ISBN 978-981-14-1684-2
Proceedings of 2019 the 9th International Workshop on Computer Science and Engineering
(WCSE 2019 SUMMER)
Hong Kong, 15-17 June, 2019, pp. 130-135
doi: 10.18178/wcse.2019.06.019

Application of an Ensemble Learning based Classifier in Crime Prediction

Rui Lu [1+], Linying Li [2]

[1] Information Department, Liaoning Police College, China
[2] Software College, Dalian University of Foreign Languages, China

Abstract. As an outstanding issue for police, crime prediction has been paid widely attention by researchers. Based on Ensemble Learning method, this paper applies random forest to classifier and introduces a crime prediction method to deeply explore the characteristics of criminal suspects to achieve the purpose of crime prevention. According to calculating attribute importance, the method keeps the important attributes in the algorithm. The reduced attribute set is used to train the random forest model to obtain the crime prediction classifier. The crime data was applied to the proposed classifier which is evaluated by the precision and recall. The experimental results show that the presented classifier is effective.

Keywords: crime prediction classifier, ensemble learning algorithm, random forest, data mining

1. Introduction

Currently, criminal cases in China are increasing and becoming more and more complex. The crime data show exponential growth in the amount of data, and the forms of data are complex and diverse. The application of crime big data needs practical quantitative analysis and forecasting application. At the same time, the non-publicity of crime data makes it difficult to obtain crime data, which also limits the development of crime prediction research. In contrast, data mining methods have shown good performance in different areas of prediction applications.

Research shows that the application of criminal case, victim and suspect data to data mining can help to find hidden patterns and provide decision support for law enforcement [1]. According to the historical data, Li et al. [2] trains the SVM-based feature prediction model, and calculates the feature similarity with the candidate suspect database, so as to predict the suspect. According to the research of the public security department, the crimes committed by criminals depend on some basic person attributes. These attributes are of great significance to discover criminal suspects after a case [3].For the crime prediction problem, sun et al. [4] propose an improved random forest classifier. Using random forest algorithm, Wang et al. [5] present a crime risk prediction model to find the characteristics of criminal suspects. Luiz et al. [6] propose several methods to predict crime and measure the existing correlations between crime and urban metrics. Elena et al. [7] combine Benford's Law and Machine Learning to detect money laundering criminals from actual Spanish court cases. Aiming at insurance fraud detection, Li et al. [8] propose a multiple classifier system based on the random forest, principle component analysis and potential nearest neighbour methods.

This paper presents a random forest-based crime prediction model which aims to deeply explore the characteristics of criminal suspects. Firstly, the order of attribute importance is calculated according to the historical crime data. Then the obtained attribute set is used to train the random forest model and finally obtain the crime prediction model. At last, the simulation shows that compared to SVM and naive Bayesian method, the random forest-based crime prediction model is move accurate and efficient.

+ Corresponding author. Tel.: + 86-411-8672-9635; fax: +86-411-8672-9111.
E-mail address: luruilly@sina.com.

2. Theory Method

Random Forest which uses decision tree as base learner to construct bagging ensemble, is a typical ensemble learning method. s. The variable importance comes from random forest algorithm can be used as attribute reduction method for data. Therefore, it has been widely used in various classification, regression, prediction, feature selection and outlier detection problems in recent years [9,10].

Definition 1: For the given combination classification model $\{h_1(X), h_2(X), \cdots h_k(X)\}$, the margin function is defined as

$$mg(X,Y) = av_k I(h_k(X) = Y) - \max_{j \neq k} av_k I(h_k(X) = j) \tag{1}$$

In equation (1), (X,Y) presents data sets subject to random distribution， $I(\cdot)$ is characteristic function. The margin function measures the interval between the average number of correct and wrong classifications.

Definition 2: The generalization error of combination classification model is defined as

$$PE^* = P_{X,Y}(mg(X,Y) < 0) \tag{2}$$

In equation (2), the subscript X,Y means the probability P covers the space of X and Y.

