Membrane Technology
ISSN (online): 1873-4049

Risk Prediction Model of Recidivism Based on Stacking Algorithm

Jia Yao*

School of Information Engineering, Shandong University, Qingdao, Shandong, China
**Corresponding Author.*

Abstract:
At present, the impact of AI on research in the field of social sciences is gradually attracting attention from the academic community. It has become a trend and made some progress to evaluate the risk of recidivism by quantitative means. However, actuarial assessment, dynamic risk factor assessment and other means take model driven as the basic method, and use human subjective hypothesis to fit the data with unknown distribution pattern, which reduces the objectivity, stability and accuracy of the conclusion to a certain extent, this is the bottleneck which restricting the research in this field. Based on the idea of data-driven, the author uses the dynamic tracking data of 26020 released persons within three years, combined with an ensemble learning method of Stacking, to establish a Intelligent recidivism risk prediction model. This scheme can overcome the limitations of artificial prediction method; mine the potential association relations and mapping rules in the data. At the same time, the author confirms this model is a more reliable and robust method for recidivism prediction than Random Forest model in terms of prediction accuracy, ACU area, elimination of over fitting, small sample learning and other indicators. This model achieves comprehensive performance improvement by adding multiple weak classifiers, so it will be a more efficient prediction method of recidivism based on big data in the future.

Keywords: data driven, recidivism risk prediction, ensemble learning, bagging, stacking

INTRODUCTION

AI has become the driving force behind a new round of technological and industrial revolution. It not only affects the natural science field but also promotes the transformation of social science research methods. Repeated crime, also known as recidivism, refers to the act of being sentenced for violating the criminal law and being sentenced for a crime again after returning to society with the end of sentence or release. The traditional risk assessment of prisoners' recidivism is using the scale evaluation method which based on model driven. Firstly, random samples are selected, and then the statistical method is used for feature extraction, finally, the indicators such as whether the prisoners are dangerous, dangerous degree and risk type are quantified[1-2]. As shown in Figure 1, from a worldwide perspective, the research in the field of recidivism risk prediction can be roughly divided into five stages: 1. From the 1950s to the 1970s, in this stage, quantitative analysis based on deductive reasoning has not been widely used. Based on the intention of description and understanding, the researchers make a follow-up observation on the certain quantity criminal individuals for a certain period of time, and summarize the research conclusions as views and experiences. 2. From the second stage, that is, after the 1980s, quantitative research has been widely used in the field of recidivism risk prediction. In this stage, the research pays more attention to the strict quantitative experiments, adds more factual basis and restrictive conditions, and its research conclusions are more widely recognized. At the same time, the introduction of the control group effectively improved the persuasion of the research conclusions and views. 3. The 1990s is the third stage, in this stage the risk of recidivism is comprehensively assessed by combining static risk with dynamic demand. 4. After the 21st century, the fourth research stage is the combination of recidivism risk assessment and individual analysis. The research in this stage is mainly based on the "Risk-Demand-Response (RNR)" model standard established by the United Kingdom and the United States. At this stage, quantitative prediction tools are widely used, such as CHR-20, which is used to assess the risk of violent crime, and Reactions on Display (RoD), which is used to assess the risk of recidivism of mental disorders. These individualized quantitative research methods have been preliminarily recognized by the academic community and applied in judicial practice to a certain extent, but there is still no consensus on its reliability [3-6]. 5. After years of development, the quantitative research on the risk prediction of recidivism has made remarkable achievements. Its optimistic prospect has been recognized by the academic and judicial circles, but it still can not make breakthrough. It can see that the quantitative analysis method has been widely used in the research field of recidivism prediction since the 1980s. Through the use of statistical and mathematical analysis, the ability to abstract and summarize the research objectives has been improved, and various causal relationships within the phenomenon of recidivism can be accurately analyzed. The basic operation process is to first establish a mathematical model based on experience, and then use the hypothetical model to fit the collected data, calculate a variety of indicators and values of the target object, and finally make an analysis conclusion. The source power of the research method is still model driven [7]. However, it is very difficult to accurately fit the real data with the pre assumed model, and there are also great accidental factors, which are the main reasons for the long-term stagnation of research in this field. With the rapid development of Internet technology and the improvement of computer computing ability, people can use

the network to obtain large-scale data, and through computer-aided completion of in-depth exploration of data, deeply grasp the distribution law of data, and create new computable methods based on the characteristics of data to complete the prediction of recidivism risk [8]. Therefore it can believe that the fifth stage of development is to adopt a data-driven assessment approach, which excavates objective principles and basis in a certain scale of sample space, accurately predicts the recidivism risk of target individuals, and promotes the research in this field to make breakthrough progress with the help of big data and machine learning.

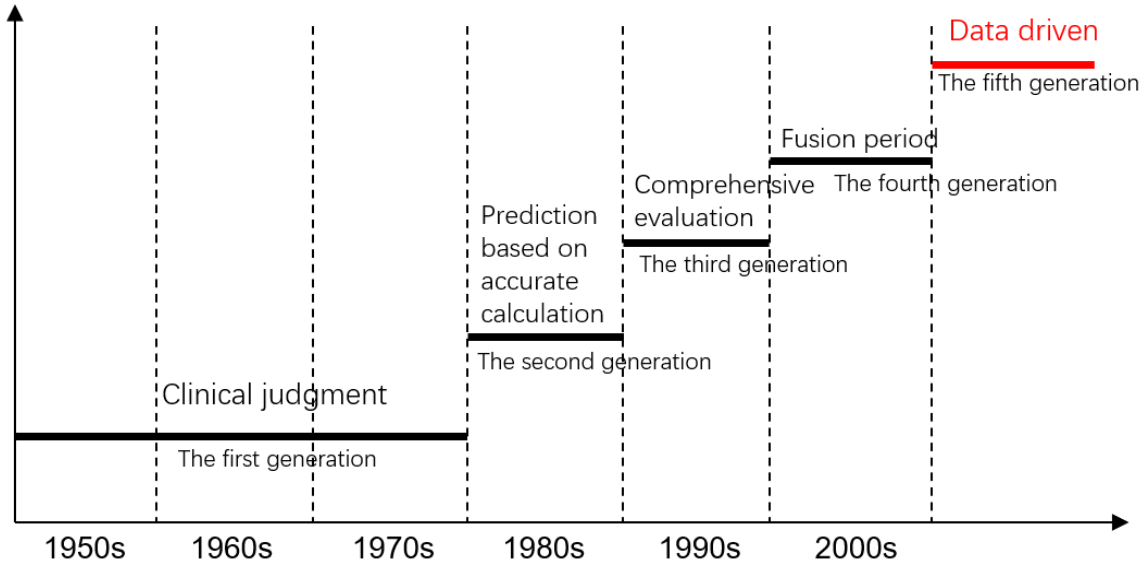

Figure 1. The development of research on recidivism risk prediction

Machine learning is a new discipline based on pattern recognition and computer learning theory. Its basic idea is to construct statistical model by using large-scale training data, that is, to learn how to complete the task from the data, and then judge and predict the events in the real world through the experience gained by learning. Statistical model is composed of training data and the process of predicting new data based on training data. The task of machine learning is to find some optimization program to minimize the error of the model based on the training data set. Different optimization procedures are called learners or classifiers, such as Decision Tree classifier (DTC), K-Nearest Neighbor (KNN) classifier, Naive Bayesian Model (NBM) classifier, etc. [6,9]. For constant $a > 0$, a learner B selects a part of samples from the input data set D for training, and obtains the prediction model h, if the prediction error of h is less than $\frac{1}{2} - a$, that is, the prediction accuracy of h is more than half (the performance is better than random guess), then B is called the weak learner of Boolean function f (0-1 problem to judge whether an individual has recidivism risk). In this task of recidivism risk prediction, DTC, KNN and NBM are all weak learners [10-11]. Although it can be determined that the prediction performance of weak learners is better than random guess, there is still room for improvement. If there is a polynomial learning algorithm that can learn a specific concept (recidivism risk prediction) and obtain better performance than the weak learner, the concept is considered to be strongly learnable, and this polynomial learning algorithm is defined as a strong learner. As shown in Figure 2, $X_1, X_2 \ldots\ldots X_n$ represents the data set needed by the training model, and $K_1, K_2 \ldots\ldots K_n$ represents the different weak learners trained based on the data set. Ensemble learning can construct and combine multiple weak learners to form a new strong learner to complete the learning task and obtain better performance. This process can be called multi-classifier system or committee-based learning. Bagging and Stacking all belong to the classification strategies of ensemble learning [12-14]. In recent years, due to their excellent performance, Bagging and Stacking have attracted extensive attention from academia and industry [15]. This work is to use Bagging and Stacking strategies to construct strong classifiers on the basis of weak learners. The author uses the relevant information of a large number of prisoners', a set of rules which can judge whether the independent individual has the tendency of recidivism is established. At the same time, the subjective and objective factors leading to recidivism have been found out to some extent, which can guide the formulation of relevant social policies to reduce the crime rate, so as to open up a new idea to deal with the related problems in the judicial field with the concept of data driven [16-17].

