Research on Construction of Key Personnel Judgment Model Through Data Label

Wei Li¹, Changmin Chen¹* and Yi Li²*

¹Wuhan FiberHome Digital Technology Co., Ltd, China Information Communication Technologies Group Corporation, Wuhan, China
²School of Mathematical Sciences, Huaibei Normal University, Huaibei, China
Email: liweidata@foxmail.com; cmchen@fhzz.com.cn; Whut_ly_forever@163.com

Abstract. Key personnel are important business objects in the management and control of public security elements. Through big data and machine learning technology, it is of great significance to model the years of police’s experience on research and judgment for the actual combat and the construction of Public Security Prevention and Control System. Aiming at the problem of complicated data type and low quantization degree of public security data, which makes it inconvenient to use machine learning for model analysis directly, a method of constructing public security personnel label system is proposed in this paper to transform complex business data into quantized label features, and on this basis, a key personnel judgment model is constructed by using FP-Growth and XGBoost algorithm. Through comparative analysis, the research results show that the judgment model proposed in this paper is better than the traditional business integral model and logistic regression algorithm in Accuracy, Precision and Recall, which can better serve the actual police combat.

Keywords. Key personnel; XGBoost algorithm; FP-Growth algorithm; label system.

1. Introduction

As China’s socialist construction enters a new era, the complexity and difficulty of public security management become increasingly prominent, public security and public order face new challenges. General secretary Xi Jinping, at the Central Conference On Political And Legal Affairs and the Seminar For Leading Officials At Provincial And Ministerial Levels Focusing On Preventing and Defusing Major Risks By Sticking To The Bottom Line, stressed that “we should accelerate the construction of a three-dimensional and information-based public security prevention and control system, maintain a strong deterrent against criminal offenses, and enhance people’s sense of security”, and he further clarified the direction and point of the public security prevention and control work. Key personnel are important business objects in the control of public security, and there are potential threats to the stability of social security. From the historical work experience of frontline police, it can be easily seen that the key personnel’ crimes present common characteristics like escaping crime, gangs, same household registration and etc. Therefore, it can help the police to dig out potential key personnel and prevent risks in advance by analyzing personnel’s behavior characteristics with certain technical means.

Many domestic and foreign scholars keep doing research on key personnel control and social risk prevention. Saarikkom and Kivivuori explored the influencing factors of juvenile delinquency through questionnaire survey and multivariate logistic regression test, it was found that traffic congestion, alcohol abuse, single father and low education were all important factors to make them hostile to the police [1]. Based on the fuzzy k-means clustering method and the actual data on the Indian government
website, Tomar and Manjhvar studied the growth and influencing factors of crime, so as to better predict and prevent it in advance [2]. Dutta, Gupta and Narayan focused on the risk of credit card fraud and proposed a data mining method based on public detection and peak detection algorithms to detect credit card application fraudsters, which made up for the disadvantages of traditional manual judgment [3]. Al-ajlan and Ykhlef analyzed the data left by people’s interactions on social networks and smart phones, and introduced a suspicious behavior analysis model based on social networks to classified criminal behaviors, suspicious behaviors and normal behaviors [4]. Wei and Yang combined big data with personnel judgmental high risk of terrorism, and built indicators based on statistical analysis and business experience [5]. Cai conducted a comprehensive study on the risks of social drug users by using the comparative research method, investigation method, analytic hierarchy process and etc. He screened and determined evaluation indicators based on the Delphi method, calculated indicators weight through hierarchical judgment matrices and expert scores, and verified the accuracy of the risk assessment model [6]. Throughout the research status, scholars have little research on the quantitative characteristics of key personnel and crime risk, and there is not enough research on individual behavior evaluation at the micro level, which is the starting point of this paper.

Aiming at the problem of complicated data type and low quantization degree of public security data, which makes it inconvenient to use machine learning for model analysis directly, this paper firstly transforms complex business data into quantized label features by off-line big data computing technology, then builds the public security personnel label system, and constructs a key personnel judgement model based on FP-Growth and XGBoost, which solidifies the business experience of the police from the data level by using big data and machine learning technology. And it would assist the police to make rapid and accurate judgment in actual combat.

2. Model Design

2.1. A Calculation Method of Personnel Feature Labels Based on Off-Line Big Data Computing Technology

With acceleration of the public security informatization construction, the business data has become multiple and miscellaneous due to the rapid accumulation, so that simply listing the historical records cannot achieve the purpose of data governance and integration. Public security data label is a concept proposed in recent two years, it is the symbol representation and the characteristic value of public security business objects, which reflects the usefulness of public security business data. The public security data label system contains descriptive label attributes in multiple dimensions and depicts the characteristics for business objects, which can form the objects portrait and better serve the business application [7].

