iJournals: International Journal of Software & Hardware Research in Engineering
ISSN-2347-4890
Volume 6 Issue 8 August, 2018

Prediction of Criminal Tendency of High-risk Personnel Based on Combination of Principal Component Analysis and Support Vector Machine

Author: Fang Yang[1]; Chunxue Wu[2]; Naixue Xiong[3]; Yan Wu[4]

Affiliation: School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology[1,2]; Department of Mathematics and Computer Science, Northeastern State University [3]; Public and Environmental Affairs, Indiana University Bloomington[4]

E-mail: yangfang930718@163.com[1]; wcx@usst.edu.cn[2]; xiongnaixue@gmai.coml[3]; wuyan8910@126.com[4]

ABSTRACT

The research on the big data in the security and protection industry has been increasingly recognized as the hotspot in case of the rapid development of the big data. This paper mainly focuses on addressing the problem that predicts the criminal tendency of the high-risk personnel based on the recorded behavior data of the high-risk personnel. Since reducing the dimension of data will improve the accuracy of the classification directly, we propose a predictive model that combines Principal Component Analysis (PCA) and Support Vector Machine (SVM) to predict the criminal tendency of high-risk personnel. In the model, at first, PCA is used to reduce the dimension of data preprocessed. And then SVM with different kernel functions will predict whether the high-risk personnel have a criminal tendency or not, moreover, the loss penalty parameter can be obtained by the k-fold cross-validation which could ensure the minimum prediction error. At last, theoretical analysis and simulation results prove that the accuracy could vary with the different kernel function in SVM. In addition, through the evaluation of Receiver Operating Characteristic (ROC) curve, the calibration diagram and the lift chart, the SVM under the radial basis kernel function outperforms the other kernel functions in terms of prediction accuracy.

Keywords: Big data, High-risk personnel, Accuracy, Principal component analysis, Support vector machine, K-fold cross-validation

1. INTRODUCTION

With the explosive growth of the global data, the big data has attracted considerable attentions from different research communities in recent years and those tremendous data could be generated and accumulated in all walks of life, such as the medicine, Internet of Things (IoT), and the industry [1]. To extract useful information from the massive data, data mining is usually used in which the data mining could help to obtain the information and patterns hidden in the data. The distributed mining for massive data usually depends on some technologies of cloud computing, such as the distributed processing, the distributed database and the cloud storage, as well as the virtualization technologies. However, how to handle these massive heterogeneous data in highly distributed environments becomes a huge challenge, especially in cloud platforms. Therefore, in [2] authors provided a cloud computing based functional framework that identifies the acquisition, management, processing and mining areas of IoT big data, and defined and described several associated technical modules in terms of their key characteristics and capabilities, as well as analyzed the current research in IoT application, identified the challenges and opportunities associated with IoT big data research. In recent years, with the rapid development of smart citiy, intelligent transportation and other industries, large-scale integration, large-scale networking, and cloud technologies have pushed the security and protection industry into the era of big data, i.e., the big data in the security and protection industry. In particular, video surveillance has become one of the main data sources of the big data in the security and protection industry, where the number of unstructured video data produced is huge and the growing speed is so fast, thus causing a series of problems in the application of the big data and even changing the application form of the security and protection industry. For instance, the traditional monitoring method was purely manual viewing of videos, while the big data of the security and protection can use the intelligent computer identification to analyze videos and pictures, so as to obtain information [3]. Those unstructured data, like videos and pictures, can be transformed into the structured data computer-recognized, and their feature can be extracted used to analyze and compare. In fact, it is necessary to unified organize and manage the massive data by both promoting the fusion of structured and unstructured data and building a data model. In such a huge data in security and protection industry, the daily behaviors and harmful behaviors of high-risk personnel can be recorded, by using them, the public security agencies can track the trajectory of high-risk personnel and focus on their abnormal actions by the identity identification, timed positioning, and behavior reproduction, thus further assessing their risks.