Definition 3: If the number of classifications in the forest increases, according to the law of large Numbers, the generalization error of the combinatorial classification model converges to

$$P_{X,Y}\{P_\theta(h(X,\theta) = Y) - \max_{j \neq Y} P_\theta(h(X,\theta) = j) < 0\} \tag{3}$$

θ is the parameter vector of a single decision tree.

By perturbing the attribute value in the out of bag (OOB) data, the influence of the attribute on the classification result can be judged. The greater the influence is, the more important the attribute is [10].

Definition 4: The attribute importance (AI) of attribute j in tree m is defined as

$$AI_m(j) = \frac{\sum_{i \in L_m} I(y^i = y_i^m)}{|L_m|} - \frac{\sum_{i \in L_{m,j}} I(y^i = y_{i,j}^m)}{|L_{m,j}|} \tag{4}$$

In equation (4), L_m presents the OOB data in tree m, y_i^m is the prediction results before the perturbation. $L_{m,j}$ presents the new OOB data after the perturbing attribute j, $y_{i,j}^m$ is the prediction results in dataset $L_{m,j}$, y^i presents the actual classification value in dataset L_m . If the attribute j does not appear in the tree m , then considers $AI_m(j) = 0$.

Definition 5: Attribute importance based on classification accuracy of OOB is defined as mean decrease in accuracy (MDA) which equals to the discrepancy of average accuracy before and after slight perturbation of external bag data independent variables. AI is computed by

$$MDA = \frac{1}{M}\sum_{m=1}^{M} AI_m(j) \tag{5}$$

Equation (5) illustrates the contribution of attribute importance to classification model, and is used as heuristic information for attribute reduction.

3. Ensemble Learning based Classifier

The characteristics of criminal suspect is a part of the characteristics of criminal cases. The analysis procedure for suspect characteristics need to be related to crime characteristics. The analysis process for classification is shown in Fig.1.

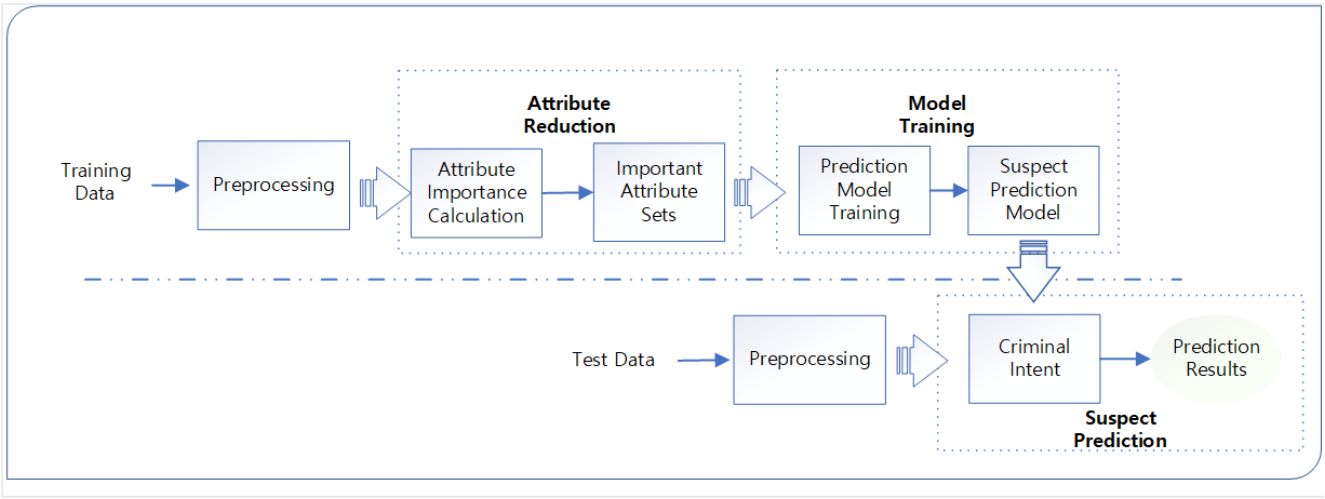

Fig 1: An attribute reduction-based classification for suspects

Redundant data attributes unrelated to prediction operations are removed, while attribute values are generalized and missing values are processed, etc. The purpose is to improve data quality and make it suitable for the input and operation requirements of the model.

