Application of Data Mining Technology in the Analysis of Criminal Laws

Ying Zhou, Pancheng Li

Jiangxi Police College, Nanchang City, Jiangxi Province, 330003, China

Keywords: big data; data mining; criminal law; association rules

Abstract: At present, the public security organizations in China have carried out a new revolution in the modern police service of public security informatization. The construction of a public security information system has become an inevitable trend in the fight against illegal criminal activities. However, information construction work for many years has accumulated a large amount of original data. The large scale of data and complex data structures have become a difficult problem for public security information system. Based on the author's learning and practical experience, this paper first analyzed the data mining technology and its association rules, and then studied the application of graph data mining in criminal laws. Finally, the article summarized the data mining process.

1. Introduction

The informatization construction has made rapid progress and established a vertical and horizontal public security information network. The police services have fully realized digital management and accumulated a large amount of basic business data. Some comprehensive applications are also gradually carried out. However, the basic business data which is accumulated in the work at present is only used for simple and primary applications, such as query, update and statistics[1]. How to find out the regular information behind the data through data mining technology, provide services for various management tasks, offer scientific basis for leadership decision-making and supply technical support and reference for actual combat of public security is worthy of further discussion.

2. Data Mining Technology and Its Association Rules

2.1 Data mining technology.

Data mining is one of the most cutting-edge research directions in database and information decision-making in the world. The Gartner Report lists five key technologies that will have an important impact in the industry in the next three to five years. Among them, knowledge discovery in database (KDD) and artificial intelligence rank first. At the same time, this report includes parallel computer architecture research and KDD in the 10 new technology areas that companies should invest in over the next five years. Data mining can be divided into three phases: data preparation, data mining and interpretation of results. At present, data mining technology has been applied in many industries and has achieved certain results, including astronomy, biomedicine, health care, DNA analysis, finance, insurance, retail industry and telecommunications[2]. Early data mining applications focused on helping companies improve their competitiveness. With the popularity of data mining, its application areas are also expanding. The development of the information industry provides a broad space for data mining.

2.2 Association rules.

Based on the information of offenders and criminals, security police can analyze the relationship between the age, education level, occupation and other characteristics of the offenders and the occurrence of certain types of cases, and the relationship between the types of the case and the location of the case, the time of the case, the object and the characteristics of the crime, so as to identify high-risk groups, high-risk locations and high-incidence time. Association rules are the

  DOI: 10.25236/dpaic.2018.019

most typical things in many types of knowledge in data mining. In 1993, Agrawal first proposed it to find the customer purchase model in the sales of goods when analyzing the market basket problem. Later, many researchers conducted a lot of studies on the mining of association rules. They optimized the original algorithm, such as introducing random sampling and parallel thinking, to improve the efficiency of algorithm mining rules. Because the association rules requires to mine and process a lot of data, the common methods includes the reduction of the number of database scans, division, sampling and incremental techniques to improve the efficiency of the algorithm. Databases are usually large in scale and are often distributed across several sites. Parallel algorithms can improve the efficiency of mining. In many applications, valuable association rules between data items often appear in relatively high concept layers, and it is difficult to find useful association rules in lower concept layers[3].

3. Application of Graph Data Mining in Criminal Laws

3.1 Data preparation.

This paper derived the detailed information of all the criminals in a certain area since 2000 from the database of the national crime management system in the public security network. The total amount of data has reached several hundreds of thousands of items, running in the Oracle9i database. The main field names in the system of criminal offender information are name, gender, birthplace, date of birth, ethnicity, education level, occupation, address, case type, height, nickname, expertise, body type, face feature, accent, foot length, footwear size, body surface marking, tooth characteristics, special features, means of committing a crime, criminal tools, crime characteristics, selection of time, selection of premises, selection of objects, selection of items, disposal of stolen goods, fleeing categories and fleeing range. The table serves as the main analysis object in this paper[4].