In fact, the big data prediction plays important roles in the application aspects of big data for statistical search, comparison, clustering, and classification that can be further utilized in the various fields. However, since the high dimension and complex structure of the big data, both of them will have an effect on the accuracy of the big data prediction, thus, how to reduce the dimension becomes a big challenge. To cope with it, many researchers proposed the method of reducing the data dimension, such as Linear Discriminant Analysis (LDA), LDA can establish the criterion by a training sample of a known category and will classify the data of unknown categories by predicting variables. In [4], authors proposed that LDA preserved as much of the class discriminatory information as much as possible to guarantee the accuracy of the big data prediction, and it is also worth to note that LDA may overfit data which means that it only performs the better efficiency for a particular data set, thus lacks generality. Due to the advantages of the less indicator selection and easier to implement, Principal Components Analysis (PCA) has considered to be another promising solution for reducing the data dimension in the big data prediction. In [5], authors pointed out that the PCA algorithm can remove more noise signals, reduce the influence of noise and other factors, thereby increasing the signal-to-noise ratio, and further obtain better detection results. Therefore, the accuracy of the big data prediction can be improved by using the PCA to reduce the data dimension in advance, on the other hand, to select an appropriate classification theory is another important issue to guarantee the accuracy. Support vector machine (SVM) is widely used to the pattern recognition, the classification and the regression analysis, and it also attracts more attentions because of its faster data processing speed and better accuracy. Such as in [6], authors presented that the use of SVM in conventional Brillouin optical time domain analyzer to directly extract temperature would achieve the faster speed and better accuracy than the conventional curve fitting methods. And except for the better accuracy, SVM exhibits a data processing speed faster than the commonly used the lorentzian curve fitting by two orders of magnitude. Based on the above analysis, to guarantee the accuracy, we propose a predictive model, i.e., the prediction of criminal tendency of high-risk personnel based on combination of Principal Component Analysis (PCA) and SVM. In this model, firstly, the PCA is used to analyze the principal component of data to reduce the dimension of data and further remove the redundant data. And then the SVM with different kernel functions can predict whether high-risk personnel have a criminal tendency or not, moreover, the loss penalty parameter can be obtained by the k-fold cross-validation which could ensure the minimum prediction error. Finally, the validation result shows that the accuracy will vary with the different kernel functions in SVM, in which the confusion matrix is used to describe the performance of the classification model, and through the evaluation of the ROC curve, the calibration chart and the lift diagram, the SVM under the radial basis kernel function outperforms the other kernel functions in terms of prediction accuracy.

The main contributions of this study can be considered as follows:

1. We propose the prediction of criminal tendency of high-risk personnel based on combination of PCA and SVM, and verify its performance under different kernel functions, as well as demonstrate the radial basis kernel function is the optimal with the highest accuracy.

2. The proposed model can be applied to the field of the big data in the security and protection industry to predict the classification for the criminal tendency of high-risk personnel.

2. RELATED WORK

At present, scholars mainly study and predict crimes in terms of the distribution of crime hotspots, the management and the control mechanisms of high-risk personnel. Yu Hongzhi et al. [7] proposed the improved fuzzy BP neural network method to predict the crimes because of the complex and influencing factors of crimes. Almanie et al. [8] presented an Apriori algorithm to obtain frequent crime patterns, and adopted decision trees and naive Bayes classifier methods to help predict crime events at specific time and location. S. Carolin Jeeva et al. [9] pointed out the association rules mining (apriori and predictive apriori) for online crimes such as phishing, and the rules obtained are interpreted to emphasize the features that are more prevalent in phishing URLs. Sujatha and Ezhilmaran [10] provided an effective stress intensity factor mining algorithm for predicting crime locations. Sun Feifei et al. [11] used a model-combining classifier named Random Forest, combined with the application of machine learning technology in crime prediction, and proposed an improved classification algorithm for predicting crimes.

Since Vapnik [12] established a complete SVM theory in 1995, this method has attracted attention both at home and abroad. Support Vector Machine can be regarded as a generalized linear classifier. Its basic idea is to transform the input space into a high-dimensional feature space through nonlinear transformation and find the optimal linear interface in the new space. It is widely used in comprehensive evaluation, prediction and other fields. For example, Hu Haiqing and others [13] assessed the credit risk of SMEs. Chi Guotai and others [14] established a credit evaluation model for small loans for farmers through support vector machine. Liu Liang and others [15] presented a multi-level fuzzy comprehensive evaluation method and situation analysis method to propose a qualitative and quantitative comprehensive warning model for emergencies. Mouad Zouina et al. [16] classify URLs using support vector machines and 5-fold cross-validation. Alireza Souri and others [17] point out data mining technology to detect malware, which helps researchers fully understand the field of malware detection. Tuba Parlar et al. [18] use the classifiers such as support vector machines and decision trees to propose new feature selection methods with better performance than other feature selection methods.

Currently there are relatively few studies on the management and control mechanisms of high-risk personnel. The research on high-risk personnel focuses on the status quo and the causes of the exploration phase. There are few studies on risk identification and judgment, and they are in the initial stage. There is no specific measurement method and system, and there is no established and authoritative measurement method and system. Because of its good properties, Support Vector Machine can be widely applied to the prediction field, but the attention to high-risk personnel is insufficient. Therefore, in this paper, based on the principal component analysis method, a large amount of information is used to establish a predictive model using support vector machine to predict whether a high-risk person has a criminal tendency.

