Random Forest Algorithm Based on Genetic Algorithm Optimization for Property-Related Crime Prediction

Tuo Shi [a], Gao He[b] and Yulei Mu[c]

Beijing Police College Beijing, China

[a]stshi8808@sina.com, [b]47536488@qq.com, [c]maurice623@163.com

Abstract. The property-related crime is an important type of crime, which affects the stability of social order, and presents the characteristic of "high-occurrence but low-breakage". In order to realize the precise prevention against the property-related crime, this paper constructs a large-scale environmental factors data set based on real official data, and proposes a random forest algorithm based on genetic algorithm optimization. This algorithm constructs a classification model according to the data of environmental factors that affect the property-related crime, and predicts the trend of the property-related crime in the region. Experimental results show that, the algorithm proposed in this paper achieves a good performance in forecasting the trend of property-related crime, and significantly improves the efficiency of searching for the optimal solution with low computational complexity. It has a strong application prospect in forecasting property-related crime, it could provide early warning to the future trend of the crime.

Keywords: Genetic Algorithm, Random Forest, property-related property-related crime Prediction, intelligence-led policing.

1. Introduction

The property-related property-related crime is a major problem, which affects the stability of social order[1].In order to effectively combat the multiple cases of the property-related crime , “realizing and forecasting the trend of the property-related crime in various criminal spaces”has become a key issue to be solved urgently. In the traditional analysis, the early warning of crime relies on the experience of police officers or the artificial statistical law to give qualitative conclusions[2].In the past, through using information technology to conduct research on property-related crime prediction, the classification accuracy rate was always unsatisfactory. On the one hand, the data quality and feature scale are limited, and on the other hand, the classification performance of the selected model is limited. Therefore, this paper explores the solutions for the above two aspects: First, through a large number of official data,75-dimensional features are chosen to build a large-scale data resources that affecting the property-related crime.Secondly, aiming at warning task of property-related crime, an algorithm named GA-RF is proposed. The random forest algorithm combines Bagging and random subspace techniques. It is an advanced classification algorithm based on decision tree combination. In order to obtain better combination accuracy[3], it is necessary to ensure the accuracy and diversity of a single decision tree. Then, "how to improve the accuracy of the decision tree combination based on reducing the search complexity" has become a key issue. In order to solve the above problems, genetic algorithm is selected to improve the accuracy of integrated classification from the set of decision trees in random forests—constituting the “optimal decision tree set”, eliminating the unhelpful improvement of the accuracy of integrated classification. The reason why the genetic algorithm is used to optimize the random forest algorithm is that the genetic algorithm can effectively use the randomization and parallelization technology to efficiently search globally for a coded parameter space, and quickly obtain the parameter optimization combination while maintaining the diversity of the population, also avoid to falling into the local optimal solution and over-fitting[4][5]. Finally, this paper constructs a large-scale real property-related crime data set, and sets up comparative experiments to verifies the proposed algorithms. The experimental results show that the proposed algorithm can achieve higher performance and application prospects in crime prediction.

The structure of the rest paper is arranged as follows: section 2 summarizes the existing information-based property-related property-related crime prevention schemes and random forest improvement algorithms; section 3 sorts out the environmental factors affecting the property-related crimes of invading property-related crimes, and builds large based on these factors. Scale data set; section 4 is on the basis of summarizing genetic algorithm and random forest algorithm, explains how the algorithm reduces the random deep forest algorithm search space, and then introduces the random forest algorithm based on genetic algorithm optimization proposed in this paper; section 5 is based on the above The data set is constructed to compare the algorithm proposed in this paper with other comparison algorithms and analyze the experimental results.

2. Construction of Data Set on Environmental Factors Influencing the Property-related Crime

2.1 Analysis of Environmental Factors

Firstly, on the basis of experts' experience, with the real data of public security organs, this paper summarizes the factors on economic, transportation, factors affecting property property-related crimepopulation, region, energy and facilities aspects, with a total Indicators of environmental of 75 potential environmental features. The related indicators are as shown in Table 1.

Table 1 This paper constructs a dataset for influencing factors of property-related crime

Wholesale and retail business units	Wholesale and retail business units	Number of residential service, repair and other service units
Number of health and social work units	Number of large legal entities	Number of medium-sized legal entities
......	......	......

