Computer Science and Information Systems 17(3):705–715 https://doi.org/10.2298/CSIS200107016Z

A New Model for Predicting the Attributes of Suspects

Chuyue Zhang[1], Xiaofan Zhao[1*], Manchun Cai[1], Dawei Wang[2] and Luzhe Cao[1]

[1] School of Information Technology and Cyber Security
People's Public Security University of China
Beijing, China
zhaoxiaofan@ppsuc.edu.cn
[2] College of Criminology
People's Public Security University of China
Beijing, China
649661694@qq.com

Abstract. In this paper, we propose a new model to predict the age and number of suspects through the feature modeling of historical data. We discrete the case information into values of 20 dimensions. After feature selection, we use 9 machine learning algorithms and Deep Neural Networks to extract the numerical features. In addition, we use Convolutional Neural Networks and Long Short-Term Memory to extract the text features of case description. These two types of features are fused and fed into fully connected layer and softmax layer. This work is an extension of our short conference proceeding paper. The experimental results show that the new model improved accuracy by 3% in predicting the number of suspects and improved accuracy by 12% in predicting the number of suspects. To the best of our knowledge, it is the first time to combine machine learning and deep learning in crime prediction.

Keywords: crime prediction, suspect prediction, machine learning, deep learning.

1. Introduction

Criminal activities show certain distribution characteristics in time and space, which indicates that criminal activities are not completely random. The offenders often carry out the crime selectively according to the time and place. It is a hot topic to discover the law of crime from criminal activities.

With the development of statistics and criminology, the cognition and analysis of criminal behavior are gradually improved. Big data has many applications in the field of public security and criminology. Crime prediction provides assistance for crime prevention, public security prevention and control, case detection and police decision-making and has become a hot research topic nowadays.

In this paper, deep learning algorithms and machine learning algorithms are applied to the prediction of criminal suspects. Through a large number of historical data, the relationship between the information of the case and the attributes of the suspect is mined. Compared with the traditional person-oriented case investigation, the model based on big data can provide police officers with more objective and comprehensive auxiliary information and help them find the perpetrators as soon as possible. The effective and accurate information of the suspect's characteristics is of great significance for the rapid detection of the case. This can not only reduce the police useless work, but also help the victims to recover the loss. For the construction of a harmonious society, it has great significance.

The main work of this paper is to use the historical data of the theft crime to predict the age of the suspect in the theft case through deep learning. Based on the idea of ensemble learning, this paper combines deep learning with machine learning, and uses textual data and numerical data to predict the age of suspects. When a new crime occurs, it can provide the predicted results of age and number.

2. Related Works

Police have embraced predictive analytics and data-driven metrics to improve law enforcement tactics, practice, and strategy [1]. Predictive policing [2] methods fall into four general categories: predicting the occurrence of crime, predicting the suspects, predicting the possibility of recommitting the crime and predicting the victims of crimes. This paper focuses on predicting the suspects.

The information analysis of the traditional case solving mainly depends on the experience of the investigators. Crime has the characteristics of time and space aggregation [3]. Melo S N et al. [4] analyzed that crime has certain regularity and stability over a long period of time. Sagovsky et al. [5] analyzed the relationship between season, temperature and property crimes. Almanie et al. [6] analyzed the economic factors, demographic factors for crime and used Decision Tree classifier and Naive Bayesian classifier to predict crime types. Michael Oyinloye et al. [7] analyzed the influence of age, income and education on criminal behavior. Song et al. [8] analyzed the impact of different border areas on different types of crime. For the suspect's prediction, TOLLENAAR N et al [9] use statistical method to predicate general recidivism, violent recidivism and sexual recidivism. Based on the case information and victim information, LI Ronggang et al. [10] use Support Vector Machine algorithm to predict the suspect's gender, age, race, etc. Based on date and location, crime type, criminal ID and the acquaintances, Vural MS et al. [11] use Naive Bayesian Model to predict criminal of particular crime incident. Based on the features of criminals in criminal case, SUN Feifei et al. [12] uses random forest model to predict possible suspects. But some of the input data in these studies is only known after solving the case, such as suspect age, criminal history, acquaintances, etc. In the real situation, we only know the objective information of the case, and we could predict the characteristics of suspects from it. In this paper, we only use case information obtained by investigators to predict the number of suspects.