3. METHODS

3.1 System framework

In the framework, firstly, when generating the data simulation, the data can be preprocessed by data cleaning, data transformation, data protocol, and so on. Because of the high dimension of the dataset, the correlation between the data also increases, which may lead to reduce prediction performance of the model. Therefore, the main component analysis method is used to reduce the dimension of the data, so that the correlation of the data set after dimension reduction will decrease, and the prediction performance will further increase. Then, support vector machine with different kernel functions are used to predict the classification results, and the optimal parameters can be obtained by 10-fold cross-validation to calculate sensitivity, specificity, accuracy, and confusion matrix. Finally, the model is evaluated by the receiver operating characteristic curve, the calibration diagram and the lift diagram, moreover, the prediction performances of SVM under different kernel functions are compared.

Based on the above analysis, the high-risk personnel risk early-warning model can be established, which is divided into data preprocessing, model tuning, and model evaluation, as shown in Figure 1.

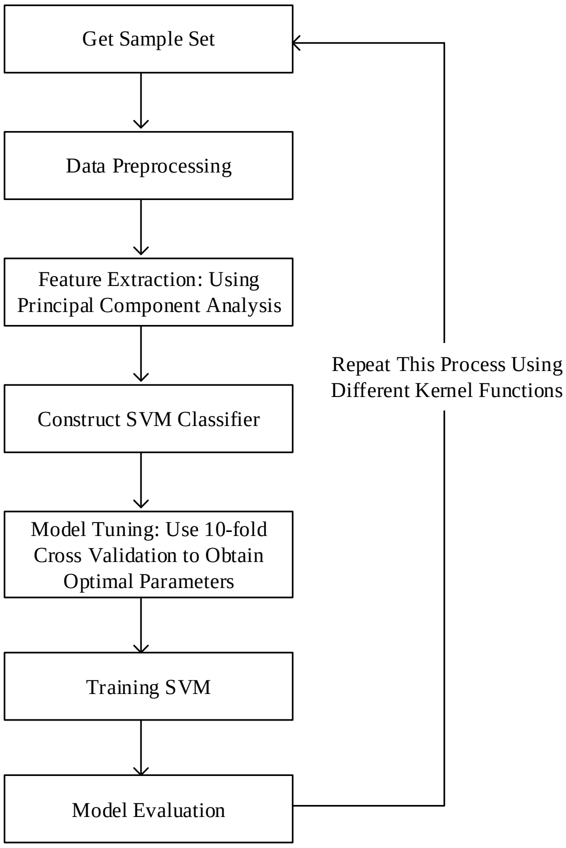

Fig 1: The framework of proposed model

3.2 Proposed methodology

Support vector machine has been used to deal with classification problems. However, as the amount of data increases, the dimension becomes higher and higher, which makes the classification effect less optimistic. The principal component analysis can be used to reduce the dimension of data, therefore, this paper combines the principal component analysis and the support vector machine to deal with high-risk personnel classification, in which the optimal parameter values can be obtained by using the 10-fold cross-validation under the different kernel functions.

3.2.1 Principal component analysis

Principal component analysis is a common technique used to the data dimension reduction, which aims to find a linear combination of some original predictor variables and capture most of the predictor variance. The first principal component is the combination that captures the variance with the greatest predictor variable among all possible linear combinations. The subsequent principal components are in turn irrelevant to the previous principal components and capture the linear combination with the maximum predictor variable among the residual variances.

Consequently, the main idea of the principal component analysis is to construct a linearly independent covariance matrix between the features based on the original sample data, and calculate the eigenvalues of the covariance matrix, as well as construct an optimal projection matrix based on the larger number of eigenvalues and the corresponding eigenvectors, so that the original sample data is mapped to the new feature space according to the projection matrix, thus achieving the goal of dimension reduction [19]. The feature parameter reduction process is as follows:

1) Suppose there are *r* sample features, the total number of samples is *N*, that is, each sample is *r*-dimension, and each sample is a one-dimension row vector, and $X=(a_1,a_2,\ldots,a_r)$ constitutes a sample matrix *X* with *N* rows and *r* columns.

2) Calculate the average vector of all samples:

$$m=1/N\sum_{i=1}^{N}x_i \quad (1)$$

3) Calculate the covariance matrix:

$$C_x=1/N\sum_{i=1}^{N}(x_i-m)^T(x_i-m) \quad (2)$$

4) Find the eigenvalues of the covariance matrix C_x. λ_1, λ_2, . . . , λ_r ($\lambda_1>\lambda_2>\ldots>\lambda_r$) correspond to the feature vector α_1, α_2, . . . , α_r, respectively.

5) Sort the eigenvalues in a decreasing order and calculate the cumulative contribution rate of the first *p* principal components. It can measure the extent of which the new composite component interprets the original data. Generally, the cumulative contribution rate is greater than 85% [20], and it can be calculated by:

$$\eta(p)=\sum_{i=1}^{p}\lambda_i/\sum_{i=1}^{r}\lambda_i,\ p<r \quad (3)$$

6) The transformation matrix *T* is constructed based on the feature vector corresponding to the previous *p* largest eigenvalue $T=(b_1,b_2,\ldots,b_p),\ p<r$.