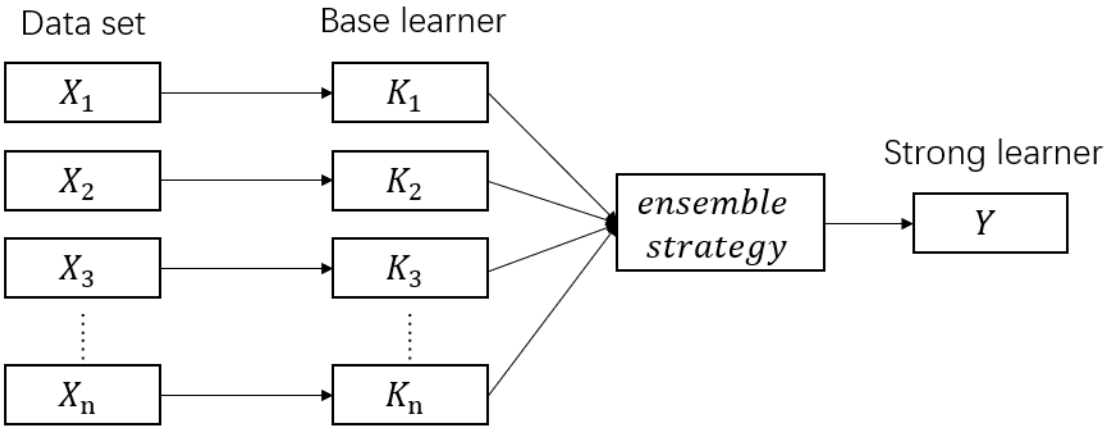

Figure 2. Construction process of strong learner

PREVIOUS WORK

Bagging algorithm is a group learning algorithm proposed by Leo Breiman in 1996. Different models usually do not produce exactly the same error on the test set. Based on this point of view, Bagging algorithm can train several different models separately and let all models vote to complete the output of test sample prediction conclusion, so as to improve the accuracy and stability of single weak classifier prediction. Random Forest (RF) is the evolution of Bagging algorithm, it makes voting through multiple decision tree models to complete the final decision, and weak learners complete the training process based on different sample subspaces, and there is no dependency among different decision tree models [18].

Decision tree is a prediction model of attribute structure, which represents a mapping relationship between object attributes and object values. Its basic idea is to induce a set of judgment rules from the training set, so that this rule has less conflict with the data set and has good generalization ability. The loss function is regularized maximum likelihood function, which is expressed as follows:

$$C_\alpha(T) = \sum_{t=1}^{T} N_t H_t(T) + \alpha|T| \tag{1}$$

$|T|$ represents the number of leaf nodes, t represents the leaf nodes of the decision tree T, t including N_t sample points, $H_t(T)$ represents the information gain or Gini coefficient of the leaf nodes, and $\alpha \geq 0$. Based on the expression, it can see that each node in the decision tree has an impure index, the information gain or Gini coefficient is usually used as the impure index. Each time the feature with the highest impure is selected for splitting, so the impure index of the child node must be lower than that of the parent node. The construction process of decision tree is around the optimization of an impure index [19]. In 1948, Claude Elwood Shannon, the father of information theory, put forward the concept of information entropy. Assuming that the sample set D contains m different samples, in the target task $m=2$, that is, individuals with recidivism risk and individuals without recidivism risk, based on a certain characteristic, the proportion of each type of sample is $p_k(k=1,2.....m)$, so the information entropy of set D is defined as follows:

$$Ent(D) = -\sum_{k=1}^{m} p_k \log_2 p_k \tag{2}$$

The decision tree of ID3 model takes information gain as the index of impure, information gain represents the degree to which the uncertainty of the sample set is reduced when the information of a feature is known [20-21]. Expressed as:

$$Gain(D) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v) \tag{3}$$

That is, information entropy minus conditional entropy, conditional entropy represents the entropy of the set divided according to a certain feature. The greater the information gain of the feature, the more concise the tree can be by using this feature as the root node of the tree. The decision tree of CART model takes Gini coefficient as the impure index. Assuming that the sample set D contains m different kinds of samples, $p_k(k=1,2.....m)$ it represents the probability that the sample belongs to

the category k [22], then the Gini coefficient is expressed as follows:

$$Gini(D) = \sum_{k=1}^{m} p_k(1-p_k) = 1 - \sum_{k=1}^{m} p_k^2 \quad (4)$$

As shown in Figure 3, Random Forest is a model composed of many decision trees. Its randomness is reflected in the training set (sample subspace) of random sampling, each decision tree classifier learns on different subsets of the sample space. On the other hand, it is reflected in the random feature subset, that is, when each node is divided in the decision tree construction, a group of features are randomly selected instead of using the complete feature space. Because the random sampling is used in the process of constructing the training set and node segmentation, the trained model has small variance and strong generalization ability. The prediction conclusion of Random Forest is based on the prediction results of all decision trees. The number of votes obtained by different classification results is counted, and the category with the most votes is taken as the final prediction conclusion. It can see that each single decision tree is a weak classifier, and the Random Forest algorithm integrates them into a strong learner. Generally, random forest can obtain higher classification accuracy than single optimal decision tree [23-24].

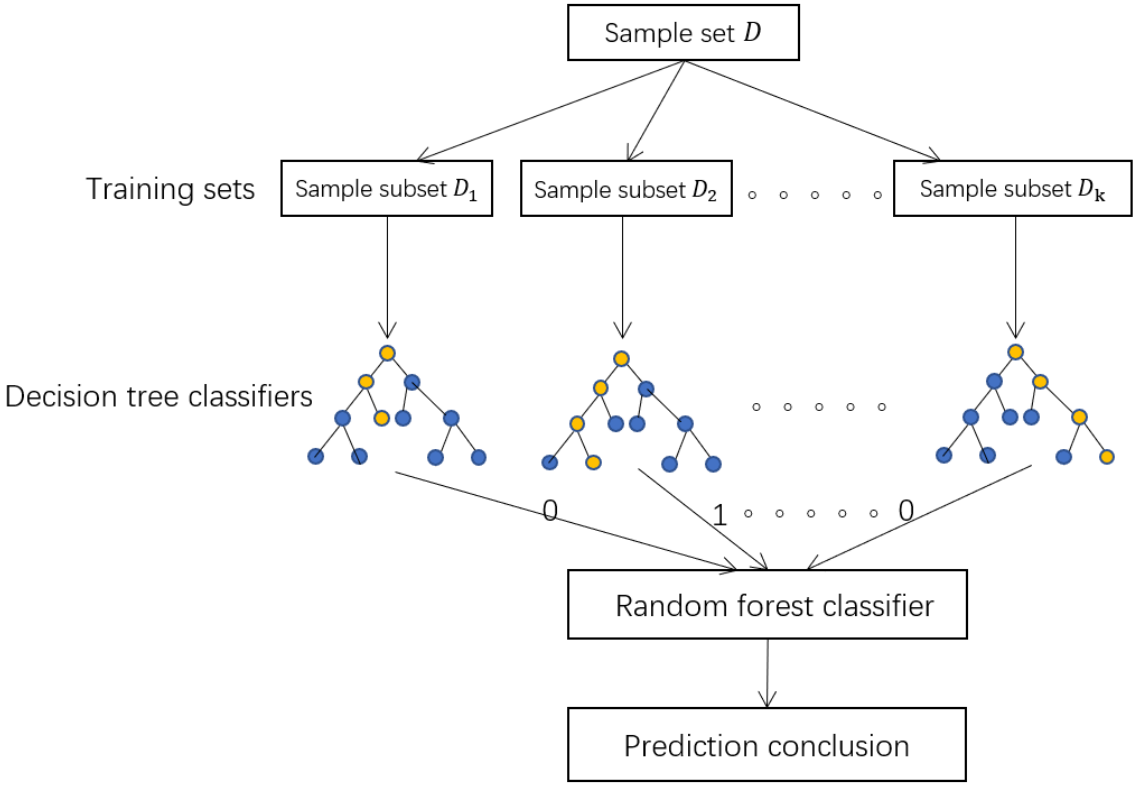

Figure 3. Implementation process of random forest algorithm

Stacking strategy trains a meta-model to combine the heterogeneous weak learners and obtain the final prediction results according to the prediction results of different weak learners. The goal of Bagging strategy is to obtain an integration model with smaller variance than its components, and the goal of Stacking strategy is to generate an integration model with lower bias than its components. Stacking strategy was proposed by David h. Wolpert in 1992, and Leo Breiman proposed "stacked regressions" by combining generalized linear model and stacking strategy in 1996. Thereafter Mark J. Van der Laan from UC Berkeley theoretically proved the effectiveness of stacking method in 2007. In recent years, breakthroughs have been made in the breadth and depth of Stacking strategy application, which is reflected in the emergence of new data allocation strategies and the emergence of deep-seated stacking (more than three layers) [25]. How to use a new Stacking strategy to complete the task of recidivism risk prediction and compare its performance with Bagging strategy will be showed in the following chapters of this paper.

