Keikhaee, S., *et al.* (2014) Detecting Suspicion Information on the Web Using Crime Data Mining Techniques. *International Journal of Advanced Computer Science and Information Technology*, **3**, 32-41.

[3] Kennedy, L.W., Caplan, J.M., Piza, E.L., *et al.* (2016) Vulnerability and Exposure to crime: Applying Risk Terrain Modeling to the Study of Assault in Chicago. *Applied Spatial Analysis and Policy*, **9**, 529-548. https://doi.org/10.1007/s12061-015-9165-z

[4] Baumgartner, K., Ferrari, S. and Palermo, G. (2008) Constructing Bayesian Networks for Criminal Profiling from Limited Data. *Knowledge-Based Systems*, **21**, 563-572. https://doi.org/10.1016/j.knosys.2008.03.019

[5] Babakura, A., Sulaiman, M.N. and Yusuf, M.A. (2014) Improved Method of Classification Algorithms for Crime Prediction. 2014 *International Symposium on Biometrics and Security Technologies*, Kuala Lumpur, 26-27 August 2014, 250-255. https://doi.org/10.1109/ISBAST.2014.7013130

[6] Chen, H., Chung, W., Xu, J.J., *et al.* (2004) Crime Data Mining: A General Framework and Some Examples. *Computer*, **37**, 50-56. https://doi.org/10.1109/MC.2004.1297301

[7] Nath, S.V. (2006) Crime Pattern Detection Using Data Mining. 2006 *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops*, Hong Kong, 18-22 December 2006, 41-44. https://doi.org/10.1109/WI-IATW.2006.55

[8] Wang, T., Rudin, C., Wagner, D., *et al.* (2015) Finding Patterns with a Rotten Core: Data Mining for Crime Series with Cores. *Big Data*, **3**, 3-21. https://doi.org/10.1089/big.2014.0021

[9] Rasekh, A.H., Liaghat, Z. and Mahdavi, A. (2012) Predict Edges in Fliker Social Network Using Data Mining Method. I*ntelligent Information Management*, **4**, 60-65. https://doi.org/10.4236/iim.2012.43009

[10] Li, J., Peng, W., Tao, L., *et al.* (2014) Social Network User Influence Sense-Making and Dynamics Prediction. *Expert Systems with Applications*, **41**, 5115-5124. https://doi.org/10.1016/j.eswa.2014.02.038