Attribute reduction is an important step in the prediction method. By calculating the importance degree of attributes, the attributes less associated with the prediction results are removed, and only the important attributes are retained to participate in the calculation, so as to reduce the computational load of the algorithm and improve the practicability of the algorithm.

After attribute reduction, the training data set enters the training process of model. In this paper, a training method based on random forest is designed to obtain the suspect decision model.

In the suspect prediction stage, the preprocessed test data are input into the prediction model, and the criminal tendency of each test set sample is calculated, so as to draw an identification conclusion.

4. Prediction Model based on Ensemble Learning

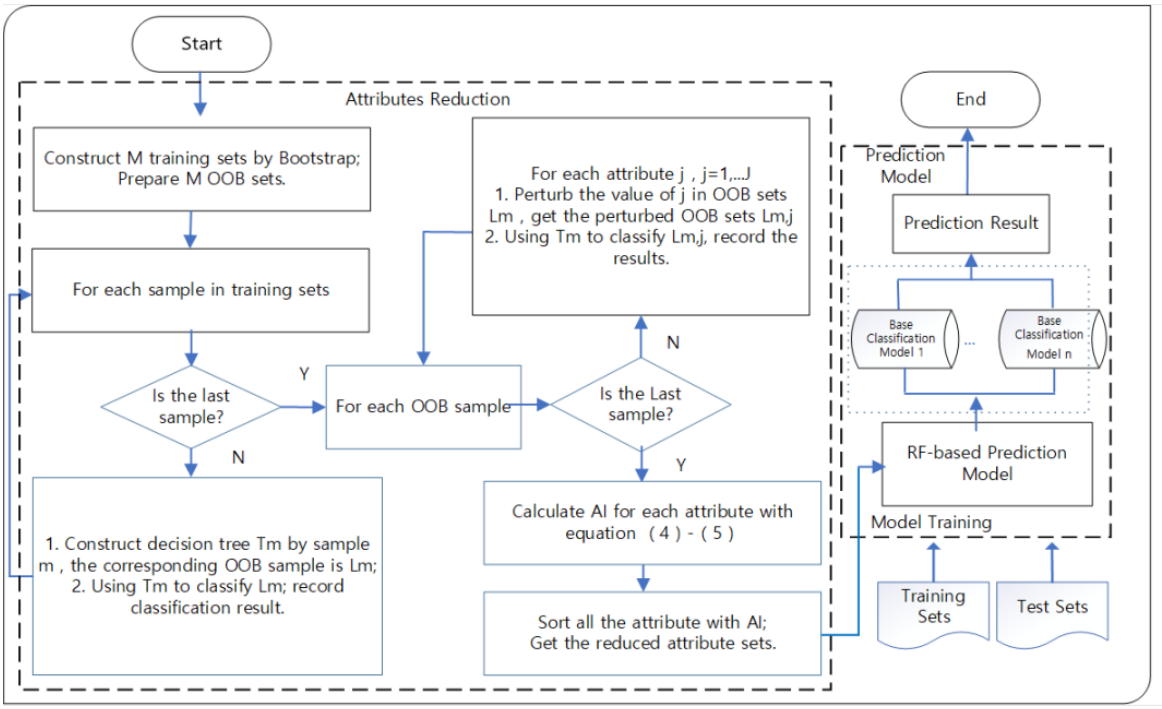

Fig.2: Prediction model based on random forest

The prediction model is shown in Fig.2. In this model, attribution reduction method takes the following steps

(a) Given the training set, bootstrap sample generates M ($m=1,2,...,M$) new training set, the other data set that is not extracted is called as OOB data L_m.