FEATURE ENGINEERING

The training data for the intelligent model designed by this study is based on criminals who served sentences in Iowa, USA

from 2010 to 2015, as well as follow-up data on recidivists from 2013 to 2018, and data on recidivism within 3 years after their first release from prison , the data has been published on the Kaggle platform(https://www.kaggle.com). The author use Principal Component Analysis (PCA) to reduce the dimensionality of the data by finding the direction with the highest variance in the data, visualize high-dimensional sample feature space data into two dimensions to better understand and interpret the data. As shown in Figure 4, the sample space contains 26020 independent sample points. The sample points indicate the persons who are serving their sentences or have been released after serving their sentences. The red marking points refer to the persons who have recidivism behavior in the three-year tracking period, representing the high-risk individuals with recidivism tendency. The yellow marking points refer to the persons without recidivism, representing the low-risk individuals with recidivism tendency. Here the horizontal and vertical coordinate values only have geometric spatial significance and do not have physical significance of real world.

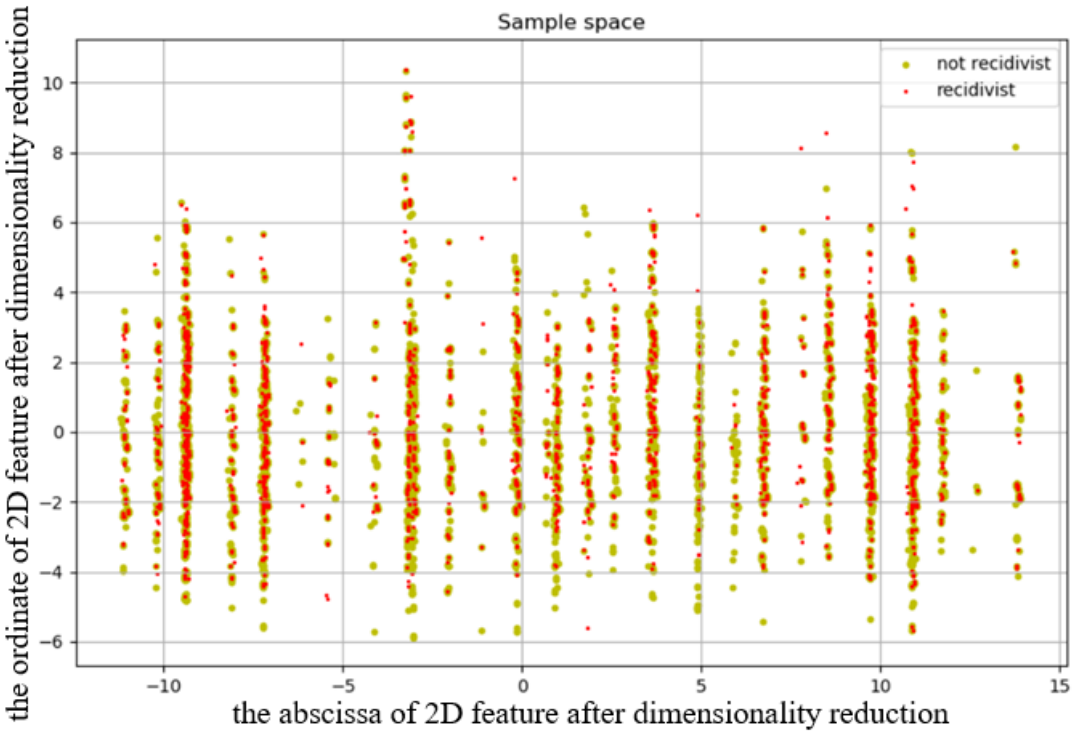

Figure 4. Sample space

The possibility of an independent individual's recidivism is related to social environment, physiological factors, psychological factors, prison management factors, and social factors after being released from prison. As shown in Figure 5, it is the feature space of an independent individual after quantitative treatment, which covers 10 features with low or high correlation with recidivism tendency, the horizontal axis represents ten different features, and the vertical axis represents the range of sample coverage in each feature space. As to the correlation between these characteristics and recidivism, it is difficult to make accurate judgment or quantitative conclusion from the perspective of criminal psychology or social psychology. Therefore, it will be an arduous task to establish a prediction model of recidivism based on artificial experience, and there is the bottleneck that is difficult to break through. It will be a brand-new attempt to find association rules and establish scientific and objective prediction method by machine learning in the background of big data. Meanwhile, it has optimistic prospects. This work is to make some efforts in this direction. The personal information of relevant personnel collected in the data set includes: 1.The time span of tracking, that is, the time span from the time when the person is released to the end of tracking. In this project, the tracking period of all personnel is 3 years, and the risk degree of recidivism is marked by whether the person has committed a crime or not during the tracking period. 2. Race - ethnicity of the person concerned indicates the correlation between different cultural background and living habits and recidivism. 3. The author divided the age of relevant personnel into five different age groups, such as teenagers, middle-aged and old people. Age factor in crime is one of the factors that affect the formation of criminal psychology. Crime statistics in various countries show that the age of high incidence of crime is mostly between 16 and 20 years old,this age group has the characteristics of emotional instability and impulsiveness, and is vulnerable to adverse factors, so that they go back to the road of crime after being released. The crime rate of the elderly over 60 years old and the children under 14 years old is relatively low. 4. The sentencing levels obtained by the relevant personnel before they are released, such as Serious Misdemeanor (one year's sentence), Aggravated Misdemeanor (two years' sentence), Simple

Misdemeanor (30 days' sentence), etc. 5. The conviction types of related personnel, such as Drug, Violence, Public Order, etc. 6. The conviction sub-types of related personnel, such as Burglary, Theft, Assault, etc. 7.The jurisdiction in which the relevant person lives during the tracking period. There are differences in management measures, social security and personnel quality in different regions. It is generally believed that there is a causal relationship between social living environment and recidivism tendency. 8&9.Specific reasons for the release of relevant personnel, such as Parole, Discharged End of Sentence, Special Sentence, etc. 10.In order to reduce the crime rate of parole, bail and released prisoners, the federal government of the United States will take relevant education and assistance policies for these target groups, but its effectiveness has not been accurately concluded [26-29].

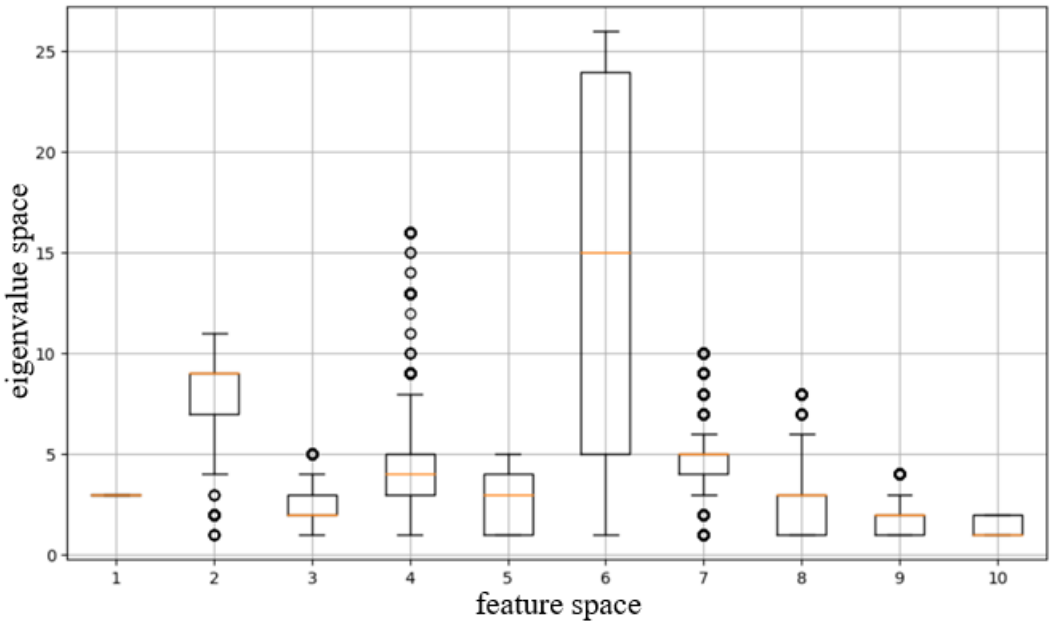

Figure 5. Feature space

METHODS AND EXPERIMENTAL RESULTS

Random Forest

Random Forest is a combination of decision tree and Bagging. Bagging is constructed based on decision tree and randomness is introduced in the generation process of decision tree [30]. The execution process of the algorithm is expressed as follows:

a. Existing data set:

$$D = \{x_{i1}, x_{i2}, x_{i3}.....x_{in}, y_i\}(i \in [1, m]) \tag{5}$$

In the target task $m = 26020$, y_i is label, indicating the crime situation of 26020 released personnel within three years, that is, whether there is a crime again, $n = 10$ represents 10 features in the feature space that are related to recidivism. Sampling space $(m \times n)^{m \times n}$ can be generated by sampling with return.

b. A weak learner decision tree is constructed, based on each sampled data set:

$$d_j = \{x_{i1}, x_{i2}, x_{i3}.....x_{ik}, y_i\}(i \in [1, m], k < n) \tag{6}$$

According to the process mentioned above, different decision trees are generated after training. The prediction result of each decision tree for unknown samples is $h_j(x)$.

c. After t times of training, the following conditions were met:

$$H(x) = \max \sum_{t=1}^{T} \phi(h_j(x) = y) \tag{7}$$

Among this formula, ϕ means a decision fusion algorithm. The author adopt the relative majority voting method, that is, the author takes the more prediction conclusion of all decision trees as the final prediction conclusion [31].