The business object of this article is people, especially the key personnel. By integrating mass data such as basic personnel information, key personnel information on record, and personnel trajectory information, ETL tool is used to extract relevant business data from various business systems to HDFS of the big data platform, which is loaded through Hive data warehouse to form a big data resource pool, and then label calculation is carried out on this basis. Label calculation is divided into two steps. First, feature precomputation, that is, extracting the features of interest from the original business data through mapping, statistical calculation and etc., then constructing the feature database as the attribute support of the actual label value; The second step is feature labeling, that is, by defining marking rules, features are extracted into label information, and personnel label database is built. Label calculation mainly depends on off-line computing. The construction process of feature and label database is showed in figure 1.

On the basis of the big data resource pool, with map-reduce computing technology, Hive data is loaded and feature computing rules are transformed to off-line calculation and analysis for SQL. The calculation rules include explicit rules and implicit rules. Explicit rules refer to the feature calculation rules that can map directly such as age and gender. Implicit rules refer to the calculation rules that require certain statistical calculations or logical transformations such as traveling or staying in same
hotel with key personnel. The feature calculation results are stored in Hive to form the personnel feature database. Then the appropriate threshold or rules are defined, and feature data is transformed into label data to form the personnel label database. Based on the label database in Hive, data is synced to application database such as Hbase and Elasticsearch, which could support the application like rapid search. By the way, features and labels above are stored in a wide table.

![Big Data Resource Pool](Hive)

- Feature Database (Hive)
- Label Database (Hive)
- Feature Calculation
- Label Calculation
- Application Database (Hbase, Elasticsearch)
- ETL Extraction

**Figure 1.** The construction process of feature and label database.

Business data is continuously generated every day, so the labels are constantly updated to ensure the accuracy and timeliness of data. The label increment strategy includes two forms: the full-volume run batch and the incremental run batch. The full-volume run batch apply to scene that original table data volume is not big. After the data access increment extraction task is completed, the feature database task and the label database task are called successively. New feature and label table is named with date suffix, then rename and delete the history table, remove the date suffix next. At last, label increment refresh process is completed. The incremental run batch apply to situation that original table data volume is large, which data is stored in partitioned tables. After the data access increment extraction task is completed, the feature database task is called to mark the incremental data of the big data resource pool, and the result is appended to the feature database. Then the feature database data is aggregated, so that the newly added features and historical features are merged to form a new feature database. The rest table transform steps are same as before.

The above data access task, feature tasks and label tasks are in sequence, so Oozie, a big data component, is used to configure the hierarchical structure of scheduling tasks to form a workflow, enabling smooth and effective execution of tasks, thus completing the calculation of feature labels and laying the foundation for the judgment model.

### 2.2. The Design of Key Personnel Judgment Model

In this paper, the key personnel judgment model algorithm is designed as figure 2.

The process of labeling data has been completed by off-line big data calculation, which transforms the data into a unified format. The number of people in the data is denoted as N, the number of labels is denoted as M, then the size of the data set is N×M. In order to improve the accuracy of the model, this paper firstly selects features through the FP-Growth association rule algorithm [8], and automatically eliminates the feature labels that are not highly relevant to the model. NUM(X) refers to the number of X in the dataset, NUM(Y) refers to the number of Y in the dataset, NUM(X∩Y) refers to the number of X and Y in the dataset, and N refers to the total number of the dataset, then:

**Support:** the probability that the subset X and the subset Y occur simultaneously in the dataset.

\[
support = \frac{P(Y|X)}{P(ALL)} = \frac{NUM(X \cap Y)}{N}
\]  

(1)
Confidence: the probability that the subset of X and the subset of Y occur simultaneously accounts for the probability that the subset of X occurs.

\[
\text{conf} = P(Y|X) = \frac{\text{NUM}(X \cap Y)}{\text{NUM}(X)}
\]  

(2)

Frequent item set: is the set of support greater than or equal to the minimum support.

Lift: the probability that the subset of X and the subset of Y happen at the same time accounts for the probability that the subset of Y happens. The Lift measures the strength of the association between frequent item sets X and frequent item sets Y. When the Lift is greater than or equal to 1, there is a strong correlation between X and Y, and when the Lift is less than 1, there is no correlation between X and Y.

\[
\text{lift} = \frac{P(Y|X)}{P(Y)} = \frac{\text{NUM}(X \cap Y) \cdot N}{\text{NUM}(X) \cdot \text{NUM}(Y)}
\]

(3)

In this paper, subset Y refers to the value of a certain key personnel label which is 1, subset X refers to the label combination that label value is 1. For example, Support refers to the ratio to the total that the number of key personnel labels with value of 1 and other label values equal to 1. The FP-Growth association rule algorithm is adopted. The input is N×M data set with minimum support. The minimum support is expressed as min_sup, and its range is (0, 1). In this algorithm, take min_sup=0.001. The output is \(\{\text{Label combination}, \text{A certain key personnel label, Confidence, Lift}\}\), written as \(\{X, Y, \text{conf, lift}\}\), the conf and the lift are both in (0, 1), \(\text{"X→Y"}\) indicates a strongly related rule. The algorithm calculation process is divided into the following steps.