(b) Do the following operations with every sample set m ($m=1,2,...,M$) : construct M decision trees T_m, the corresponding OOB set is donated as L_m.

(c) Use decision tree T_m to classify data set L_m, record the classification results as y_i^m.

(d) Perturb every OOB data set: for j, $j=1,2,...,J$,perturb the value of attribute j in L_m and then get the disturbed data set $L_{m,j}$.

(e) Apply T_m to classify perturbed data set $L_{m,j}$ and then get the classification results $y_{i,j}^m$.

(f) After perturbation of the attribute values of each OOB data sets, the attribute importance of each attribute is calculated by equation (4) and (5).

(g) According to the attribute importance of each attribute, the reduced set of important attributes is formed and input into the prediction model.

In the stage of model training and model prediction, the model is constructed with random forest idea. In the training stage, the training data set enters the model for attribute reduction, and then the random forest method is applied for model training, thus generating n base classification models. The test data set is input into each base classification model for classification, and then the predicted results are determined by voting.

5. Experiment and Analysis

The experimental data are derived from part of the records of the information of the criminals who have been desensitized, and are used to mine the evidence relationship between the attribute characteristics of suspects and the risk of crime, so as to obtain the suspects with high degree of suspicion, and finally achieve the purpose of crime prevention and decision support.

The input information of the model is the characteristics of criminal personnel, including age, family status, education level, employed, ex-convict, specialty, resident, gender, height, weight and financial status. The education level is subdivided into the categories of elementary school, middle school, high school, bachelor degree, master degree and doctor degree. The output information of the model is the classification result of the "degree of crime" of the criminal suspect with two categories {general, serious}.

5.1. Data preprocessing

Data preprocessing is one of the key steps to improve data quality. According to the characteristics of experimental data, we need to deal with missing values in data sets, fill missing values as much as possible, and delete records that cannot fill missing values. After data preprocessing, 2021 valid records were finally extracted, including 1036 in "general" category and 985 in "serious" category. Some of the quantified data are shown in Table 1. The method described in Section 3.2 is used to reduce attributes and obtain the MDA values. After calculation and reduction, {Age, Family Status, Education Level, Employed, Ex-convict, Specialty, Resident} are found to be important attributes.

Table 1: Some quantified data

Age	Family Status	Education Level	Employed	Criminal Record	Specialty	Resident	Criminal Degree
3	1	1	0	1	1	0	1
2	2	2	1	0	1	1	1
2	3	4	1	1	0	1	1
1	3	2	1	1	1	1	0
3	2	1	1	1	1	1	1

5.2. Experiment result

Table 2: The experiment result

#	P	R	F_β
1	0.8654	0.7872	0.8097
2	0.8611	0.8123	0.8267
3	0.8824	0.8762	0.8781
4	0.8378	0.7817	0.7981
5	0.8735	0.8201	0.8358
6	0.8619	0.8163	0.8298
7	0.8731	0.7873	0.8118
8	0.8661	0.88521	0.8792
9	0.8772	0.87152	0.8733
10	0.8315	0.7809	0.7958

The control variable method was used to adjust parameters to achieve a better prediction accuracy. The number of trees is 200, and the number of selected candidate variables for each split is 3. The precision P and recall R of the model are indicators to measure the performance of the model. In suspect prediction scenario, it is desirable to miss as few suspects as possible, so the recall R is more important [9]. Considering the precision and recall, we use weight harmonic average (F1) as the performance measure for the model. The general form of F1 is F_β, which can express different preferences for precision and recall. The experiment results are shown in Table 2.