7) By using $Y=\mathrm{XT}$, the first *p* principal components are calculated, and the original *r*-dimension feature can be transformed into a *p*-dimensional feature, in this way, it not only preserves the original information, but also reduces the dimension.

3.2.2 Support vector machine

Support vector machine is a statistical model that was first developed by Vladimir Vapnik. In the following years, the model has evolved into one of the most flexible and effective machine learning tools [21]. The mechanism of SVM is to find an optimal classification hyperplane that satisfies the classification requirements, so that the hyperplane can maximize the blank area on both sides of the hyperplane while ensuring the classification accuracy.

In this paper, the Radial, Polynomial, and Sigmoid models will be used to build the model.

1）The number of dimensions of the feature space after the radial basis kernel function transformation is actually infinite, which is due to the infinite number of terms in its expansion [22]. Its mathematical expression is

$$\exp(-\sigma \| x-u \|^2) \tag{4}$$

2）The polynomial kernel function actually transforms the feature space into a higher-dimensional space [23]. The polynomial kernel function is a kind of the global kernel functions, which allows data points with long distances to affect the value of the kernel function. It is worth to note that both the greater degree and the higher dimension of the map will lead to a more amount of computation. If the degree is too large, the learning complexity becomes so high correspondingly, thus occurring the phenomenon of over fitting easily. Its mathematical expression is

$$(scale(x'u)+1)^{degree} \tag{5}$$

3）S-type kernel functions, Sigmoid kernel functions are derived from neural networks and are widely used in deep learning and machine learning [24]. When using the Sigmoid function as a kernel function, a multi-layer perceptron neural network is implemented by the support vector machine. Its mathematical expression is

$$\tanh(scale(x'u)+1) \tag{6}$$

3.2.3 K-fold cross-validation

Because the classifier is trained on a given data set, it may only obtain higher classification accuracy for this training data set than other independent data sets. To avoid overfitting, the k-fold cross-validation will be used to tune the model [25]. In this method, the samples are randomly divided into k equal size subsets. In the first step, the model is fitted with all except the first subset (first fold), then the first set of samples reserved is predicted and its corresponding results will be used to evaluate the model. Next, the first subset is returned to the training set, the second subset is reserved for model evaluation, and so on. The resulting k model evaluation results will be summarized (usually the mean and standard deviation), and then it can be inferred the relationship between the optimal parameters and model performance. In this paper, we use the 10-fold cross-validation to determine the loss penalty parameter with the minimum prediction error.

4. BUILDING MODEL

4.1 Data preprocessing

In data mining algorithm engineering applications, data preprocessing technology plays an important role in the data mining to guarantee the accurate models and high quality data. The common data preprocessing process includes data cleaning, data transformation, data integration, and data protocol [26]. Data cleaning refers to the elimination of irrelevant data, filling in missing values, smooth noise data, identifying and deleting outliers on target data samples. Data transformation is to transform the data representation into another equivalent form, which makes the data more standardized and improves data mining efficiency. The data transformation mainly reduces the dimension of the data, eliminates the invalid attributes, finds the really useful feature attributes, and correctly judges the attribute types, so the data mining can be processed and calculated later. Some of the attributes in this article contain a large number of different and unordered values, which should be generalized to a higher level of concept and reduce the number of different values of the attribute. Data integration mainly aims to the processing operations of merging target data samples that exist in different databases and cause data conflicts and inconsistencies. Data protocol represents the process of maximizing the processing of target data samples on the basis of the useful features of the mining target and the understanding of the content of the data itself. It mainly includes the selection of reference attributes of data samples and data processing of data sampling [27].

4.2 Feature extraction

In order to reduce the dimension of the feature, this paper uses the principal component analysis to remove the redundant data. Since each feature unit is not uniform, data standardization is required before the data is reduced, and the data is transformed into -1 to +1. By counting the cumulative contribution rate, it can be concluded that the cumulative contribution rate of the first 8 principal components can reach 85%. Therefore, this paper selects the first 8 principal components to reduce the 14-dimension to 8-dimension on the basis of retaining the original information, it is shown in Figure 2.

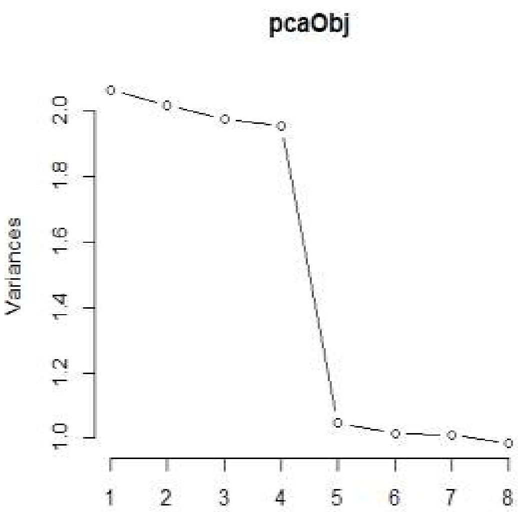

Fig 2: Main ingredient gravel

It can be seen that the most obvious change in slope is at the fifth principal component, therefore, the first four principal components should be preserved. We observe the cumulative contribution rate of the principal component to ensure that the principal component can explain at least 85% of the total variance, as shown in Table 1.