3.2 Data extraction.

According to the actual needs of the research, the relevant data is taken from the above data preparation stage to establish data tables for data analysis and processing: the personnel information table (ryxxb) and the case information table (ajxxb). Because of the limited space, data records are omitted. The basic structure is shown in Table 1 and Table 2.

Table 1 Personnel information table (ryxxb) sample data

Date of birth	Educational level	Occupation	Case type
19711024	60	170	18
19710875	70	040	19

Table 2 Case information table (ajxxb) sample data

Case type	Selection of time	Selection of location	Selection of object	Means of committing a crime	Location of committing a crime
18	43	627	21	6102	51
19	31	451	22	2304	03

3.3 Data arrangement.

(1) Data arrangement of personnel information table (ryxxb). For the sake of simplicity, only a portion of the sample data is extracted. Table 1 lists two records for reference. In order to facilitate the conversion and analysis of graph data, the date of birth in Table 1 is converted to age, which is generalized into youth, middle age and old age; the educational level is generalized into four categories: college or above, secondary school or high school, junior high school, and elementary school and below. Unqualified data in the table is cleared, such as empty content and error content.

(2) Data arrangement of the case information table (ajxxb). Table 2 shows some sample data of

the case information table (ajxxb). The six related fields in Table 2 are all dictionary entries, such as case 18 → hijacking case, selection of time 43 → noon, selection of location 627 → office, selection of object 36 → female youth, means of committing a crime 2304 → climbing from the window, the characteristics of the crime 51 → committing a crime in the neighbourhood. Due to 308 categories of the means of committing a crime, the item was generalized and the article took the main categories. Unqualified data in the table is cleared, such as empty content and error content.

3.4 Data graph representation.

In order to graphically represent the information about the offender and the crime, the apex is used to represent the offender, and the node information on the map represents the case. If the perpetrator has the same information on the time, selection of location, selection of object and means of committing a crime, then there is naturally a side connection, indicating that the two offenders may have contacts or gangs. If a, b, c and d respectively represent theft, robbery, rogue and injury cases, and w, x, y and z respectively indicate the time, location, object and means of committing a crime, perpetrators with the same attributes in the case information table forms a picture, as shown in Figure 1.

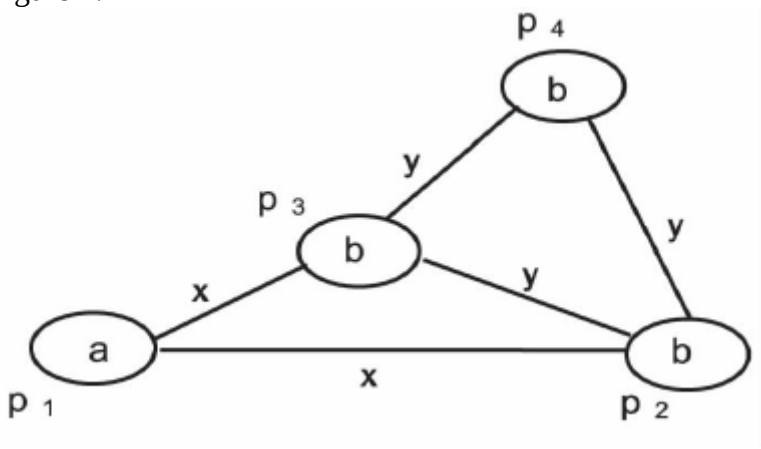

Figure 1 Association graph of criminals

The four apexes P1, P2, P3 and P4 in Figure 1 represent four different criminals. Among them, P1 and P2 represent thief and robber respectively, but the two criminals chose the same place; P2, P3 and P4 represent robbers with the same objects[5].