Stacking

As shown in Figure 6, the stacking strategy is a two-layer structure. In the first layer, eight different weak classifiers are used to train the training samples, and then the prediction results are used as the training samples of the second layer. In the second layer, Multinomial Naive Bayes classifier (MultinomialNB) is used as a meta- model to integrate weak learners. The weak classifiers as follows:

I. A Random Forest classifier based on ID3 pattern decision tree.

II. A Random Forest classifier based on CART pattern decision tree.

III. A ID3 model decision tree.

IV. A CART model decision tree.

V. K-nearest neighbor classifier (KNN) is a nonparametric learning algorithm. When it is necessary to predict what category an unknown sample belongs to, count the categories of k known samples closest to it, the category with more quantity is the prediction conclusion [32]. Euclidean distance is used to calculate the distance:

$$d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \tag{8}$$

In the formula, x is the unknown sample, y is the known sample, n is the size of the feature space, covering ten different factors that have correlation with the recidivism in the data set. The algorithm can be interpreted as that existing an unknown individual, and a certain group of people has high similarity with this individual. If the majority of the group is the personnel without recidivism within three years after release, the individual can be judged to be the low risk of recidivism.

VI. Naive Bayes is a generative model, which is widely used in document classification and spam filtering. Given a particular individual in this task, he has some characteristics as $(feature_1, feature_2 .....feature_n)$. It covers the age of the individual, the type of crime he has committed, and the reasons for his release,etc.. There are two different classification labels, indicating low risk of recidivism and high risk of recidivism. The goal of this task is to determine which label this individual belongs to, if satisfied:

$$p(label1 \mid feature_1, feature_2 .....feature_n) > p(label2 \mid feature_1, feature_2 .....feature_n) \tag{9}$$

It can be considered that this individual belongs to the low-risk group of recidivism. Naive Bayes uses maximum a posteriori probability estimation method, the joint distribution $P(lable, features)$ of feature space X and target space Y is trained by learning, the distribution of posterior probability $P(lable \mid features)$ is derived. Then, the maximum likelihood estimation method is used to estimate the posterior probability $P(lable_{new} \mid features_{new})$ of the new target [33]. The calculation method is expressed as follows:

$$p(label_{new} \mid features_{new}) = \frac{p(features_{new} \mid label_{new}) \times P(label_{new})}{p(features_{new})} \tag{10}$$

For a priori probability $p(feature \mid label)$ on different data sets, can choose different Naive Bayes classifiers to obtain better performance, but it is still difficult to accurately grasp the distribution law of data and quantify the features reasonably on this basis, usually, only try. It can effectively integrate the advantages of various Naive Bayes classifiers through the integration strategy to achieve more accurate prediction. When Bernoulli Naive Bayes classifier is used, the feature attribute should be quantized as an independent Boolean type value. The conditional probability of one feature is expressed as follows:

$$p(X \mid C_k) = \prod_{i=1}^{n} p_{ki}^{x_i}(1-p_{ki})^{(1-x_i)} \tag{11}$$

Where the size of the sub sample space is n, C_k is the type of the sample. If the proportion of individuals over 30 years old in the sample subspace of recidivists is 20%, the conditional probability value is 0.2.

VII. When Gaussian Naive Bayes classifier is used; conditional probability of the feature is expressed as:

$$P(X = x \mid C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{\frac{-(x-\mu_k)^2}{2\sigma_k^2}} \tag{12}$$

If the average age of the sample subspace without recidivism within three years is 32 years old, then $\mu_k = 32 . \sigma_k$ represents the variance of a certain feature value in the sample subspace to measure the dispersion of the data. If the 30-year-old group accounts for 5% of the group with recidivism within three years, the conditional probability value is 0.05.

VIII. Logistic Regression classifier is usually used to deal with binary classification problems, especially for data with strong correlation between features and tags. It has the characteristics of simple principle, strong anti noise and parallelization. In the long-term practice, it has been proved that the model can obtain better performance than the tree model classifier on small-scale dataset. The loss function of the classifier is derived based on maximum likelihood estimation, which is used to measure the information loss of the model with parameter θ when fitting the training data, so that the training model and training data can be better fitted, the loss function is expressed as:

$$J(\theta) = -\sum_{i=1}^{m} (y_i \times \log(y_\theta(x_i) + (1 - y_i) \times \log(1 - y_\theta(x_i)) \tag{13}$$

θ is a group of parameters solved, m is the number of training samples, y_i is the label of training samples, that is, whether there is crime again in the tracking period, and $y_\theta(x_i)$ is the predicted value, that is, the return value of prediction model calculated based on parameter θ, x_i represents the characteristic value of sample i. In order to avoid better performance of the model in the training set and poor performance in the verification set, the author further add a penalty term to the loss function to adjust the effect of the model fitting. The new loss function is expressed as follows:

$$J(\theta)_{L2} = C \times J(\theta) + \sqrt{\sum_{j=1}^{n} \theta_j^2} \, (j > 1) \tag{14}$$

$J(\theta)$ is the original loss function, C is the super parameter controlling the degree of regularization, n is the size of the parameter space, and the added norm $L2$ is the square root of the sum of squares of each parameter in the parameter vector [34-35].

Meta-model: Multinomial Naive Bayes classifier uses the prediction results of eight base learners in the first layer of Stacking framework as the training set to train again. Therefore, each sample in the second level training set has eight features, which represent the prediction results of different classifiers in the first layer. The author think that the new feature space is more in line with the form of polynomial distribution. At the same time, the author uses smooth processing in the calculation of prior probability $P(C_k)$ and conditional probability $P(x_i \mid C_k)$, so as to avoid that because a certain feature has not appeared in the training set resulting in $P(x_i \mid C_k) = 0$ and a posterior probability of 0. The calculation formula is expressed as follows:

$$P(C_k) = \frac{N_{C_k} + \alpha}{N + k\alpha} \tag{15}$$

Where the total number of samples is N, k is the number of categories, N_{C_k} is the number of samples with category C_k, and α is the smooth value.

$$P(x_i \mid C_k) = \frac{N_{x_i, C_k} + \alpha}{N_{C_k} + n\alpha} \tag{16}$$

Where the dimension of the feature is n, N_{x_i, C_k} is the number of samples with the ith characteristic value is x_i in the sample space of class C_k [36]. α is the smoothing value. The author set α to 1, that is, Laplace smoothing. Or Lidstone smoothing can be used, and the value range of α should be set to (0, 1).