Table 1. Variations in the interpretation of the original data for each component and cumulative distribution of the variation

	Standard Deviation	Variance Ratio /%	Cumulative Contribution Rate /%
PC1	1.4370	14.75	14.75
PC2	1.4203	14.41	29.16
PC3	1.4057	14.41	43.27
PC4	1.3985	13.97	57.24
PC5	1.02292	7.474	64.718
PC6	1.00764	7.252	71.97
PC7	1.00555	7.222	79.193
PC8	0.9921	7.03	86.22

The table shows the standard deviation, variance ratio, and cumulative contribution rate of the first eight principal components. It can be seen that the variance ratio of the first four principal components is larger than that of other principal components, but the cumulative contribution rate is less than 85%. However, the cumulative contribution rate of the first eight principal components is more than 85%, and the principal component gravel chart should retain the first eight principal components.

4.3 Tuning model

Through the 10-fold cross-validation, the loss penalty parameter is obtained when the data set's prediction error is minimized. Figure 3 shows the loss penalty parameters obtained for different radial gamma kernel functions. The darker the color, the smaller prediction error it will be. The gamma values and C under different kernel functions is shown in Table 2. Thus, it can be concluded that when both the gamma value is equal to 0.1 and C is equal to 10, the model is optimal.

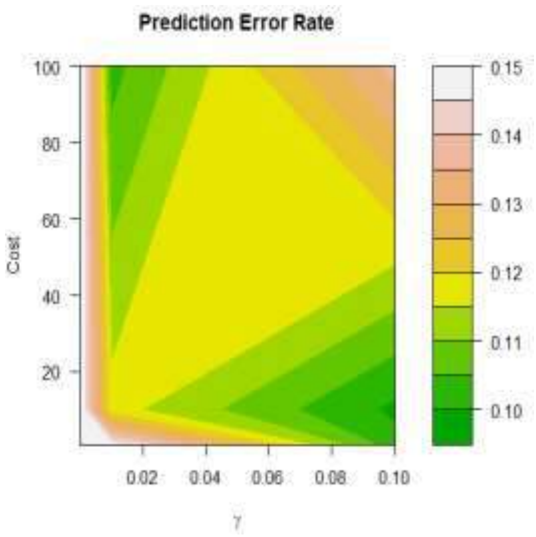

Fig 3: Diagram of loss penalty parameters obtained with different gamma values for radial basis kernel functions

Table 2. Parameter values for different kernel functions

	Gamma	C
Radial	0.1	10
Polynomial	0.1	1
Sigmoid	1e-06	1

The detailed design of prediction model of criminal tendency of high-risk personnel based on combination of PCA and SVM is shown in Algorithm 1.

Algorithm 1. The prediction model of criminal tendency of high-risk personnel based on combination of PCA and SVM

Input: Sample set data0 // The data0 is a matrix of *N* row and *r* column

1: **begin**

2: calculate the mean value; // According to Equation (1)

3: calculate the covariance matrix; // According to Equation (2)

4: calculate the eigenvalues and the feature vector;

5: sort the eigenvalues in a decreasing order;
6: save the first *p* feature vectors with larger eigenvalues; // According to Equation (3)
7: output the sample set data1; // The data1 is a matrix of *N* row and *p* column
8: spilt the data1 into the train set data2 and test set data3;
9: svm. train (data2, Y, kernel type, C, gamma); // Y is the result of the classification
10: svm. test (data3, Yt, kernel type)
11: **end**

5. EXPERIMENTS AND RESULTS

5.1 Modeling tool

The implementation tool used in this paper is R language [28]. R is an open source, free and excellent statistical software. It originated from AT&T Bell Laboratories' S language and has a stronger object-oriented function than other statistical or math-specific programming languages. The software is widely used abroad. For example, Google and Facebook are using it for data analysis work.

5.2 Data preparation

Based on the sensitivity of real-life data, 3333 pieces of data were generated by simulation in this paper, and the same feature attributes as the real data were retained. Among these data, 2315 data are training sets. This paper will use re-sampling techniques such as k-fold cross-validation to fit the data, and the remaining 1018 will be used as the test set. For example, there is a "hometown" of redundant attributes that has nothing to do with the prediction results. This attribute has no practical significance for the study of this article.

	Radial basis function		Polynomial kernel function		S-type kernel function	
	Yes	No	Yes	No	Yes	No
Yes	95	24	88	15	0	0
No	46	853	53	862	141	877

Because this article only studies the situation in a certain area, it has nothing to do with the "original" property, so it is deleted. For example, the data in the "home address full name" generalized to "whether it is a local person".

