Security Monitoring Data Fusion Method Based on ARIMA and LS-SVM

Security Monitoring Data Fusion Method Based on ARIMA and LS-SVM

Kaiwen Xu, Jin Yu, Yanzhu Hu and Xinbo Ai*

School of Automation, Beijing University of Posts and Telecommunications, Beijing 100876, China

*Corresponding author e-mail: axb@bupt.edu.cn

Abstract. Using the Autoregressive Integrated Moving Average Model (ARIMA) and least squares support vector machine model (LS-SVM), the data of the security monitoring data obtained during the security supervision process is data fusion, and the data is reduced by data. After the components are analyzed, the accident prediction is performed based on the improvement of data processing efficiency. Finally, the main data analysis (PCA) of the 15-dimensional data is used to reduce the dimension to the 7-dimensional data based on the accuracy of the information. After that, the data fusion technology is used to fuse the data to establish the ARIMA-LS-SVM combination. The model uses the combined forecasting model to predict and analyze the safety production accidents, and uses the actual data to verify. The results show that the data fusion technology can improve the efficiency of data processing. The model fits the time series of safety accidents well. The high prediction accuracy can help the company's safety production accident prediction in the future.

1. Introduction

In the production process of the enterprise, the safety of the production process has always been a key topic of discussion. Although the amount of big data collected in security supervision is large, most of the production security management systems are subsystems that are independent of each other, resource sharing and data structures and formats are not the same, although they are collected and presented. The amount of data is large, but it is difficult to achieve a unified combination of these fragmented data information.Regarding the integration of safety production data, the most critical part is the data fusion of the data mastered. There are many types of data about safety production monitoring, and the Chinese vocabulary, English vocabulary, voice signal, picture and other data are more different. Data fusion is the focus of future research.

At present, some researchers in the world have studied related data fusion technologies and algorithms. Aisha Siddiqa focused on related technologies such as heterogeneity and scalability of data [1]. Furqan Alam and other mathematical methods combined with a specific IoT environment, comprehensive analysis of the Internet of Things data fusion technology and applications [2]. In the multi-source data fusion technology, Sadia Din comprehensively analyzes the decision-making of data through centralized clustering and distributed clustering techniques [3]. Zhao Yuling Ma Ke proposed D-S theory and Bayesian theory and data fusion model [4-5]. Liu Xiaowu proposed a support vector machine SVM as a fusion model to fuse heterogeneous sensor data [6]. Enrico Bocca Ann S. Barry

and so on through the fusion of heterogeneous data to determine the level of security risk [7-8]. Kellyn Rein proposed a method of combining high-level and low-level data through multiple models to achieve more accurate analysis decisions [9]. S. Sutor combine heterogeneous data containing multiple signals to enhance the video security surveillance effect [10]. Liu Mixia and Nicklaus A. Giacobe based on network data security, using D-S evidence theory to analyze the data [11-12]. Wang Huiqiang explored a big data fusion system based on artificial neural networks [13]. A.R. Newman studied the throughput problem in the data fusion process and improved it to enhance the performance of the data test [14]. Zhenyu Lu analyzed the relationship between urban and urban security through data fusion of urban planning and urban security monitoring [15]. However, in the above data fusion research, the targeted areas do not involve safe production, and the data fusion method does not distinguish between the high frequency and low frequency characteristics of the signal.

In order to achieve better data and more accurate integration, this paper adopts the method of hierarchical processing. First, after eliminating the useless information such as dirty data for the pre-processing of the big data of the security supervision, the data is migrated and sorted, and the data fusion processing is performed while considering the redundancy and noise of the data; the corresponding picture information is processed. After dimension reduction processing, it is merged with data, and data mining and data analysis are performed on the data through LS-SVM and ARIMA hierarchical model. The effective information analyzed by the improved algorithm model is used for prediction, and the information is used for inference and decision making. Finally, through the consistency test, the fusion goal is achieved while having good performance.

2. Multi-source heterogeneous data fusion and safety production prediction model research method

Since the big data of safety production comes from different data sources. The data difference is very significant for the analysis of the data and the future safety rating and safety supervision decision. Impact. ARIMA is a classic time series prediction algorithm, which is suitable for solving linear problems. However, when the data is a nonlinear model, the ability to deal with nonlinearity by the SVM model can make up for this defect. Based on this paper, a data fusion hierarchical model including time series model (ARIMA) and least squares support vector machine (LS-SVM) algorithm model is constructed. The principle are shown in Figure 1, and a large number of data is used to build a training model for large scale. Sample training. And through the test, training and test of the sample to test the validity and robustness of the model, select a better data fusion classification model for safety production rating.