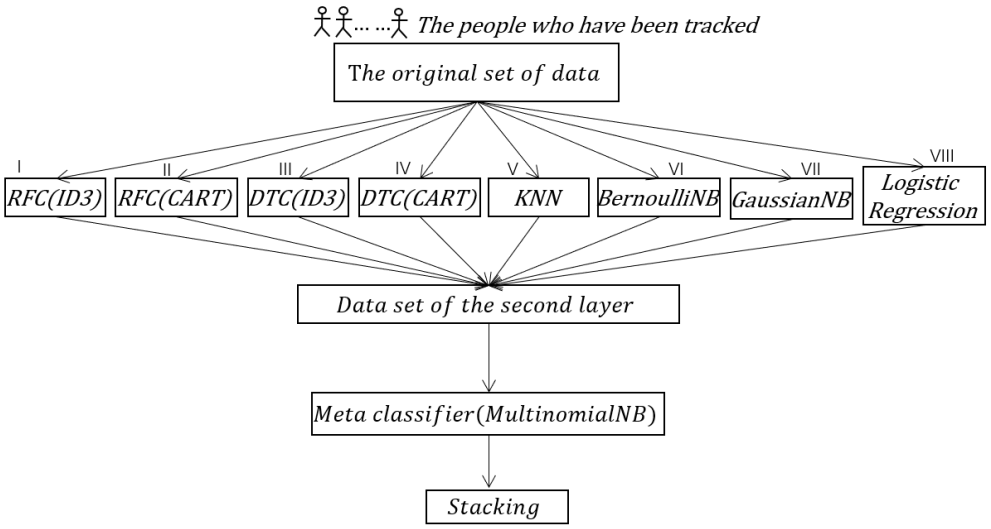

Figure 6. The structure of stacking algorithm

The basic idea of Stacking strategy can be understood as: first, the first layer of primary learners are trained by using the initial training set, and then a new data set is generated based on the prediction results of primary learners to train the second level secondary learners. As shown in Figure 7, the author takes 67% of the total data as the training set of the first layer, that is, the original training set, and 33% of the total data as the test set. In the process of constructing the second level feature (met feature), in order to avoid over fitting and information leakage, the original training set is divided by the method of five fold cross validation. After the original training set is divided, it is recorded as $train1, train2 ..... train5$. First, the data of $train2 - train5$ is used to train the base model, then the trained model is used to predict the data of $train1$, and the prediction result is used as the meta feature generated by the base model1(RFC (ID3)). In the same way, the author uses the data of $train1, train3, train4, train5$ to train the basic model, and then use the trained model to predict the data of $train2$. The prediction results are used as the Meta features generated by the base model1 from $train1, train3, train4, train5$. The above operations are repeated to obtain the Meta features generated by the base model 1 for the whole original training set. The same method is used to generate Meta features for other base models, such as RFC (CART) and Logistic Regression, to form the Meta feature set for the training of meta-model of the second layer (MultinomialNB). Each base model predicts the original features of the test set after one training, and five training based on different data can predict five groups of different conclusions, and take their mean value to construct new features of the test set [37-38].

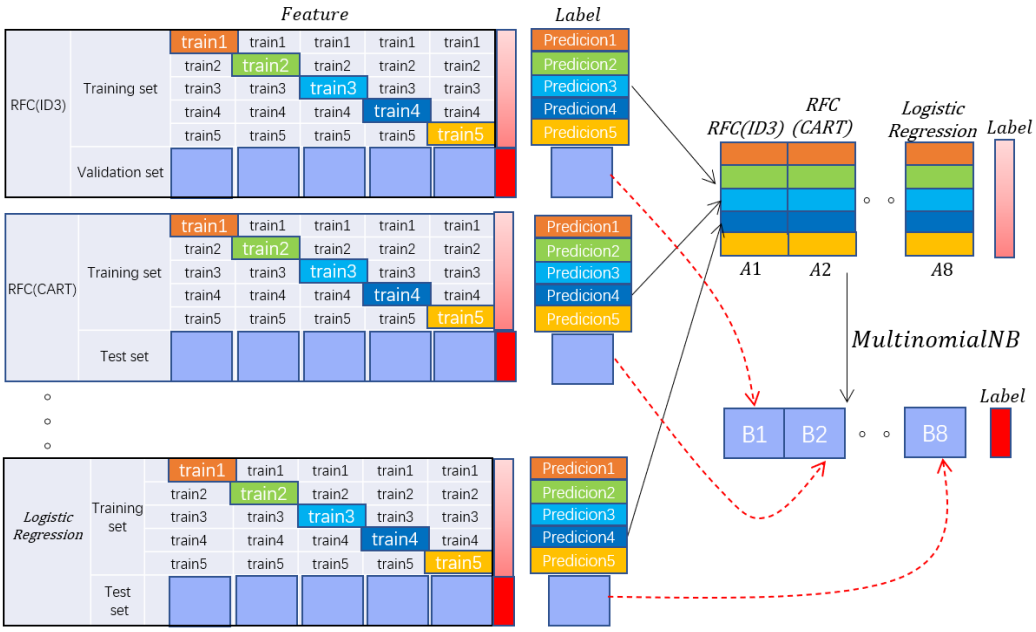

Figure 7. The execution flow of stacking algorithm

The execution process of the algorithm is expressed as follows:

Input training set $D = \{(x_1, y_1), (x_2, y_2).....(x_m, y_m)\}$, m is the number of training samples, x_i is the characteristics of the person being tracked, y_i is the label to indicate whether the tracked person has recidivism within three years.

(I) Select the junior learning algorithm $\xi_1, \xi_2.....\xi_T, T = 8$, which are Random Forest classifier, decision tree classifier, Logistic Regression classifier and so on.

(II) Select the secondary learning algorithm ξ , the author uses Multinomial Naive Bayes classifier.

(III) The output is:

$$H(x) = h^{'}(h_1(x), h_2(x).....h_T(x)) \tag{17}$$

Based on the prediction results of the primary learning algorithm, the secondary learning algorithm integrates all the classifiers into strong learners to complete the prediction with higher classification accuracy.

(IV) $for\ t = 1,2...T\ do\ h^{'} = \xi(D^{'})$

(V) $h_t = \xi_t(D)$

(VI) $end\ \ for$

(VII) $D^{'} = \ Ø$

(VIII) $for\ i = 1,2.....m\ do$

(IX) $for\ t = 1,2...T\ do$

(X) $z_{it} = h_t(x_i)$

(XI) $end\ \ for$

(XII) $D^{'} = D^{'} \cup ((z_{i1}, z_{i2}.....z_{iT}), y_i)$

(XIII) $end\ \ for$

(XIV) $h = \xi^{'}(D)$

Based on Bagging and Stacking strategies, the author constructs a recidivism risk prediction model. The following work will divide the sample space randomly, using 33% of the samples as the test set and 67% of the samples as the training set. Ten groups of independent experiments are conducted to evaluate the performance of different classifiers. As shown in Figure 8, Stacking model has a significant improvement in recognition accuracy compared with weak classifier decision tree model, and its recognition accuracy rate can reach 0.665 ± 0.05, while the recognition accuracy rate of decision tree model is only 0.645 ± 0.005.The distribution of data in the data set and the correlation between data are very complex, it is difficult to accurately grasp the data by manual methods. Different weak classifiers will observe the data from different angles, as mentioned above, single decision tree algorithm pays more attention to the impure index when node splitting, KNN algorithm pays more attention to the distance relationship between different sample points, and Logical Regression algorithm uses the loss function evaluates the fitting degree of the models, so they can only play their respective advantages for specific data distribution patterns. The author uses the Stacking strategy to fuse the observation results of different classifiers, and observe the data space and data structure from various angles, by this method will obtain higher classification accuracy in the target task. At the same time, this advantage is widely applicable and does not depend on the specific data distribution form. Compared with the weak classifier, the random forest model does not achieve significant performance improvement. The Random Forest randomly selects part of the data to establish multiple decision trees, and the final prediction results refer to the

results of different decision trees. Although in most cases, it can avoid the prediction inaccuracy of single decision tree caused by the influence of outliers, the weights of all prediction functions are the same The results show that if the performance of multiple decision trees trained based on different samples and features are basically the same, Random Forest will be difficult to achieve performance improvement. Therefore, in terms of prediction accuracy, Stacking is a more ideal ensemble strategy for the target tasks [39-41].