2.2 Building a DataSet

For the dataset, the jurisdiction of police station is basically corresponding to the area of street township level, so this paper takes 331 streets and towns in a city of our country as the research object. At the same time, this paper obtains the environmental factors related to the property-related crime of embezzlement through open and professional channels. The dataset involves a total of 1,277,814 cases of property-related crime, with a total of 75 dimensions of features. At the same time, according to the distribution of the data, the regions are divided into 6 levels: highest, higher, high, low, lower, lowest. Taking 331 streets and towns as the research object, this paper takes the type of each region on property-related crime as prediction target.

3. Random Forest Algorithm Based on Genetic Algorithm Optimization

In order to improve the prediction accuracy, this paper improves the classification accuracy of random forest algorithm. with the help of genetic algorithm, it puts forward a new algorithm for predicting the trend of property-related crime,using genetic algorithm to search the decision tree of random forest and generating the most beneficial decision tree combination, so that the precision of integrated classification is promoted. These decision tree knots are eventually synthesized into the new integrated classifier.

Random forests introduce the concept of integrated voting-by growing a set of non-pruning decision trees (each decision tree is randomly sampled from the training data by self-help method) and allowing them to vote for most categories. The introduction of ensemble strategies enables random forests to reach lower variance at the same time[6].

Random forest involves the selection of algorithm parameters, its performance also depends on the choice of algorithm parameters, but the setting of parameters often rely on manual experience, so how to automatically optimize the parameter and quickly select the features with low computational cost is an urgent problem to be solved--this paper introduces genetic algorithm to overcome this challenge.

Genetic algorithm (Genetic algorithm, GA) is a general-purpose optimization search algorithm based on Darwin's natural selection and genetic theory in biological systems. In genetic algorithms, the individual of a population is called a "chromosome (chromosome)",which corresponding to a solutionfor particular problem. In this article, each chromosome represents a "decision tree combination", so the length of each chromosome is the number of decision tree, if a bit has a value of "1", the decision tree is retained, and the value is "0", which indicates that the decision tree is deprecated.

In summary, this paper proposes a random forest algorithm based on genetic algorithm optimization. This algorithm combines genetic algorithm and random forest to select the optimal combination of parameters for classification[7]. The core parameters of the random forest algorithm based on genetic algorithm optimization proposed in this paper include: population size, crossover probability $P_{crossover}$、 Mutation probability P_{Mution}、 fitness function (Fitness Function), termination conditions, etc. The execution process of the algorithm includes the following seven steps:

Step 1: Generate an initial random forest decision tree set. Use the Bootstrap to process the training set to obtain the self-help sample training set; use the obtained self-help sample training set to train a decision tree; at each node of the decision tree, randomly select m features, and select a feature from the information gain as the current node. Repeating the above steps until a decision tree is constructed, and then you can get the decision-tree set of the random forest could be gotten.

Step 2: Define the fitness function. The fitness function is used to measure the quality of a chromosome, that is, to measure the quality of a solution (the optimal decision tree set). To some extent, fitness functions act as “environmental” decisions—as the execution of genetic algorithms, populations continue to evolve, continually producing individuals with high fitness function values (ie “adaptation to the environment”). In order to directly optimize the final result, this paper selects the calculation formula of the final integrated classifier accuracy to calculate the fitness function of each decision tree combination, which is defined as follows:

$$F = 1 - \frac{1}{M}\sum_{i=1}^{M}[f(x_i) - y_i]^2 \tag{1}$$

Among them, $f(x_i)$ represents the first in all test samples i classification results of the samples, y_i express the label of the sample.

Step 3: If the population evolution process reaches the termination condition, record the optimal chromosome as the final result (“optimal decision tree combination”), and terminate the whole process; if the population evolution process does not reach the termination condition, proceed to the next step .

Step 4: Select the process.Using the fitness function to model with the “Survival of the Fittest” principle and select the individuals with the highest fitness function value as the “parents” of next generation – because these individuals have higher probability to derive better decisions. A tree combination used to perform the crossover and mutation processes described below.

The probability that an individual x is selected as the next generation of "parents" is:

$$p(x) = \frac{F(x)}{\sum_{x'} F(x')} \tag{2}$$

Wherein F(x) and F(x') represent the fitness of individual x and individual x', respectively.

Step 5: Cross process. In the crossover phase, each bit of the "parent" chromosome selected in the previous step is exchanged with the crossover probability P_Crossover - for each bit, a value within the interval (0, 1) is randomly generated, if the value is less than P_Crossover, then the value of this bit of two chromosomes is not exchanged; if the value is greater than or equal to P_Crossover, the value of this bit of the two chromosomes is exchanged. The crossover process can accelerate the discovery of new elite chromosome individuals.