Step 1-1: Select a certain key personnel type as the target. The labeled N×M data set is divided, and the label column corresponding to a certain category of key personnel is selected, marked as m, and the remaining m-1 column is marked as \(\{l_1,l_2,\cdots,l_{m-1}\}\).

Step 1-2: Construct FP-tree and frequent item sets. FP-tree is a specific structure that store the data set, and frequent item sets are found based on it, which is a collection of elements that are often found together.

**Figure 2.** The algorithm process of key personnel judgment model.
(1) Traverse N×M datasets, get a label set with a number of frequent item elements of 1, delete the label combinations less than the minimum support defined, and then arrange the remained items in descending order of their support.

(2) Traverse the N×M dataset, create the item header table (descending from the top), and the FP-tree.

(3) For each frequent item, find its conditional pattern base, recursively call the tree structure, and delete item that less than the minimum support. If it ended up with a tree structure of single path, then listed all the combinations directly. Otherwise, continue to call the tree structure until a single path is formed. Simultaneously calculated the conf and lift of each rule which is formed as the key personnel label m and the label combination that is selected from \(\{l_1, l_2, \ldots, l_{m-1}\}\).

Step 1-3: Select strongly related label combinations from frequent item sets. Select the rules as the final result from Step 1-2 that Y equals m, X contains only one element, and lift is greater than or equal to 1. Union those X together as a set, which includes k labels that are strongly related to label m, is written as \(\{I_1, I_2, \ldots, I_k\}\). It also is the input to the final judgement model.

Key personnel judgement model is a binary classification model. In this paper, XGBoost algorithm is used for model training. XGBoost is an abbreviation of Extreme Gradient Boosting, which is an improved optimization of GBDT [9]. Different from GBDT, which only uses the first derivative information, In order to balance the decline of the objective function and the complexity of the model, XGBoost performs the second-order Taylor expansion on the loss function, and adds the regular term to obtain the optimal solution beside the objective function, which are helpful to avoid overfitting. The algorithm first uses the training set to train a tree, and uses the tree to predict the training set to get the predicted value for each sample, then gets the bias between the prediction and the actual. Next, train the second tree to use bias as the target instead of actual. After the two trees are trained, the bias of each sample can be obtained again, then the third tree can be further trained, and so on. When the training times reach the preset value or convergence, stop training. The specific implementation steps of the judge model are as follows.

Step 2-1: Sampling. Select a certain category of key personnel label m as the prediction target. Model input is label sequence \(\{I_1, I_2, \ldots, I_k\}\), and there need to take a sample from this N×(k+1) dataset as follows.

(1) In reality, the number of key personnel is far less than the number of normal people. Therefore, any label m equal to 1 in the dataset should be taken out as a positive sample. The number is denoted as n.

(2) When the value of n is less than or equal to k, it indicates that the number of key personnel is too small and the model has a high degree of freedom, which is not suitable for modeling. Otherwise, n data are randomly selected from the dataset those label m is equal to 0. By this method, the information entropy is the maximum and the model is the most universal. The two types of data were combined to form the training sample of the model, which size is 2×n×(k+1), and it is randomly divided into training set and test set according to a certain proportion.

Step 2-2: Training the XGBoost classifier. In order to ensure the accuracy of the model at an effective computing speed, GridSearch method is used in this paper for circular training, with parameters including learning_rate, n_estimators and max_depth. Accuracy, Precision and Recall are chosen as evaluation indicators, and they are used to select the final model which performs better both in train set and test set.

3. Results Analysis

3.1. The Result of Public Security Personnel Label System

The public security personnel label system designed in this paper is divided into three layers. The first layer is divided into basic attribute, relationship attribute, behavior attribute and high-risk attribute. The second layer of basic attribute is divided into gender, age, place of household registration, educational level, religion, and etc. The second layer of relationship attribute contains relationship with key
personnel. The second layer of behavior attribute is divided into abnormal behavior, frequent trace, and etc. The second layer of high-risk attribute contains personnel category. Similarly, the above second layers can be divided into third layer according, which is a specific label value in general. For example, the third layer of age includes the teenager, youth, middle-age and elder. The results of the label system are shown in figure 3.