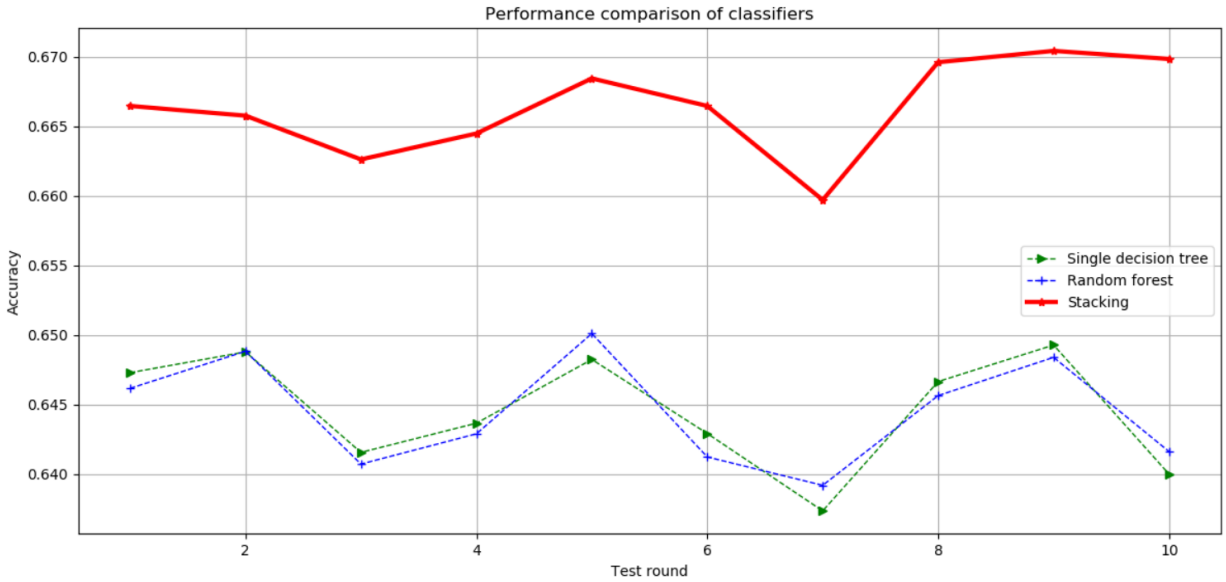

Figure 8. Comparison of prediction accuracy based on Stacking and Random Forest

As shown in Figure 9, the comparison of the over fitting performance between the Stacking model and the Random Forest model shows that the prediction accuracy of the Stacking model on the training set and the test set is basically the same based on the different segmentation ratio of the training set and the test set, and the prediction accuracy of the Random Forest model on the test set is slightly lower than that on the training set. In comparison, Stacking model not only has better prediction accuracy, but also has better performance in eliminating over fitting. A single decision tree is more sensitive to data changes, and it is easy to over fit the data set with a lot of noise. The trend of over fitting can be weakened by continual adding decision trees to the Random Forest model. The variance of generalization error of the model (reflecting the insensitivity that training data can not be effectively learned by the model) will gradually decrease, but the bias (reflecting the stereotyped memory ability of the model to the training data) did not change significantly. Therefore, with the expansion of the scale of the model, the degree of over fitting will eventually approach a specific value and cannot be continuously improved. It can see that there is still a slight over fitting phenomenon in the Random Forest model, at the same time, it is impossible to eliminate over fitting only by ensemble strategy. In the Stacking model structure, there are several weak classifiers in the first layer, among which KNN and single decision tree are high square difference algorithms. Although they can fit the training data well, they also learn a considerable number of noise points. The Logistic Regression classifier is a high bias algorithm, which uses the assumed data distribution to fit the data, although it can be used in some cases but it lacks flexibility, there are quite accidental factors in performance, and the generalization is low. The advantage of Stacking is to integrate the advantages of various weak classifiers, and effectively overcome the defects of different weak classifiers. By this method, the Bias-Variance-Tradeoff is realized through the ensemble method different from Bagging, which overcomes the model over fitting problem to a certain extent without losing accuracy and flexibility. Furthermore, with the decrease of training sample size, the performance of Random Forest model tends to decline, while the performance of Stacking model is relatively stable. This also shows that Stacking model is more suitable for learning under the condition of small amount of data [42-44].

In this task, it need to find high-risk individuals of recidivism as far as possible, and further take relevant intervention measures to these personnel. The completion effect is expressed by quantitative index TPR, and the higher the index, the better the system performance. At the same time, the author don't want to misjudge the low-risk individuals as the high-risk individuals, the FPR is the index to measure the system misjudgment rate. The lower the index is, the better the system performance is. These two indicators restrict each other. If the system is sensitive enough, it can identify the high-risk group of recidivism without significant characteristics, but the accompanying result is that the FPR index increases, that is, the risk of misjudgment will also increase at the same time. The goal of designing a strong classifier is to find a balance between the two indexes.

Taking TPR as the vertical axis and FPR as the horizontal axis, the ROC space is obtained, and the AUC value is the area covered by the ROC curve. As shown in Figure 10, based on the target task, both Stacking model and Random forest model exceed 0.6 on ACU index, and weak classifiers such as single decision tree (DTC), KNN, and Naive Bayes (MultinomialNB) cannot achieve this goal, which shows that the ensemble strategy can achieve higher recognition accuracy and lower miscalculation rate. The kernel function is used to map the data to the high-dimensional space in the model of Support Vector Machine classifier (RBF_ svc), and the linear inseparable problem between the low-risk samples and the high-risk samples in the original space is successfully solved, the ACU index reaches 0.602. It is not only a small sample learning method with high robustness, but also has better performance than other weak classifiers in the target task, but compared with the ensemble strategy, especially Stacking model, there is still a little gap. This result can be explained as: Stacking algorithm effectively ensemble different weak classifiers through training set reconstruction, and successfully breaks through the performance limit of a single weak classifier, it is impossible for SVM classifier to achieve this goal only by adjusting parameters. Boosting model is a different ensemble learning strategy, the difference between Boosting model and Bagging model is that the basic learner uses the weighted data to train. During the training process, the weight of error classification sample points will be increased, and the weight of correct classification sample points will be reduced. In the Figure 10, it can see that the performance of AdaBoost classifier is significantly improved compared with weak classifier, which is close to the performance of Stacking model, and better than Random Forest classifier, because it further introduces gradient promotion strategy on the basis of Bagging model. Expanding the research scope and depth of Boosting strategy in this field, comparing with other ensemble strategies in different performance indexes, and seeking reasonable explanation based on performance differences will be one of the key contents of the next work [45-46].

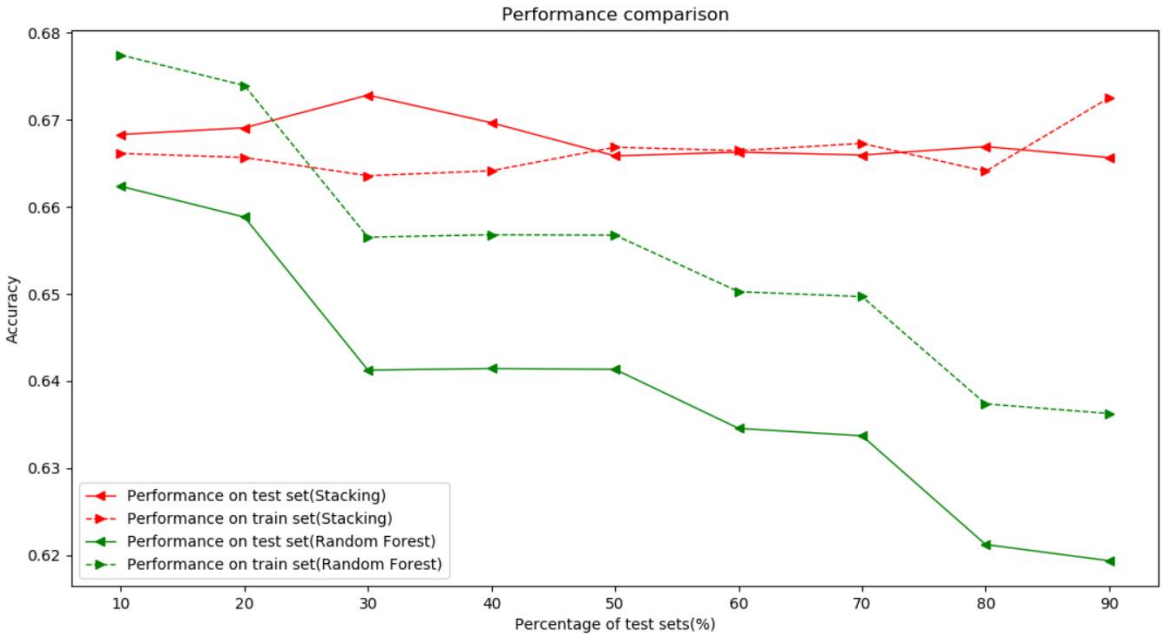

Figure 9. Comparison of prediction performance based on training set and test set

The goal is not only to accurately predict the possibility of recidivism of unknown samples, but also to find out some important factors leading to recidivism, and to establish corresponding prevention and control measures and Countermeasures to reduce the crime rate. As shown in Figure 11, the comparison of the importance of different characteristics in the prediction process of Random Forest model shows that all the features participate in the prediction of the possibility of recidivism to some extent. This shows that recidivism is the result of the combination of static and dynamic factors, individual factors and external environmental influence, and the comprehensive operation of various mechanisms. Objectively speaking, it is quite difficult to accurately predict it. The research on the correlation between age, race, crime type and recidivism has achieved preliminary results, but there is still no broad consensus in the academic community. The researchers tend to think that the recidivism is the result of a variety of different internal and external factors, so predicting through artificial experience based on different isolated influencing factors will be difficult to solve this problem efficiently. With the improvement of computer hardware ability and the development of statistical theory, based on large-scale personnel tracking data, the author establishes a data-driven prediction model, which can effectively explore the potential correlation of data, and open up a more optimistic research path for the research of recidivism risk prediction. During the execution of Stacking model, in order to obtain higher classification accuracy, the second level meta-classifier uses the prediction results of the first level basic classifier to train the

model, so the original features will not directly participate in the final decision. The researchers can not observe the importance of the original features in the decision-making process and guide the actual work in different application scenarios based on this, which reduces the interpretability and application value of Stacking model to a certain extent [47-48].