5.3 Experimental results

The sensitivity of the model is the ratio that is accurately determined as "occurrence" in all observed "occurring" samples. The sensitivity is sometimes referred to as the true positive rate. Specificity refers to the rate at which the observed sample is not accurately judged as "not occurring". The false positive rate is defined as 1 minus specificity. Table 3 shows the sensitivity, specificity, and accuracy under different kernel functions.

Table 3. Sensitivity, specificity, and accuracy for different kernel functions

	Sen	Spe	Acc
Radial	0.674	0.973	93.1%
Polynomial	0.624	0.983	93.3%
Sigmoid	0	1	86.2%

A common method of describing classification model performance is the confusion matrix [29]. The confusion matrix is a simple cross-tabulation of observation and prediction classes. The elements on the diagonal represent the correct predictions, and the off-diagonal elements represent the number of samples corresponding to each possible false positive. The following shows the confusion matrix classified in the case of different kernel functions. The confusion matrix under different kernel functions are shown in Table 4 below.

1）Radial basis kernel function: The support vector opportunity is used by default to use the radial basis kernel function, where Yes represents a criminal tendency sample, and No represents a sample without criminal tendency.

2）Polynomial kernel function: The polynomial kernel is suitable for orthonormalization (vector orthogonality and modulo 1) data.

3）S-type kernel function: The theoretical basis of support vector machine (convex quadratic programming) determines that the S-type kernel function is ultimately obtained as a global optimum rather than a local optimum, and it also ensures a good generalization ability of unknown samples.

Table 4． Confusion matrix under different kernel function

5.4 Performance evaluation of different kernel functions

The ROC curve is a method of using the probability to compare models, which determines a valid threshold given a series of consecutive data points. A value that exceeds this threshold represents a specific event [30].The closer the ROC curve is to the upper left corner, the higher the accuracy of the classification model will have. AUC is the area under the ROC curve. The better the classification performance, the closer the ROC is to the upper left corner and the larger the AUC is. In this test set, the area under the ROC curve for the radial basis kernel function is the largest (0.9285), while the area for the S-type kernel function is the smallest (0.7967). The ROC curves for these three models are shown in Figure 4.

One method of assessing probabilistic quality is a calibration graph. For a given data set, the calibration graph shows the amount of contrast between the observed probability of the event and the predicted class probability [31]. Score the model using the data set as a test set and create a calibration plot, as shown in Figure 5. The calibration chart shows the midpoint of the box as the x-axis and the observation event rate as the y-axis plot. If the point falls on a diagonal along 45°, the model has a good calibration rate. It can be seen from the figure that the S-type kernel function deviates sharply from the diagonal, and the radial basis kernel function almost coincides with the diagonal line. Therefore, from the calibration chart, the prediction performance under the S-type kernel function is the worst, and the radial basis function has the best prediction performance.

Lift charts are a visual tool for assessing the ability of a model to determine classification data sets. Assume that there is a set of samples that use the event class probability to score M events. When the samples are sorted according to the target class probability, sorting of target samples is higher than that of the non-target samples. The role of the lift chart is to rank the samples according to the score, and calculate the cumulative target sample rate as more samples are evaluated. The horizontal axis of the lifted graph is the cumulative test sample percentage and the vertical axis is the cumulative gain/boost. Figure 6 is a diagram of the elevation of the test set probability for different kernel functions. Similar to the ROC curve, by comparing the lifting graphs of different kernel functions to search for the most suitable model, the area under the graph curve can be used to quantify the model effect. It can be seen from the figure that the model prediction performance under the radial basis function is optimal, while the model prediction performance under the S-type kernel function is the lowest.