The object of this paper is the crime of theft. We extract case features such as the time of the case, loss amount, method, places and so on, and turn them into numerical data according to certain rules. In order to make up for the feature loss caused by data dispersion, we added more comprehensive text information on case description. In addition, we combine machine learning and deep learning in training numerical data. For text data, we use Convolutional Neural Network (CNN) to capture the local features and Long Short-Term Memory (LSTM) to capture text sequence features. When a new case occurs, the model can provide investigators with the prediction of age of suspects.

3. Data Preprocessing

In this paper, the time, place, loss amount, methods and case categories are preprocessed and labeled as shown in tables 1 to 6.

Table 1. The data processing rules of case categories.

Label	Categories	Number
1	pickpocketing	5500
2	theft of property in the car	4872
3	household theft	5500
4	theft of non-motor vehicles	5500

Table 2. The data processing rules of time period.

Label	Time (24 hours)	Note
1	0:00~8:00	Including 8:00
2	8:00~12:00	Including 12:00
3	12:00~14:00	Including 14:00
4	14:00~18:00	Including 18:00
5	18:00~22:00	Including 22:00
6	22:00~24:00	Including 24:00

Table 3. The data processing rules of places.

Label	Categories	Instructions
1	Residential	Residence, dormitory
2	Traffic area	Subway station, bus, roadside, expressway, etc.
3	Office area	Schools, hospitals, parking lots, offices, etc.
4	Entertainment area	Shopping malls, vegetable markets, street shops, catering places, Internet cafes and so on.

Table 4. The data processing rules of methods.

Label	Categories	Instructions
1	Others	Stolen goods from trucks, Unarmed climbing, Decoy
2	Pickpocketing	Pickpocketing, By the way to steal
3	Technical unlock	Insert card unlock, Poke the lock, Tin foil unlock, Technical unlock
4	From the window to enter	Break Windows, Break glass
5	From the roof to enter	From the roof, From the vent
6	From the wall to enter	Climb over the wall, Break walls
7	From the door to enter (Door damage)	Expand seam, Destroy anti-theft net (column), Destroy the door body
8	From the door to enter (No damage to the door)	Deceive into the door, Follow others into the door
9	Violence unlocked	Break the lock core, Pliers cut the lock, Break the lock
10	Violence hit the car	Smashing car windows, Pry the trunk of the car
11	Theft through the window	Get the key through the window with a pole, Open the window and steal
12	Stolen vehicle	Towing, dragged, Traction

Table 5. The data processing rules of loss amount.

Label	money (RMB)	Note
1	0~900	Including 900
2	900~1880	Including 1880
3	1880~3600	Including 3600
4	3600~1000000	More than 3600

Table 6. The data processing rules of weather.

Weather	sunny	Rain or snow	Others
Daytime weather	1	2	3
Night weather	1	2	3

The age predicted in this article is the suspect's age at the time of the crime. The prediction results are designed as a four-category problem. Considering the equilibrium of the data distribution, we take 1/4, median and 3/4 of the loss amount as the dividing point and divide the loss amount ages into four categories. The first quartile is 25. The median is 32. The third quartile is 42.

The number of suspects in this data set is extremely unbalanced. In order to balance the data, this paper unifies the label of more than or equal to two people committing the crime as 2, and the label of single person committing the crime as 1.

According to different time scales, the time information extracted includes year, quarter, month, ten days, day, week and time period. The location is described by latitude, longitude, administrative districts and place types. The day temperature, night temperature, day wind, night wind, day weather and night weather are also added to this model as auxiliary information. Example of pre-processed data are shown in table 7. Please refer to the conference paper for detailed description.

Table 7. Example of pre-processed data.

inputs	numerical data	year	2018	category	1	week	5	night weather	1
		month	5	method	9	time period	5	day temperature	10
		day	20	longtitude	11.11	loss amount	3	night temperature	0
		quarter	2	latitude	11.11	districts	2	day wind	3
		ten-days	3	day weather	1	places	3	night wind	3
	text data	On May 20, 2018, 20, alarm, * * * id number * * * * *, tel * * * *, address * * * *, the alarm person lost * * * in * * * *, The lost items were valued at * * *.							
outputs	age	2							
	number	1							

4. Model

We remove the feature with the lowest contribution by mutual information and Chi-square test, in order to reduce the influence of irrelevant factors on the prediction results.