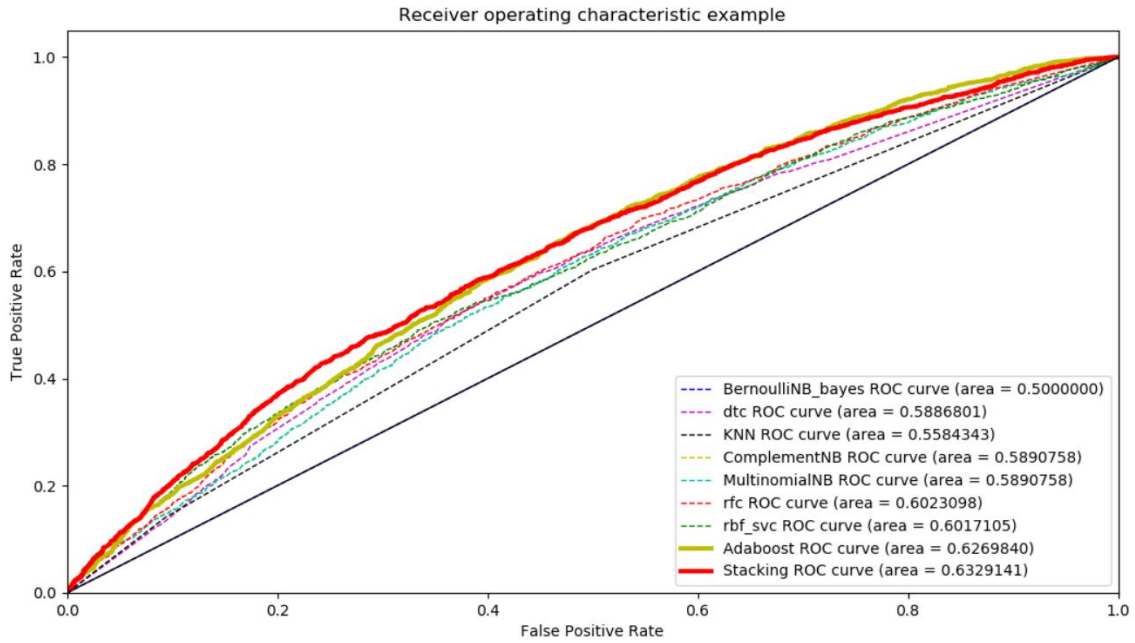

Figure 10. Comparison of different classifiers based on ROC curve

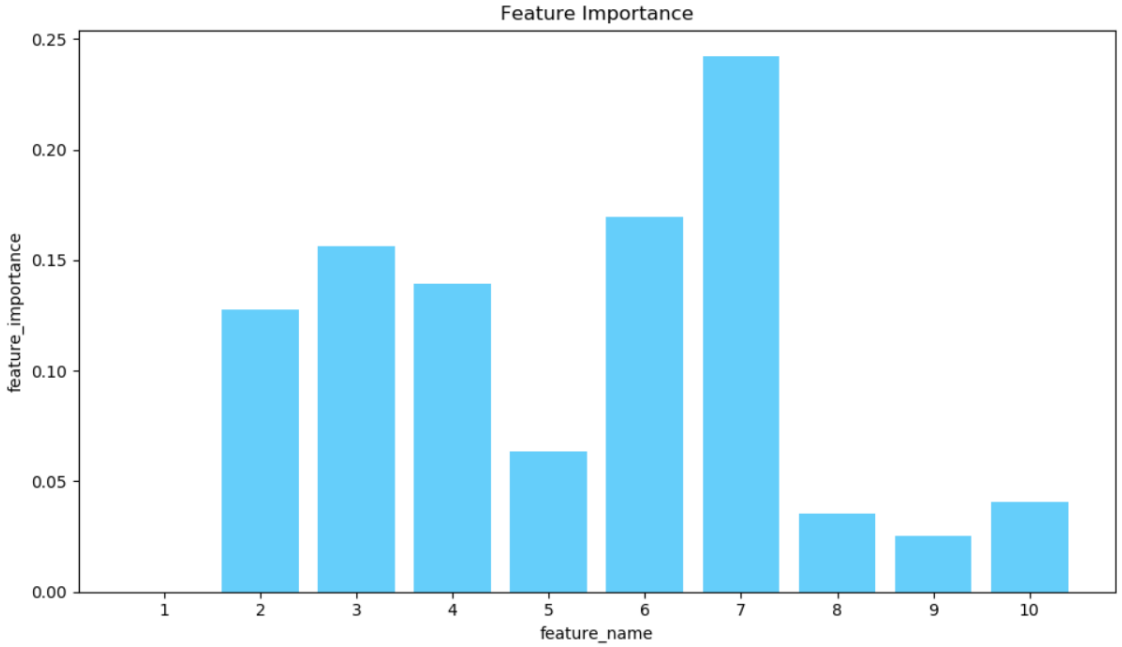

Figure 11. Comparison of the importance of different features in decision making process

CONCLUSION

The paper shows that the recidivism is the result of the interaction of known or unknown internal and external factors such as the offender's personal psychological factors, personality traits, environmental changes, etc., so it is a difficult task to find the causal relationship between recidivism behavior and various features and establish association rules by artificial prediction. Since the 1980s, most of the researches in this field have focused on making various types of assessment scales by model driven method. Although the qualitative research has evolved into quantitative research, the development of the scale still follows the traditional psychological research process of group control, questionnaire survey and statistical analysis. The basic idea is to establish the scale based on the researchers' subjective hypothesis a specific prediction model, by fitting the unknown data through the model, can not deeply mine the distribution law and potential association relationship of the data; there is

considerable contingency and the interference of the researcher's subjective factors. Even if the relatively ideal reliability and validity are obtained in some sample spaces, the model will also have a large generalization error.

With the improvement of computer hardware performance, more researchers tend to think that solving such problems should be driven by data or the problem itself, rather than relying on model driven. In recent years, the risk assessment of recidivism based on data-driven has gradually attracted the attention of academia. The research method of machine learning is to analyze the data by algorithm, construct statistical model through training data, learn from the data and establish mapping rules for prediction. However, researchers mostly use weak classification such as Support Vector Machine, single decision tree, Naive Bayes and so on, although these methods have their own advantages in different data background, the generalization ability is insufficient and the space for performance improvement is limited. Based on ensemble learning strategy, the author builds a two-layer Stacking model. The author uses KNN, single decision tree, Naive Bayes and other weak classifiers as the first level learners, and reconstruct the feature space by K-fold cross validation. And then the Multinomial Naive Bayes model is used as the meta-classifier, and the strong classifier with better performance is trained in the second layer based on the new feature space. The new fusion model can solve the Bias-Variance-Tradeoff and TPR-FPR-Tradeoff problems. It is not only significantly better than many classical weak classifiers, but also better than the Random Forest model based on Bagging strategy in the prediction accuracy, ACU area, and elimination of over fitting and small sample learning Maybe. At present, the prediction accuracy of this model has exceeded that of artificial experience, prediction accuracy of artificial experience is only about 60%, but there is still room for improvement. At the level of training data, can continue to add features related to recidivism tendency, such as behavior trajectory, correspondence, network behavior, etc., but the acquisition of this information will involve the violation of personal privacy to a certain extent, which increases the difficulty for further research. At the same time, human physiological information such as gene map, bioelectrical signal and medical image can also be added to the feature space of training samples, although the direct correlation between these information and recidivism has not been widely recognized by the academic community. At the algorithm level, can optimize the combination structure of basic classifiers in the first layer of Stacking model, select more ideal meta-classifier in the second layer, and try to increase the depth of Stacking model. Boosting strategy is also based on the idea of ensemble learning. In this research, it has also achieved better performance, which can be used as a means of further research. The above will be the next step research contents on recidivism risk prediction which based on data driven.

REFERENCE

[1] Amon, S., Putkonen, H., Weizmann-Henelius, G., Fernandez Arias, P., Klier, C.M., 2019. Gender differences in legal outcomes of filicide in Austria and Finland. Arch. Wom. Ment. Health 22 (1), 165–172.

[2] Brady, P.Q., Reyns, B.W., 2020. A focal concerns perspective on prosecutorial decision making in cases of intimate partner stalking. Crim. Justice Behav. 47 (6), 733–748.

[3] Chopin, J., Beauregard, E., 2019. Sexual homicide of children: a new classification. Int. J. Offender Ther. Comp. Criminol. 63 (9), 1681–1704.

[4] Chen, C., Bai, Y., Wang, R., 2019. Online political efficacy and political participation: a mediation analysis based on the evidence from Taiwan. New Media Soc. 21 (8), 1667–1696.

[5] Lee, S., Xenos, M., 2022. Incidental news exposure via social media and political participation: evidence of reciprocal effects. New Media Soc. 24 (1), 178–201.

[6] Oser, J., Grinson, A., Boulianne, S., Halperin, E., 2022. How political efficacy relates to online and offline political participation: a multilevel meta-analysis. Polit. Commun. 39 (5), 607–633.