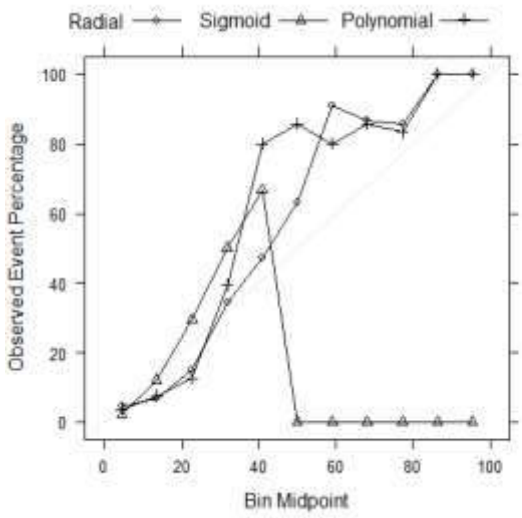

Fig 4: ROC curves for different kernel functions

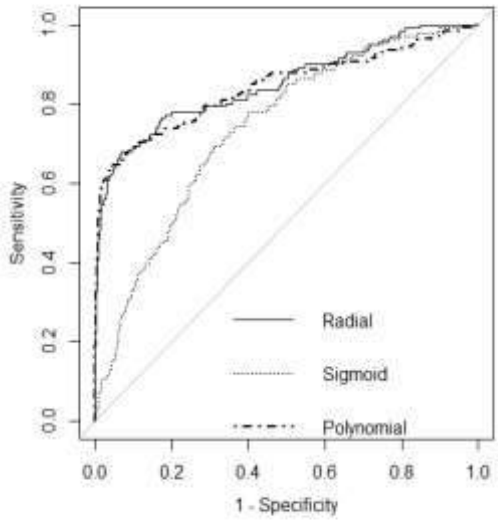

Fig 5: Calibration plot for different kernel functions

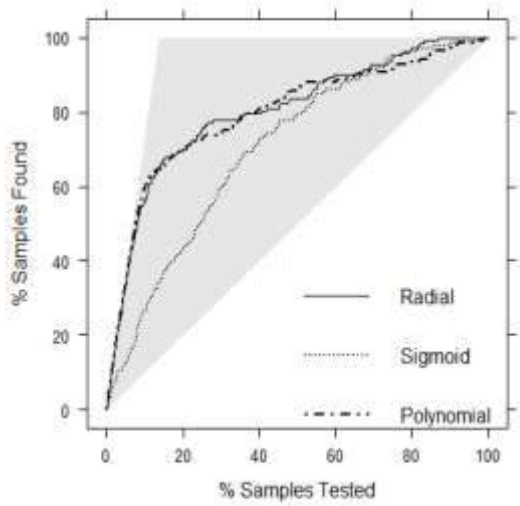

Fig 6: Lift chart with different kernel functions

6. CONCLUSIONS AND FUTURE WORKS

In this paper, firstly, principal component analysis is used to reduce the dimension of the behavior data. Then, the support vector machine under different kernel functions will predict whether the high-risk personnel have criminal tendency or not. Finally, the model is tuned by the k-fold cross-validation method, and the results show that support vector machine can be used to build the early-warning risk prediction model about the high-risk personnel with a significant forecasting effect. Based on different kernel functions, the model shows different prediction accuracies and effects, particularly, the support vector machine under the radial basis kernel function has the best prediction accuracy, while the support vector machine kernel function under the S kernel function performs the worst. In fact, support vector machine can perform the excellent small sample learning and better generalization ability, and it is suitable for risk assessment of high-risk personnel. Therefore, in the proposed model, public security agency can identify high-risk groups with a high efficiency and judge whether the high-risk personnel have criminal behavior or not. The future work would build the high-risk personnel risk prediction model with higher accuracy, in addition, the decision tree can be used to excavate the general characteristics of people that have a criminal tendency.

7. ACKNOWLEDGMENTS

The authors would like to appreciate all anonymous reviewers for their insightful comments and constructive suggestions to polish this paper in high quality. This research was supported by Shanghai Science and Technology Innovation Action Plan Project (16111107502, 17511107203) and Shanghai key lab of modern optical system, as well as support from National Natural Science Foundation of China (No.61502220), and the program for tackling key problems in Henan science and technology (No.172102310636) for funding.

8. REFERENCES

[1]. X. Li and X. Li, "Big Data and Its Key Technology in the Future," in Computing in Science & Engineering, vol. 20, no. 4, pp. 75-88, Jul./Aug. 2018.

[2]. H. Cai, B. Xu, L. Jiang and A. V. Vasilakos, "IoT-Based Big Data Storage Systems in Cloud Computing: Perspectives and Challenges," in IEEE Internet of Things Journal, vol. 4, no. 1, pp. 75-87, Feb. 2017.

[3]. D. Mitra and A. Acharyya, "Big Data and Deep Learning -A review," in IJournals: International Journal of Software & Hardware Research in Engineering, vol. 4, no. 1, pp.28-30, Jan. 2016.

[4]. S. Shitole and S. Jadhav, "Recognition of handwritten devanagari characters using linear discriminant analysis," 2018 2nd International Conference on Inventive Systems and Control (ICISC), Coimbatore, 2018, pp. 100-103.

[5]. C. Sun, Y. Wang, P. Wan and Y. Du, "A cooperative spectrum sensing algorithm based on principal component analysis and K-medoids clustering," 2018 33rd Youth Academic Annual Conference of Chinese Association of Automation (YAC), Nanjing, China, 2018, pp. 835-839.

[6]. H. Wu, L. Wang, Z. Zhao, C. Shu and C. Lu, "Support Vector Machine based Differential Pulse-width Pair Brillouin Optical Time Domain Analyzer," in IEEE Photonics Journal.

[7]. H. Yu, F. Liu and K. Zou, "Improved fuzzy BP neural network and its application in crime prediction," in Journal of Liaoning Technical University, 2012.