4.1. Numerical Data

Different machine learning algorithms have different advantages in data processing. Using the idea of ensemble learning, we build a group of classifiers composed of Logistics Regression (LR), Support Vector Machine (SVM), Naive Bayesian (NB), XGBoost, k-NearestNeighbor (KNN), GradientTree, Boosting (GBDT), adaboost, Quadratic Discriminant Analysis (QDA) and Linear Discriminant Analysis(LDA), and input their output results and raw data into the neural network. Although some of the above machine learning algorithms have used the idea of ensemble learning, in this paper, the results of the above nine algorithms are used as the input of the neural network, and the weight distribution of each input is completed by the neural network. The processing of numerical data is shown in figure 1.

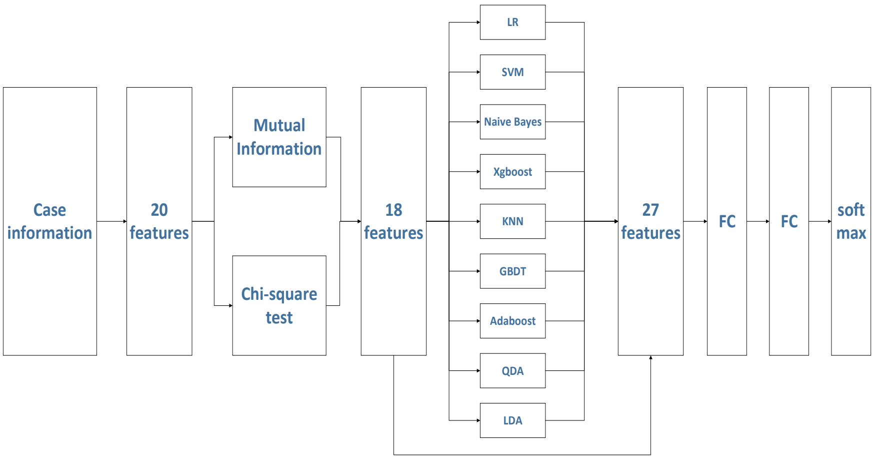

Fig. 1. Model flow for numerical data

Fully connected (FC) layer is to make linearly indivisible data become linearly separable. Formula 1 is as follows. Softmax is to make the features more distinct. Formula 2 is as follows.

$$h_{W,b}(x) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}} \left(\sum_{i=1}^{n} W_i x_i + b\right) \tag{1}$$

$$y_i = S(h)_i = \frac{e^{hi}}{\sum_{j=1}^{n} e^{hj}}, \qquad i = 1, \ldots, n \tag{2}$$

In formula 1, W_i is the weight of the neuron. b is the bias of the neuron. n is the number of neurons. In formula 2, h is the input and y_i is the probability of being in the n_{th} class.

Considering overfitting, we did not use Decision Tree (DT) and Random Forest Classifier (RFC). In the prediction of age, they can achieve 91% accuracy on the training set, but only 51% accuracy on the test set. The machine learning experimental results are shown in table 8. If the machine learning algorithm is overfitting, the data classification selected as the training set will be very good, but if a new case occurs, the classification result would be not good.

Table 8. Machine learning experimental results for the prediction of age.

Algorithm	Accuracy on the training set	Accuracy on the test set
LR	0.32	0.31
SVM	0.50	0.41
bayes	0.32	0.33
xgboost	0.32	0.32
knn	0.43	0.39

GBDT	0.36	0.35
Adaboost	0.47	0.45
LDA	0.32	0.33
QDA	0.36	0.35
RFC	0.91	0.51
DT	0.91	0.51

4.2. Text Data

The time, place, loss amount, methods and case categories of the case are all from the structured tables in the public security system. But these don't fully describe the case, such as the name of the lost item, the details of the crime scene description. Therefore, we added the case description information in the form of text.

The case description is filled in by the police officer handling the case, ranging from 5 to 150 words. This paper uses CNN and LSTM for text processing. CNN's local perception and weight sharing mechanism enables the algorithm to capture local features. The gate mechanism of LSTM selectively letting information through enables the algorithm to learn long distance dependencies. The processing of text data is shown in figure 2.