[7] J. Liu, Y. Lin, W. Ding, H. Zhang, J. Du, Fuzzy mutual information-based multilabel feature selection with label dependency and streaming labels, IEEE Trans. Fuzzy Syst. 31 (1) (2022) 77–91.

[8] J. Liu, Y. Lin, W. Ding, H. Zhang, C. Wang, J. Du, Multi-label feature selection based on label distribution and neighborhood rough set, Neurocomputing 524 (2023) 142–157.

[9] Gou, B., Xu, Y., Feng, X., 2020. State-of-health estimation and remaining-useful-life prediction for lithium-ion battery using a hybrid data-driven method. IEEE Trans. Veh. Technol. 69 (10), 10854–10867.

[10] Abdulelah A, Alwanas H, Al-Musawi AA, et al. Load-carrying capacity and mode failure simulation of beam-column joint connection: application of self-tuning machine learning model. Eng Struct 2019;194:220–9.

[11] Han, B., Ji, S., Wang, J., et al., 2021. An intelligent diagnosis framework for roller bearing fault under speed fluctuation condition. Neurocomputing 420, 171–180.

[12] Hadi S, Rigoberto B. Emerging artificial intelligence methods in structural engineering. Eng Struct 2018;171:170–89.

[13] Kyriazis, N., Papadamou, S., & Corbet, S. (2020). A systematic review of the bubble dynamics of cryptocurrency prices. Research in International Business and Finance, 54, Article 101254.

[14] L. Quan-Lun, C. Zheng-Guang, J. Feng, Prediction of oil content in oil shale by near-infrared spectroscopy based on stacking ensemble learning, Spectrosc. Spectr. Anal. 43 (4) (2023) 1030–1036.

[15] Chen T. and Guestrin C. XGBoost: A scalable tree boosting system, in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2016;August 13–17:785–794.

[16] Petoft A, Abbasi M. Children's criminal perception; lessons from neurolaw. Child Indicators Res 2022;7:1–6.

[17] Petoft A. The validation requirements of neuroscientific evidences before courts. Iran J Med Law 2021;15(56):431–43.

[18] Chu, J., Zhang, Y., & Chan, S. (2019). The adaptive market hypothesis in the high frequency cryptocurrency market. International Review of Financial Analysis, 64, Article 221231.

[19] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas.(2012).“How many trees in a random forest?”. MLDM. Springer, pp. 154–168.

[20] T. G. Dietterich. (2020).“Ensemble methods in machine learning”. Multiple classifier systems. Springe, pp. 1–15.

[21] Z.Z. Cheng, et al., Temperature compensation with highly integrated ionization sensor array used in simultaneous detection of mixed gases, IEEE Trans. Ind. Electron. 69 (1) (2022) 911–920.

[22] Abdalla, Michel, Bourse, Florian, De Caro, Angelo, Pointcheval, David, 2015. Simple functional encryption schemes for inner products. In: Katz, Jonathan (Ed.), Public-Key Cryptography – PKC 2015. Springer Berlin Heidelberg, Berlin, Heidelberg, pp.733–751.

[23] D. Paul, A.K. Goswarmi, R.L. Chetri, R. Roy, P. Sen, Bayesian optimization-based boosting method of fault detection in oil-immersed transformer and reactors, IEEE Trans. Ind. Appl. 58 (2) (2022) 1910–1919.

[24] M. Elsisi, M.Q. Tran, K. Mahmoud, D.E.A. Mansour, M. Lehtonen, M.M.F. Darwish, Effective IoT-based deep learning platform for online fault diagnosis of power transformers against cyberattacks and data uncertainties, Measurement 190 (2022).

[25] Berzati, A., Viera, A., Chartouny, M., et al., 2023. Exploiting intermediate value leakage in dilithium: a template-based approach. IACR Trans. Cryptogr. Hardw. Embed. Syst. 2023 (4), 188–210.

[26] Cobbina-Dungy, J. E., & Jones-Brown, D. (2023). Too much policing: Why calls are made to defund the police. Punishment and Society, 25(1), 3–20.

[27] Bichler, G., Norris, A., Dmello, J.R., Randle, J., 2017. The impact of civil gang injunctions on networked violence between the bloods and the crips. Crime Delinq. 65 (7), 875–915.

[28] Bright, D., Koskinen, J., Malm, A., 2018a. Illicit network dynamics: the formation and evolution of a drug trafficking network. J. Quant. Criminol. 35 (2), 237–258.

[29] Lauchs, M., Keast, R., Yousefpour, N., 2011. Corrupt police networks: uncovering hidden relationship patterns, functions and roles. Polic. Soc. 21 (1), 110–127. Lauchs, M., Keast, R., Chamberlain, D., 2012. Resilience of a corrupt police network: the first and second jokes in Queensland. Crime Law Soc. Change 57 (2), 195–207.

[30] Shunmugapriya P, Kanmani S. (2013). “Optimization of stacking ensemble configurations through Artificial Bee Colony algorithm”. Swarm & Evolutionary Computation, 12(12):24-32.

[31] Chen Y.J, Wong M.L, Li H. (2014). “Applying Ant Colony Optimization to configuring stacking ensembles for data mining”. Expert Systems with Applications, 41(6):2688-2702.

[32] Han, H., Zhang, Z., Cui, X., et al., 2020. Ensemble learning with member optimization for fault diagnosis of a building energy system. Energy Build. 226, 110351.

[33] Peng T, Zhang C, Zhou J, Nazir MS. An integrated framework of Bi-directional long-short term memory (BiLSTM) based on sine cosine algorithm for hourly solar radiation forecasting. Energy. 2021;221: 119887.

[34] Huotari M, Arora S, Malhi A, Framling K. Comparing seven methods for state-of- health time series prediction for the lithium-ion battery packs of forklifts. Appl Soft Comput 2021;111: 107670.

[35] Liao Z, Zhang S, Li K, Zhang G, Habetler TG. A survey of methods for monitoring and detecting thermal runaway of lithium-ion batteries. J Power Sources 2019; 436: 226879.

[36] Chang, L. L., Zhang, L. M., & Xu, X. B. (2022). Randomness-oriented multi-dimensional cloud-based belief rule base approach for complex system modeling. Expert Systems With Applications, 203, Article 117283.

[37] Ruff, L., Kauffmann, J. R., Vandermeulen, R. A., Montavon, G., Samek, W., Kloft, M., et al., (2021). A unifying review of deep and shallow anomaly detection. Proceedings of the IEEE, 109(5), 756–796.

[38] Zhang, L. M., Wu, X. G., Zhu, H., & AbouRizk, S. M. (2017b). Perceiving safety risk of buildings adjacent to tunneling excavation: An information fusion approach. Automation in Construction, 73, 88–101.

[39] Tian, X., Fan, S., Huang, W., Wang, Z., Li, J., 2020. Detection of early decay on citrus using hyperspectral transmittance imaging technology coupled with principal component analysis and improved watershed segmentation algorithms. Postharvest Biol. Technol. 161, 111071.

[40] Bian, X., Wang, K., Tan, E., Diwu, P., Zhang, F., Guo, Y., 2020. A selective ensemble preprocessing strategy for near-infrared spectral quantitative analysis of complex samples. Chemometr. Intell. Lab. Syst. 197, 103916.

[41] Sirisomboon, P., 2018. NIR spectroscopy for quality evaluation of fruits and vegetables. Mater. Today: Proc. 5, 22481–22486.

[42] Javadi S, Dahl M, Pettersson MI. Vehicle detection in aerial images based on 3D depth maps and deep neural networks. IEEE Access 2021; 9: 8381–91.

[43] Ojha A, Sahu SP, Dewangan DK. VDNet: vehicle detection network using computer vision and deep learning mechanism for intelligent vehicle system. In: Proceedings of Emerging Trends and Technologies on Intelligent Systems: ETTIS. Singapore: Springer; 2021. p. 101–13.

[44] Cheng, Y., Wu, J., Zhu, H., et al., 2020. Remaining useful life prognosis based on ensemble long short-term memory neural network. IEEE Trans. Instrum. Meas. 70, 1–12.

[45] Chen, W., Xu, H., Jia, L., & Gao, Y. (2021). Machine learning model for bitcoin exchange rate prediction using economic and technology determinants. International Journal of Forecasting, 37(1), 28–43.

[46] Y. Miao, Q. Wang, M. Chen, et al. Spatial-spectral Hyperspectral Image Classification via Multiple Random Anchor Graphs Ensemble Learning, arXiv preprint arXiv: 2103. 13710, 2021.

[47] H. Su, Y. Yu, Q. Du, et al., Ensemble learning for hyperspectral image classification using tangent collaborative representation, IEEE Trans. Geosci. Remote Sens. 58 (6) (2020) 3778–3790.

[48] M. Zhang, Y. Li, X. Liu, X. Geng, Binary relevance for multi-label learning: an overview, Front. Comput. Sci. 12 (2) (2018) 191–202.