![Figure 3. The algorithm process of key personnel judgment model.](image)

3.2. Analysis of Model Results

In this paper, drug-related personnel are taken as the research target, and the modeling data is based on the public security business data of a city. The original data has been desensitized, and the model results are only used for demonstration and analysis. The total number of positive samples is about 50,000, and the proportion of labels is shown in table 1. It can be seen that the labels of the top 10 are no-religion, middle-age, native population, male, nationwide key personnel, foreign household registration, medium height, relatives of key personnel, work with key personnel, and case-related personnel. It can reflect the common characteristics of drug-related personnel from the statistical point of view, but it is not enough to be used as the criterion.
Table 1. Top 10 labels of drug-related personnel.

| ID | Label                          | Label Proportion |
|----|--------------------------------|------------------|
| 1  | No Religion                    | 0.99921          |
| 2  | Middle-age                     | 0.98053          |
| 3  | Native Population              | 0.93934          |
| 4  | Male                           | 0.77813          |
| 5  | Nationwide key Personnel       | 0.73402          |
| 6  | Foreign Household Registration | 0.64524          |
| 7  | Medium Height                  | 0.60639          |
| 8  | Relatives of Key Personnel     | 0.56446          |
| 9  | Work with Key Personnel        | 0.53345          |
| 10 | Case-related Personnel         | 0.50648          |

In order to make the model universal, the model samples are randomly divided into two times according to 9:1, 8:2, 7:3, 6:4 and 5:5 to form training set and test set, which are numbered 1-10 in turn. Three modeling methods are used for comparative analysis, the first is the integral weight model built based on the experience of civilian police, the second is the model algorithm described in this paper, and the third is the logistic Regression (LR) algorithm. The former is used to compare the effect difference between business experience and machine learning model, while the latter is used to compare the result difference of classifier selection. The Accuracy, Precision and Recall of the three models are shown in figure 4.

Figure 4. The algorithm process of key personnel judgment model.

It can be seen in figure 4 that three models under 10 samples perform no significant difference on Accuracy, Precision and Recall. Accuracy and Precision are both more than 80%, and Recall is more than 70%, which means three models have a certain effect on drug-related personnel judgement. Moreover, the evaluation indicators of the three models in the train set and test set are similar, indicating that the model has strong generalization ability. The average Accuracy of the three models under 10 model samples is shown in table 2, in which the Accuracy of the integration model on the test set is 80.68%, the Accuracy of the algorithm in this paper is 94.04%, and the Accuracy of the LR algorithm is 82.83%. From the perspective of the index results, the results of the model developed in this paper are far better than the integral model based on business experience and LR, which can better serve research and judgment of drug-related personnel in actual combat.
Table 2. Average accuracy of three models.

| Model                          | Average accuracy on train set | Average accuracy on test set |
|-------------------------------|------------------------------|-----------------------------|
| Integral Model                | 80.74%                       | 80.68%                      |
| Key Personnel Judgment Model  | 96.08%                       | 94.04%                      |
| LR Classifier                 | 82.82%                       | 82.83%                      |

4. Conclusion
Aiming at the problem of complicated data type and low quantization degree of public security data, which makes it inconvenient to use machine learning for model analysis directly, this paper firstly transforms complex business data into quantized label features by off-line big data computing technology, then builds the public security personnel label system, and constructs a key personnel judgement model based on FP-Growth and XGBoost. Model results show that the model built in this paper performs better than the integral model based on business experience and LR on Accuracy, Precision and Recall, which can better serve the research and judgment of key personnel in actual combat.

References
[1] Saarikkom K E and Kivivuori J 2013 young people as objects of police control in a nordic context: Who are the socially visible targets? European Journal on Criminal Policy and Research 19 (4) 351-368.
[2] Tomar N and Manjhvar A K 2018 Role of Clustering in Crime Detection: Application of Fuzzy K-means Advances in Computer and Computational Sciences.
[3] Dutta S and Gupta A K and Narayan N 2018 Identity crime detection using data Mining IEEE International Conference on Computational Intelligence & Networks.
[4] Al-Ajlan M A and Ykhlef M 2018 Optimized twitter cyberbullying detection based on deep learning 2018 21st IEEE Saudi Computer Society National Computer Conference (NCC).
[5] Wei K Q and Yang Y Q 2019 Analysis of people at high risk of terrorism from the perspective of big data—A case study of 34 “Lone Wolf” terrorist attacks Journal of China Interpol College 2019 (4).
[6] Cai L B 2019 Study on Risk Assessment of Social Drug Users People’s Public Security (University of China Beijing).
[7] Zhang D W 2017 Research on the application of public security data label construction The Police Technology 6 38-41.
[8] Han J, Pei J, Yin Y, et al 2004 Mining frequent patterns without candidate generation: A frequent-pattern tree approach Data Mining and Knowledge Discovery 8 (1) 53-87.
[9] Chen T and Guestrin C 2016 XGBoost: A Scalable Tree Boosting System.