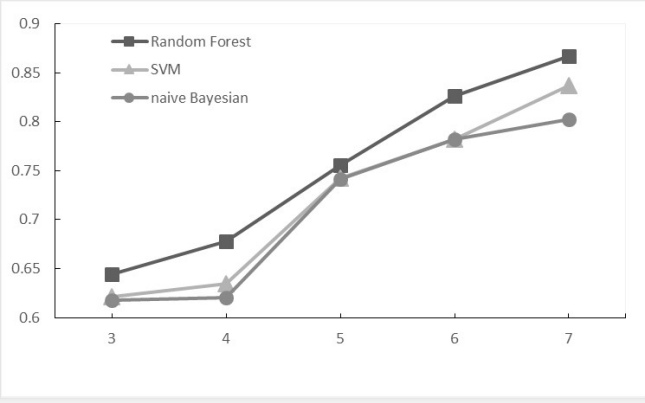

Fig.3: The results

In order to verify the performance of the random forest prediction model, SVM and naive Bayesian method were selected respectively on the Weka platform, and the operation was carried out with default parameters. The comparison of results is shown in Fig.3.The data in Table 2 and Fig.3 illustrate the feasibility of the proposed suspect prediction model. Through this model, high-risk suspects in new cases can be predicted, and the analysis results can be further compared in the relevant database, so as to realize the purpose of key research and judgment and improve the efficiency of case investigation.

6. Conclusion

Effective prediction of criminal suspects can not only achieve a rapid crackdown, but also achieve the purpose of crime prevention. Ensemble learning algorithms have been applied in different fields of prediction. In this paper, a prediction model of criminal suspects based on random forest is proposed to evaluate and reduce the attributes of criminal suspects, which effectively improves the efficiency and accuracy of the method and avoids the limitation of single decision tree classification. The model is evaluated by crime data, and the results show that the proposed model is more accurate than SVM and naive Bayesian method, and the model can be further applied to the prediction of suspects in different types of cases.

7. Acknowledgements

This work is supported by the studying project of Collaborative Innovation Center for Economics Crime Investigation and Prevention Technology, Jiangxi Province JXJZXTCX-029, by Liaoning Educational Committee Scientific Research for Youth LQ201787002, by Liaoning Educational Committee Scientific Research 2016jyt-lj02, by Liaoning Natural Science Foundation 20180550284.

8. References

[1] R Yang, S Olafsson. Classification for predicting offender affiliation with murder victims. Expert Systems with Applications, 2011,38(11):13518-13526.

[2] R. Li, C. Sun, J. Ji. Suspect characteristics prediction based on support vector machine. Computer Engineering, 2017, 43(11):198-203.

[3] S. Luo, Z. Liu, L. Guo, et al. Research on suspected culprit recognition based on probit. Transactions of Beijing Institute of Technology, 2011, 31(11):1337-1341.

[4] F. Sun, Z. Cao, X. Xiao, XIAO Xiaolei. Application of an improved random forest based classifier in crime prediction domain. Journal of Intelligence, 2014(10): 148-152.

[5] Y. Wang, Z. Guo, Y. Wang. A forecasting model of crime risk based on random forest. Journal of East China Normal University(Natural Science),. 2017(4):89-96.

[6] G. Luiz, V. Haroldo, A. Francisco. Crime prediction through urban metrics and statistical learning. Physica A: Statistical Mechanics and its Applications,2018,505: 435-443.

[7] B. Elena, A. José, M. Jose . Combining Benford's Law and machine learning to detect money laundering. An actual Spanish court case. Forensic Science International, 2018, 282:24-34.

[8] Y. Li, C. Yan, W. Liu. A principle component analysis-based random forest with the potential nearest neighbor method for automobile insurance fraud identification. Applied Soft Computing, 2018(70): 1000-1009.

[9] Z. Zhou. Machine Learning. Tsinghua University Press, 2016.

[10] D. Yao, J. Yang, X. Zhan. Feture selection algorithm based on random forest[J]. Journal of Jilin University (Engineering and Technology Editions), 2014, 44(1):137-141.