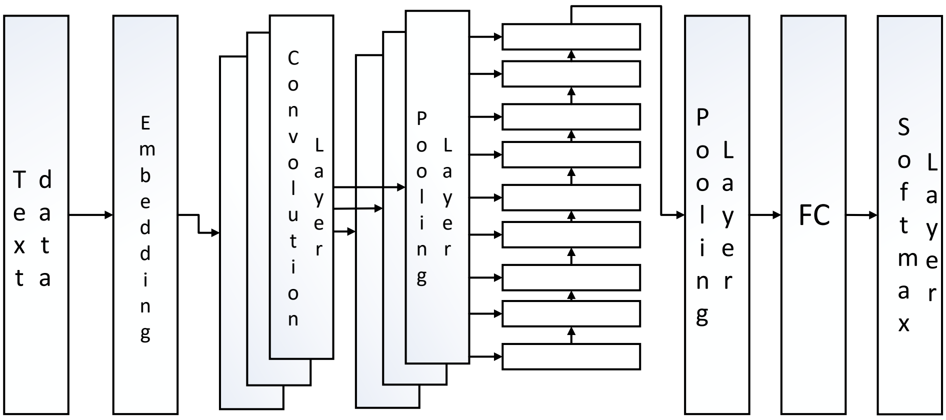

Fig. 2. Text data model structure

4.3. Model construction

In this paper, we use a parametric-matrix-based method to fuse the two types of features. The last layer of both numerical data model and text data model is softmax layer with the output shape of 4. The fusion formula is as follows.

$$X_f = W_n \circ X_n + W_t \circ X_t \tag{3}$$

$\circ$ in formula is Hadamard product, X_n and X_t are the processed numerical and textual features. W_n and W_t are the learnable parameters which is used to adjust the degrees.

After fusion, we put them into the full connection layer. Finally, classification tasks are completed through softmax layer. The model structure is shown in figure 3.

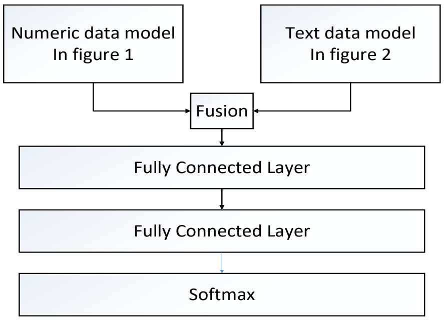

Fig. 3. Model structure

5. Experiments

5.1. Experimental Data

The data sets in this paper are all real data from the public security system, totaling 21372 pieces. We judge the practicability of this model by predicting the number and the age of suspects, named as test set 1 and test set 2 respectively.

Training sets, validation sets, and test sets are allocated according to the ratio of 7:2:1. The optimal model is determined by training set and cross validation set. The precision rate, recall rate, and F value of the test set are calculated as model evaluation.

5.2. Comparative Experiments

The comparative experiments are as follows.

- DNN (Use 20 - dimensional numerical data)
- Numerical -only model (Figure 1 model)
- Text-only model (Figure 2 model)
- Model in the conference paper

The experimental results for data 1 are shown in table 9. The experimental results for data 2 are shown in table 10.

Table 9. Experimental results for predicting the number of suspects.

Algorithm	Precision rate	Recall rate	F-values
DNN	71	73	72
numerical -only model	90	90	90
text-only model	65	64	64
Model in the conference paper	91	91	91
The new model	94	93	93

Table 9. Experimental results for predicting the age of suspects.

Algorithm	Precision rate	Recall rate	F-values
DNN	38	37	37
numerical -only model	57	56	56
text-only model	43	43	43
Model in the conference paper	53	51	52
The new model	65	65	65

5.3. Results Analysis

The experimental results are analyzed as follows:

1) This paper proves the possibility of predicting the age and number of suspects from the historical data of the case. The numerical discretization rules we define by ourselves are reasonable and feasible.

2) Compared to the conference model, the improvement of this model is the combination of machine learning and deep learning in numerical data processing. In addition, the LSTM is added. Experimental results show that these two improvements can improve accuracy by 3% in predicting the number of suspects and improved accuracy by 12% in predicting the number of suspects.