4. Data Mining Process

4.1 Graph data mining.

Analysis of the above two data tables can only analyze the relevant attributes of individual cases, such as theft, rogue, injury and robbery. This cannot find the core members of the case and the hidden laws behind the same crime information. If each crime fact is regarded as a node of the graph, the connection between criminal activities naturally forms a side, thus constructing a criminal network map. Many frequent subgraphs are mined in this map. These frequent subgraphs provide an effective basis for decision-making in identifying criminal activities and the dynamic trend of safety development. The existing public security information system cannot reflect the potential problems and alarms in these rich semantic graphs, cannot make correct predictions for the social security early warning system and cannot provide dynamic monitoring and scientific decision-making to maintain social security. The data graph GD normalization algorithm and the candidate subgraph algorithm can be referred to the literature. The following is a self-developed frequent subgraph algorithm FSubgraph to conduct experiment with case data in a city.

4.2 Experimental results and performance.

Core Duo 2.4 GHz CPU, 2G RAM, 160GB hard drive, Windows XP operating system and all

programs use Java language to run in Oracle9i database under JBuilder2006 development environment. The results and performance of algorithm GDMCR are shown in Table 3. After obtaining frequent subgraphs, the methods in the literature can be used to obtain association rules between frequent subgraphs to get knowledge. The support degree and confidence coefficient of association rules for frequent associated subgraphs are defined as follows:

$$\sup(B \to \mathrm{H}) = S(B \cup \mathrm{H}), \quad conf(B \to \mathrm{H}) = \frac{S(B \cup H)}{S(B)}$$

Here, B and H are respectively two frequent associated subgraphs, B ∪ H represents all node sets in B and H, and associated subgraph sets of all side sets in B and H. B→H means that if B is a frequent subgraph of a graph, then confidence coefficient that H is also the frequent subgraphs of the graph is conf(B→H).

Table 3 Experimental results

Support degree parameter (%)	The number of subgraphs	Operating time (s)	Memory consumption (MB)
5	40	10.1	42.1
10	22	2.1	20.5
20	14	1.2	10.7
30	9	0.7	9.8
40	5	0.4	7.0
50	0	0.3	5.9

The minimum support degree is set to 30% and the minimum confidence coefficient is 90%. The FSubgraph algorithm is used in the graph data of GD, which got 21 rules. Three valid rules are listed in Table 4.

Table 4 Case correlation data analysis results

Frequent subgraph coding	Support degree	Confidence coefficient	Encoding of subgraphs
axa	30%	90%	Both persons (the mark of both nodes is a) are thieves and chose the same crime location.
Bwbw0b	35%	92%	Three persons (the mark of three nodes is b) are robberies and the crime time is the same.
cycxyc	37%	93%	Three persons (the mark of three nodes is c) are rogues with the same object, and two persons chose the same criminal location.

5. Summary

The graph-based data mining algorithm is applied to the criminal law research and the article proposed the GDMCR algorithm to mine criminal laws based on the frequent subgraph structure with the same criminal characteristics. The algorithm can find laws and characteristics of high-risk groups, high-risk locations and core members in criminal network, which provides a powerful basis for public security organizations to detect cases to improve law enforcement efficiency and rapid response capability. This will become an important research direction for public security organizations to analyze criminal laws. How to combine it with other data mining and artificial intelligence methods, such as decision trees, neural networks, genetic algorithms and association rules mining, to analyze the case and explore the laws in the activities of criminal organizations, will be the future work.

References

[1] Ma Kai. *A Research on New Investigation Techniques in the View of Big Data* [J]. *Journal of Mudanjiang University*, 2018,27(10):26-29.

[2] Chen Changfeng. *A Research on Economic Investigation and Auditing Technology in the Context of Big Data* [J]. *Journal of Kaifeng Institute of Education*, 2018, 38(09): 295-296.

[3] Hu Sumeng. *A Research on Crime Prediction in the Context of Big Data and Cloud Computing* [J]. *Legal Vision*, 2018, (24): 164.

[4] Wang Long, Sun Bin. *An Applied Analysis of Public Security Case Database Based on Data Mining* [J]. *Gansu Science and Technology*, 2018, 34(13): 6-10.

[5] Wang Bin. *A Research on Big Data Method in Criminal Investigation* [J]. *Jingyue Journal*, 2018, (02): 67-75.