[8]. T. Almanie, R. Mirza and E. Lor, "Crime Prediction Based On Crime Types And Using Spatial And Temporal Criminal Hotspots," in Computer Science, vol. 5, no. 4, 2015.

[9]. S.C. Jeeva and E.B. Rajsingh, "Intelligent phishing url detection using association rule mining," in Human-centric Computing and Information Sciences, vol. 6, no. 1, pp. 1-19, 2016.

[10]. R. Sujatha and D. Ezhilmaran, "A new efficient SIF-based FCIL (SIFâFCIL) mining algorithm in predicting the crime locations," in Journal of Experimental & Theoretical Artificial Intelligence, vol. 28, no. 3, pp. 561-579, 2015.

[11]. F. Sun, Z. Cao, X. Xiao and S.O. Counterterrorism, "Application of an Improved Random Forest Based Classifier in Crime Prediction Domain," in Journal of Intelligence, 2014.

[12]. C. Cortes and V. Vapnik, "Support-vector networks," in Machine Learning, vol. 20, no. 3, pp. 273-297, 1995.

[13]. H. Hu, L. Zhang and D. Zhang, "Research on SMEs Credit Risk Assessment from the Perspective of Supply Chain Finance- A Comparative Study on the SVM Model and BP Model," in Management Review, 2012.

[14]. G. Chi, M. Pan and Y. Cheng, "Credit Rating Model of Small Loans for Farmers Based on

Comprehensive Discriminate Capacity," in Management Review, 2015.

[15]. L. Liu, Y. Shen, X. Cao, J. He and S. Li, "Research on Early-warning Model of Unconventional Emergencies Based on Key Information," in Management Review, 2012.

[16]. M. Zouina and B. Outtaj, "A novel lightweight URL phishing detection system using SVM and similarity index," in Human-centric Computing and Information Sciences, vol. 7, no. 1, 2017.

[17]. A. Souri and R. Hosseini, "A state-of-the-art survey of malware detection approaches using data mining techniques," in Human-centric Computing and Information Sciences, vol. 8, no. 1, 2018.

[18]. T. Parlar, S.A. Özel and F. Song, "QER: a new feature selection method for sentiment analysis," in Human-centric Computing and Information Sciences, vol. 8, no. 1, 2018.

[19]. Z. Lin, "Research on Cultivated Land Quality Evaluation Based on PCA and Decision Tree Model," in Journal of Fujian Normal University, 2008.

[20]. H. Abdi and L.J. Williams, "Principal component analysis," in Wiley Interdisciplinary Reviews Computational Statistics, vol. 2, no. 4, pp. 433-459, 2010.

[21]. T.S. Furey, N. Cristianini, N. Duffy, et al. "Support vector machine classification and validation of cancer tissue samples using microarray expression data," in Bioinformatics, vol. 16, no. 10, pp. 906-14, 2000.

[22]. Y.J. Oyang, S.C. Hwang, Y.Y. Ou, C.Y. Chen and Z.W. Chen, "Data classification with radial basis function networks based on a novel kernel density estimation algorithm," in IEEE Transactions on Neural Networks, vol. 16, no. 1, pp. 225-236, 2005.

[23]. A. Gunasekaran, "Occupant detection using support vector machines with a polynomial kernel function," in Proc Spie, pp. 270-277, 2000.

[24]. G. Camps-Valls, J.D. Martín-Guerrero, J.L. Rojo-lvarez and E. Soria-Olivas, "Fuzzy sigmoid kernel for support vector classifiers," in Neurocomputing, vol. 62, pp. 501-506, 2004.

[25]. R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in International Joint Conference on Artificial Intelligence, vol. 14, pp. 1137-1143, 1995.

[26]. H. Fang, Z. Zhang, C. Wang, et al. "A survey of big data research," in IEEE Network, vol. 29, no. 5, 2015.

[27]. Y. Tang, "Recidivism Association Rule Mining Based on Apriori Algorithm," in Command Information System and Technology, 2016.

[28]. RossIhaka and RobertGentleman, "R: A Language for Data Analysis and Graphics," in Journal of Computational & Graphical Statistics, vol. 5, no. 3, pp. 299-314, 1996.

[29]. A.M. Hay, "The derivation of global estimates from a confusion matrix," in International Journal of Remote Sensing, vol. 9, no. 8, pp. 1395-1398, 1988.

[30]. D.G. Altman and J.M. Bland, "DIAGNOSTIC-TESTS-3 - RECEIVER OPERATING CHARACTERISTIC PLOTS .7[J] ," in Bmj Clinical Research, vol. 309, no. 6948, pp. 188-188, 1994.

[31]. M. Vuk and T. Curk, "ROC curve, lift chart and calibration plot," in Metodoloski Zvezki, vol. 3, no. 3, pp. 89-108, 2006.