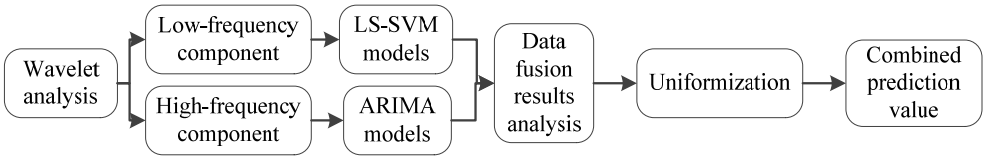

Figure 1. ARIMA and SVM combined prediction model

2.1. Data feature selection based on Pearson correlation coefficient

Due to the large number and variety of big data in the security supervision, it is necessary to preprocess the original data to ensure the operability of the data. Feature selection is the process of selecting the most relevant data features from the large-scale data and the most valuable data features for the final decision, thus reducing the dimensionality of the data. This paper selects the characteristics of the data based on the Pearson correlation coefficient.

In statistics, the Pearson correlation coefficient is a function used to calculate the correlation between two variables, and its calculation formula is:

$$\rho_{X,Y} = \frac{COV(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \tag{1}$$

Among them $COV(X,Y)$ is the covariance of vector X and vector Y. σ_X and σ_Y is the standard deviation of vector X and vector Y.

The Pearson correlation coefficient ranges from -1 to +1. When the value is 0, it means that the two variables are irrelevant. A positive value indicates a positive correlation, and a negative value indicates a negative correlation. A larger value indicates a larger correlation. The stronger the correlation, the greater the coefficient value above 0.7 is strongly correlated. After screening, the relevant characteristics of the same safety supervision are: hidden trouble investigation, hidden danger type, hidden danger grade, industry type, administrative punishment and other 15 data.

2.2. PCA-based data dimensionality analysis

Because the safety production monitoring data has a certain autocorrelation, it will affect the training effect of the model and generate noise interference. To avoid over-fitting, data dimensionality reduction processing is usually used to generate a new feature data set for training analysis. In this paper, Principal Component Analysis (PCA) is used for data dimensionality reduction. The principle is to map some high-dimensional data into low-dimensional space by means of current projection. The variance of projection dimension is the application effect of this kind of dimensionality reduction method. The main factor. When the variance of the projected dimensional data is large, the principal component analysis method can retain more original data point features with less data dimensions.

In view of the principle of principal component analysis, based on the found set of orthogonal bases, the idea of transforming the rearmost difference is maximized, thereby eliminating redundant data. Since the projection is selected with the dimension with the largest variance, the maximum amount of information can be stored in the projection space. Let $PX = Y$ .the P row elements form new features.

Characteristic decomposition for positive definite matrix XX^T. Let $P = I^T$. Then there is:

$$XX^T = IDI^T \tag{2}$$

$$C = \frac{P(XX^T)P^T}{n-1} \tag{3}$$

Where a is I unit matrix and D is a diagonal matrix characterized by diagonal elements. Since P is an orthogonal basis, then $P^T = P^{-1}$.

$$C = \frac{P(P^T DP)P^T}{n-1} = \frac{(PP^{-1})D(PP^{-1})}{n-1} = \frac{D}{n-1} \tag{4}$$

Using the PCA dimension reduction method, several vectors with the largest eigenvalue are selected as the projection direction, so that the data variance in the projection direction is the largest. The selection of the number of features in this paper can reflect the original information about 95% as the benchmark. The PCA dimensionality reduction results are shown in Figure 2. The contribution rate can reach 98% or more when the number of principal components is 7. The data fell to seven dimensions.