Step 6: Variation process. The mutation process, including the above-mentioned crossover process, enhances the randomness and recombination ability of the genetic algorithm, avoids over-fitting to a certain extent, and also increasing diversity and expands of the search range, while reducing the risk of falling into local optimum or the two progeny chromosomes obtained by the crossover process, each of the two chromosomes is inverted with a mutation probability P_Mutation - for each bit, a value within the interval (0, 1) is randomly generated, if the value is less than P_Mutation does not invert this bit; if the value is greater than or equal to P_Mutation, this bit is inverted - if the original value is 0, the mutation is 1, indicating that the decision tree has changed from a deprecated state to the selection state; if the original value of this bit is 1, the mutation is 0, indicating that the decision tree is transitioned from the selected state to the deprecated state.

Step 7: Integrate the classifier. After generating the "optimal decision tree set", all the classifiers in the optimal decision tree votes to the test sample, and the category with the most votes is the final category of the test sample. At this point, the classification and prediction task of the property-related crime-causing property-related crime trend is completed.

4. Analysis of Experimental Results.

4.1 Contrast Algorithm and Experimental Setup

In order to verify the performance of the algorithm GA-RF proposed in this paper, the traditional Random forest algorithm (RF), Xgboost, SVM-bagging and NB-adaboost four integration methods with outstanding performance in the current classification prediction are selected as the comparison algorithm. At the same time, the Random forest baseline algorithm PAR-RF with parameter optimization is selected as the comparison algorithm. The evaluation indexes used in comparison of several algorithms are accuracy, precision rate, recall rate and F1-score. For the GA-RF algorithm proposed in this paper, its main parameters involve two parts: genetic algorithm and Random forest algorithm. Firstly, for random forest algorithm, the decision tree set of it is the newly integrated individual classifier selection space, and the factors of computational efficiency are considered, also the number of decision trees is set to 100 according to experience and experiment verification. Secondly, for genetic algorithm, considering the operating efficiency and results, the population size is set to 15, the mutation rate is 0.05, and the iterative algebra is set to 50.

4.2 Experimental Results and Analysis

The results of the comparative experiments are shown in Table 2. In this paper, a significance test is used to verify the results of the experiment, superscript † and ‡ representing significant improvements to algorithm RF and algorithm PAR-RF, respectively $p < 0.05$.In order to evaluate the prediction performance of the algorithm, this paper uses the 10-fold cross validation to measure the evaluation indicators of the classification method[8].

According to Table 2, the proposed GA-RF algorithm shows good performance in the trend analysis of the property-related crime, and its accuracy is significantly higher than the other four integrated algorithms, and the F-value is also the highest. F-value of the PAR-RF method and the F-value of traditional random forest algorithm RF are increased by 5.0% and 7.7% contrast to the PAR-RF. respectively. In general, for integrated learning, the increase in integration complexity will increase the accuracy of the results, but it will also affect the operation speed. Therefore, the computational complexity of the algorithm GA-RF and the algorithm PAR-RF are compared. Table 3 shows the number of decision trees obtained by each GA-RF algorithm in the 10-fold cross-validation and the number of decision trees obtained by the algorithm PAR-RF. The experimental results show that the integration complexity of PAR-RF based on the improved grid algorithm is more than twice the genetic algorithm-based random forest algorithm GA-RF. The number of decision trees generated by one fold is relatively stable, which proves that the proposed algorithm can effectively answer the previous question - "how to improve the accuracy of decision tree combination based on reducing search complexity", and obtaining low calculation but high computational efficiency.

Table 2 Evaluation results of property-related property-related crime prediction

Algorithm	Precise rate	Accuracy rate	Recall rate	F value
SVM-Bagging	0.71	0.73	0.70	0.71
XGBOOST	0.75	0.80	0.71	0.75
NB-Adaboost	0.81	0.82	0.77	0.75
RF	0.79	0.81	0.74	0.78
PAR-RF	0.81†	0.83†	0.75†	0.80†
GA-RF(text)	0.85‡	0.88‡	0.78‡	0.84‡

Table 3 Comparison of Numbers of Decision Trees

Fold	#PAR-RF	#GA-RF
1	39	42
2	57	46
3	45	43
4	87	40
5	124	39
6	57	41
7	49	44
8	287	37
9	77	40
10	59	45
Average	88.1	41.7