3) This model is not very good at predicting the age of suspects. When a crime is committed by more than one person, this model can only recommend the most likely age interval. This simulation is not applicable to older suspects who commit crimes together with younger suspects.

6. Conclusion

The big data algorithm is used to model the historical data and guide the public security decision-making. In this paper, the age of suspects and the number of suspects are predicted by modeling the features of real case information. Different machine learning algorithms have different advantages. We combine them, assign weights through the neural network. In addition, we add text data in order to make more accurate predictions.

When new cases occur, our model can provide the police with decision support. The model achieves 94 percent accurate in predicting the number of suspects and 65 percent accurate in predicting the age of suspects. Our model has strong practical significance. How to improve the ability of predicting the suspect's age is the next step.

Acknowledgment This work was supported by National Key R&D Program Project (Grant No.2018YFC0809802), the Fundamental Research Funds for the Central Universities (Grant No.2018JKF609), Specialized Research Fund of Higher Education of China (Grant No. 2019ssky012).

References

1. Ferguson, Andrew Guthrie, Policing Predictive Policing. Washington University Law Review, Vol. 94, No. 5, 2017.
2. Perry, Walter L., Brian McInnis, Carter C. Price, Susan Smith, and John S. Hollywood, Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations. Santa Monica, CA: RAND Corporation, 2013.
3. Grubesic T H , Mack E A . Spatio-Temporal Interaction of Urban Crime[J]. journal of quantitative criminology, 2008, 24(3):285-306.
4. Melo S N, Matias L F, Andresen M A. Crime concentrations and similarities in spatial crime patterns in a Brazilian context[J]. Applied Geography, 2015, 62:314-324.
5. Sagovsky A , Johnson S D . When Does Repeat Burglary Victimisation Occur[J]. Australian and New Zealand Journal of Criminology, 2007, 40(1):1-26.
6. Almanie T , Mirza R , Lor E . Crime Prediction Based On Crime Types And Using Spatial And Temporal Criminal Hotspots[J]. Computer Science, 2015.
7. Michael, Oyinloye , Olamiju, Isaac , Otokiti, Kolade. Spatial Distribution of Crime in Akure, Nigeria: The GIS Perspectives SCIREA Journal of Geosciences. Journal of Geosciences. 2017,2. 21-38.
8. Song J , Andresen M A , Brantingham P L , et al. Crime on the edges: patterns of crime and land use change[J]. Cartography and geographic information science, 2017, 44(1):51-61.
9. TOLLENAAR N,HEIJDEN P G M V D. Which method predicts recidivism best: a comparison of statistical, machine learning and data mining predictive models [J]. Journal of the Royal Statistical Society,2013,176(2):565-584.
10. LI Ronggang, SUN Chunhua, JI Jianrui. Suspect characteristics prediction based on support vector machine[J]. Computer Engineering. 2017,43(11): 198-203.
11. Vural M S , G?K M . Criminal prediction using Naive Bayes theory[J]. Neural Computing and Applications, 2016:1-12.
12. SUN Feifei,CAO Zhuo,XIAO Xiaolei. Application of an improved random forest based classifier in crime prediction domain[J]. Journal of Intelligence,2014,33(10): 148-152.Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. In Proceedings of the 20th International Conference on Very Large Databases. Morgan Kaufmann, Santiago, Chile, 487-499. (1994)

Chuyue Zhang is a postgraduate student in People’s Public Security University of China. Her main research interest is deep learning and crime prediction. E-mail: 649661694@qq.com

Xiaofan Zhao (corresponding author) is a PhD and teacher. Her research interest is data mining. She is the corresponding author of this paper. Email: zhaoxiaofan@ppsuc.edu.cn

Manchun Cai is a PhD and associate professor. His research interest is cryptology. E-mail: caimanchun@ppsuc.edu.cn

Dawei Wang is a PhD and professor. His research interest is criminology. E-mail: wdw_ppsuc@163.com.

Luzhe Cao is a postgraduate student in People's Public Security University of China. His research interest is security situation awareness. E-mail: 867987100@qq.com

Received: January 01, 2020; Accepted: June 18, 2020