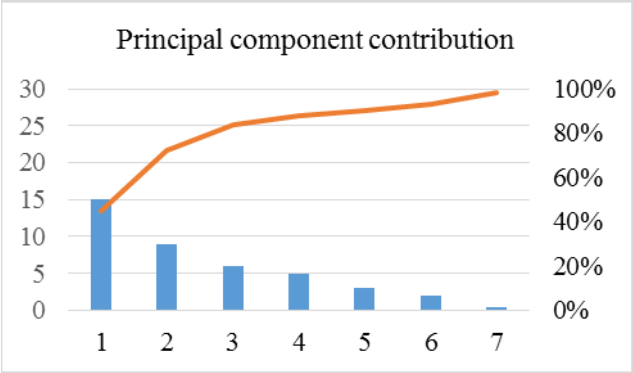

Figure 2. PCA dimensionality reduction principal component contribution rate and eigenvalue

2.3. Autoregressive Integrated Moving Average Model (ARIMA)

The Autoregressive Integrated Moving Average Model is a time series forecasting method. It is widely used in the natural and scientific fields and has the characteristics of short-term prediction accuracy. The basic idea of the ARIMA model is to use the time series generated by the prediction object as a random sequence, and use some mathematical models to achieve the effect of approximating the description. Once the appropriate mathematical model is determined, the past and present values generated by the sequence can be used. Predict future values. The precondition for using the ARIMA model is that the time series used for prediction is a stationary sequence, and it is reflected in the figure that all the data points all fluctuate up and down around a certain horizontal line.

In the ARIMA model, the smooth time series $\{y_t\}$ is satisfied

$$y_t = \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + \cdots + \Phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \tag{5}$$

In the equation. $\{\varepsilon_t\}$ is a independent and identically distributed random variable sequence. Φ_i and θ_j $(i = 1, 2, \cdots, p; j = 1, 2, \cdots, q)$ are the parameters of $\{y_t\}$ and $\{\varepsilon_t\}$ respectively. p is autoregressive parameter and q is moving average order.

In the actual context of safety production monitoring, the collected time series are generally non-stationary time series, and often exhibit periodicity or trend, such as the safety supervision situation when the accident occurs or when the inspection is intensive. Usually a lot are better. For non-stationary time series, the differential transforms should be used to transform it into a smooth time series, and then the corresponding ARIMA model is established, which is recorded as $ARIMA(p, d, q)$.

2.4. Least squares support vector machine Model (LS-SVM)

The least square support vector machine model was originally used to solve the pattern recognition problem and then extended to the regression estimation. The least square support vector machine improves and extends the support vector machine model. The least squares linear system are used as the loss function of the system, and the inequality constraint in the conventional SVM model is replaced by the equality constraint, thus transforming the quadratic programming problem into linear. The equations solve the problem, improve the training speed of the algorithm, and do not need to consider the insensitive loss function.

Suppose the sample set $(x_i, y_i) \in R^n \times R$ $(i = 1, 2, \cdots, l)$, where x_i is the i th input sample and y_i is the i th output sample. The training sample set (x_i, y_i) of the original input space R^d is mapped to a high-dimensional feature space by the nonlinear function $\Phi(\cdot)$, and the nonlinear prediction model is expressed as:

$$y = w^T \cdot \Phi(x_i) + b \tag{6}$$

Where w is the feature space weight coefficient vector, $\Phi(x_i)$ is the nonlinear map, and b is the offset.

At the same time, according to the principle of structural risk minimization, the regression problem of LS-SVM can be expressed as a constrained optimization problem:

$$\min \frac{1}{2} w^T w + \frac{1}{2} C \sum_{i=1}^{n} \xi_i^2 \tag{7}$$

$$s.t. \quad y_i = w^T \Phi(x_i) + b + \xi_i \quad i = 1, 2, \cdots, l$$

Where C is a penalty factor and ξ_i is an error variable. The constraint optimization problem can be solved by the Lagrange multiplier method, which can be obtained as follows:

$$\begin{bmatrix} 0 & I_n^T \\ I_n & \Omega + C^{-1} I_n \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \tag{8}$$

Lagrange multiplier $\alpha = [\alpha_1, \alpha_2, \cdots, \alpha_n]^T$, $\Omega = K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j)$, $y = [y_1, y_2, \cdots, y_n]^T$, $I_n = [1, 1, \cdots, 1]$, satisfy $i, j = 1, 2, \cdots, n$.

The regression estimation function can be obtained by using the least squares method to obtain α and b :

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x_j) + b \tag{9}$$

It can be seen that when estimating the regression function, the calculation process is related to $K(x_i, x_j)$ and not to $\Phi(\cdot)$. According to the characteristics of the acquired data, this paper chooses $K(x_i, x_j)$ as the radial basis function as the kernel function, and its expression is:

$$K(x_i, x_j) = \exp\left(-\frac{\left\| x_i - x_j \right\|^2}{2\sigma^2}\right) \tag{10}$$

Where σ is the core width. Therefore, the entire LS-SVM model needs to determine two parameters C and σ. In order to achieve the overall optimal effect, it is necessary to optimize these two parameters.