In addition to the high prediction accuracy, the algorithm GA-RF proposed in this paper inherits the optimization idea of genetic algorithm, so its search efficiency in parameter optimization is also higher, in order to further verify GA-RF operation. The following experiment is proposed: genetic algorithm, simulated annealing algorithm and grid search algorithm were used separately to optimize the model parameters of random forest algorithm, and compare them. Among them, the random forest algorithm based on simulated annealing algorithm is recorded as SA-RF, and the random forest algorithm based on grid search algorithm is recorded as GS-RF. Taking time expenditure as a measure, the comparison experiment results are shown in Fig. 1.

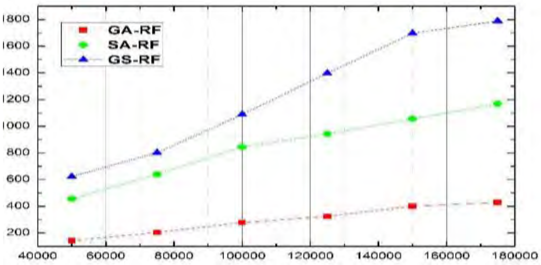

Fig. 1 Comparison of Time Cost

It can be seen from Fig. 2, that the proposed GA-RF algorithm has kept low time consumption on the data sets with different scales involved. This is because the genetic algorithm does not need to understand all the features of the problem, and the problem can be solved by the evolution mechanism. The largest time expenditure (the least efficient) is the grid search method (algorithm GS-RF), because the grid search method is more exhaustive compared with the genetic algorithm and the simulated annealing algorithm. The efficiency is very low, especially when the parameters are optimized more, the efficiency of the algorithm will further decrease. The simulated annealing based algorithm (algorithm SA-RF) has a slower convergence rate due to its slower annealing process, and the proposed algorithm GA-RF exhibits extremely strong search ability for its parallel computationality, so its advantages are more obvious; In addition, the optimization efficiency of the genetic algorithm is undoubtedly more scientific than the manual setting of parameters, without too many prior experience to select parameters.

5. Conclusion

The genetic algorithm is a general-purpose optimized algorithm based on Darwin's natural selection theory. It can efficiently search for solutions through evolutionary mechanisms without knowing all the features for the problem. According to the above-mentioned advantages of genetic algorithm, the search space of random forest algorithm is pruned and optimized. The genetic algorithm is used to “evolve” the decision tree in random forest. The satisfactory combination of decision trees proposes a random forest algorithm based on genetic optimization. Because of its integrated learning, the algorithm performances better than other classification algorithms. In addition, it uses genetic algorithm to optimize parameters, which can improve the efficiency of operation and quickly search for optimal parameters. The experiment proves that the prediction effect of the algorithm on the real data set of the property-related crime is more prominent, and effectively predicts and analyzes the trend of crime, so that the defense work against the property-related crime can be early predicted.

References

[1]. Yu Hongling, Identification of "The purpose of illegal possession" in the property-related crime of infringement of financial property-related crimes, Chinese prosecutor, 2016(18):27-30.

[2]. Liu Xiang. Research on the application of big data in the investigation of multiple infringement property-related crimes [D]. Chinese people's Public Security University, 2017.

[3]. Thanh P N, Kappas M. Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery. [J]. Sensors, 2018, 18(1):18.

[4]. Michalewicz Z. A Non-Standard Genetic Algorithm for the Nonlinear Transportation Problem[J]. Informs Journal on Computing, 2017, 3(4): 307-316.

[5]. Volkanovski A, Mavko B, Boševski T, et al. Genetic algorithm optimisation of the maintenance scheduling of generating units in a power system[J]. Reliability Engineering & System Safety, 2017, 93(6): 779-789.

[6]. Breiman L. Random Forest[J]. Machine Learning, 2001, 45: 5-32.

[7]. Friedrich T, Kötzing T, Krejca M S, et al. The Compact Genetic Algorithm is Efficient Under Extreme Gaussian Noise[J]. IEEE Transactions on Evolutionary Computation, 2017, 21 (3): 477-490.

[8]. Arias M, Mappes J, Théry M, et al. Inter-species variation in unpalatability does not explain polymorphism in a mimetic species[J]. Evolutionary Ecology, 2016, 30 (3):419-433.