2.5. Situational Research

The situation of safe production is changing all the time, so the situation of safety management must be advanced. Therefore, the time series analysis for data analysis is not applicable to the ARMA model, and the non-stationary sequence analysis is most suitable for selecting the ARIMA model.

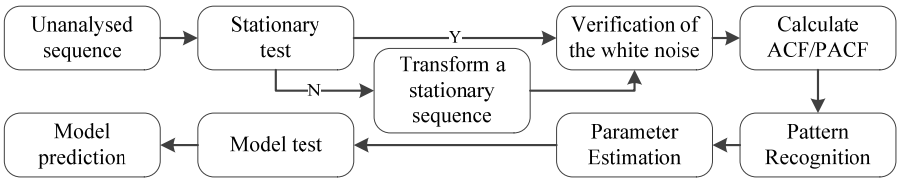

Figure 3. ARIMA algorithm model prediction flow chart

The time series of the overall data is decomposed into linear autocorrelation part and nonlinear part by wavelet decomposition, and reconstructed separately. The linear autocorrelation part adopts LS-SVM model of multiple input and multiple output for situation analysis, detail part The ARIMA model is used to analyze the situation, and finally the two parts of the situation analysis results are fitted to

obtain the final situation results, and the final results and historical data are used for analysis and judgment.

2.6. ARIMA-LS-SVM combined model modeling steps

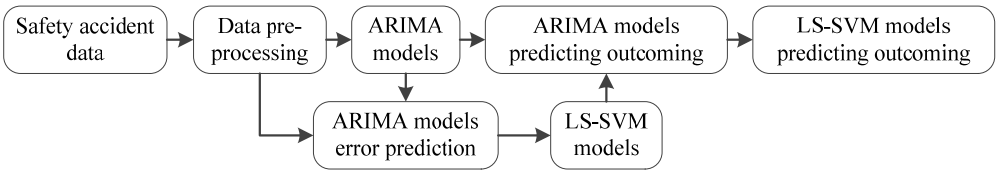

Figure 4. ARIMA-LS-SVM combination model flow chart

It is assumed that the time series $\{y_t\}$ can be regarded as consisting of a linear autocorrelation sequence $\{L_t\}$ and a nonlinear sequence $\{N_t\}$ that:

$$\{y_t\} = L_t + N_t \tag{11}$$

Sequences L_t and N_t are estimated from time series.

The first step: using ARIMA to establish a linear model for $\{y_t\}$, the predicted result is $\hat{L}_t$, which is the linear prediction part. The residual of the original sequence and $\hat{L}_t$ at time t is e_t, and the residual calculation expression is:

$$e_t = y_t - \hat{L}_t \tag{12}$$

The second step: the nonlinear relationship in the residual sequence $\{e_t\}$ implies $\{y_t\}$. That is, using the LS-SVM algorithm to establish a nonlinear model for the sequence $\{e_t\}$, the nonlinear information contained in $\{y_t\}$ can be mined. For the LS-SVM model of m input nodes, the residual calculation formula is:

$$e_t = f(e_{t-1}, e_{t-2}, e_{t-3}, \cdots, e_{t-m}) + \varepsilon_t \tag{13}$$

In there. ε_t is a random error, and the nonlinear function f can predict the approximation by the LS-SVM algorithm, and the prediction result is recorded as $\hat{N}_t$.

The third step: combine the ARIMA model with the LS-SVM model predictive value to form the final prediction result, that is, the third step: combine the ARIMA model with the LS-SVM model predictive value to form the final prediction result.

$$\hat{y}_t = \hat{L}_t + \hat{N}_t \tag{14}$$

According to the analysis of the prediction process, the ARIMA model is used to describe the linear part of the time series. The LS-SVM model mainly describes the nonlinear part. The combination of the two models can better combine their strengths, and the overall data analysis is stronger than the individual. Analysis results.

3. Experimental processing and analysis of results

3.1. Data preprocessing

In the course of the experiment, the experimental data obtained has more noise, and the dirty data and redundant data need to be processed accordingly, so as to avoid the influence and interference of the data on the experimental results, so the process of preprocessing the data is very necessary.

Table 1. Experimental data preprocessing method

Data preprocessing problem	Corresponding processing method
Missing value data	Always take the median (paragraphs with too many missing values are smoothed)
Text data	Be ignored
Archive data	Record as null or write to zero
Duplicate data	Remove duplicate data by rule
Dimension inconsistent data	Principal component analysis

3.2. Data reduction process

According to the experimental results, after the data reduction by the PCA method, the contribution rate of the new dimensionality reduction data can reach 95% or more, based on the experimental results, while retaining the authenticity and operability of the original data. Successfully reduce the complexity of the data dimension, greatly improve the efficiency of processing data through dimensionality reduction preprocessing, saving time to process useless data.

Table 2. Data dimensionality reduction experimental data

	Accuracy	Processing speed
Before data reduction	More than 99%	20.65s
After data reduction	More than 98%	9.65s

Analysis of the results: After data reduction, the accuracy rate is slightly reduced by less than 1%, and the processing speed is more than doubled. Because the amount of data is small, the eigenvalues of the data are not lost, so the processing speed and processing efficiency are improved.

3.3. Data Fusion

The main data of the safety supervision obtained after the key component is extracted by principal component analysis is as follows:

Table 3. Experimental data category table

Type of data	Related description
Administrative License	Administrative license acceptance information, administrative license approval information, and corporate license information
Report complaints	Reporting complaint information, reporting complaint acceptance information, reporting event information, reporting severity classification information, monitoring reporting complaint information reporting complaint information, reporting complaint acceptance information, reporting event information, reporting severity classification information, regulatory reporting complaint information
Hidden trouble investigation	Enterprise basic information, hidden danger type information, regional security risk information, hidden danger information in various industries
Business account	Enterprise account information
Occupational health	Personnel safety training information, dangerous goods information, risk management information, hazardous raw material product management information
Law enforcement inspection	Basic information of law enforcement inspection, basic information of administrative punishment, information of law enforcement plan, information of law enforcement personnel, list of law enforcement basis
Accident management system	Basic information on administrative punishment, accident report information, accident casualty information, accident responsibility unit information

Through data measurement, the accuracy of analysis after data fusion is higher than 98%. The accuracy of the combined analysis of the main components is less than 90%. The accuracy of each component is less than 80% without data dimensionality reduction. The analysis by data fusion is better than the analysis of the above seven principal components after separate analysis.

It can be seen from the verification results that after data fusion and data analysis, the accuracy of data analysis can be greatly improved. The analysis results are better than the separate analysis of each component and the principal component analysis without data fusion, because different components may appear between each other. Interference and impact, so component analysis without principal component analysis have the greatest error for the final safety rating analysis.

3.4. Comparison of experimental methods

Table 4. Comparison of experimental data

Experimental model selection	Accuracy	Processing speed
ARIMA Model	Less than 90%	6.25s
LS-SVM Model	Less than 90%	6.35s
ARIMA Model and LS-SVM Model	Above 99%	20.65s
Apply ARIMA Model and LS-SVM Model after data reduction	Above 98%	9.65s

Analysis of the results: After analyzing the data analysis by ARIMA and LS-SVM combined algorithm model, the accuracy is higher than the results of the analysis of big data by the two models, but the Apply ARIMA Model and LS-SVM Model after data reduction with the principal component analysis process before. In contrast, the processing speed and processing efficiency are not as good as the model of principal component analysis. Therefore, it is necessary to set the data dimensionality reduction process before the ARIMA and LS-SVM combination algorithm model.

4. Conclusion

Due to the large amount of data obtained by safety supervision, there are many factors that lead to safety production liability accidents. After research, it is found that fully exploiting the information contained in the time series can play an important role in the safety evaluation, and the data processing and other pre-processing processes can greatly improve the efficiency of data analysis, improve the accuracy and avoid data interference.

Safety supervision data can be regarded as consisting of linear and nonlinear data. Using ARIMA model to describe the linear law of time series, using LS-SVM to describe the nonlinear relationship of time series, can make a good combination of the two, gives full play to their respective Advantages to improve the efficiency of data analysis. Experiments show that the combination of ARIMA and LS-SVM algorithm has a simple structure and accurate prediction results, which is better than single ARIMA or LS-SVM algorithm. It corrects the error of single model, but it is affected by data noise. The prediction effect of the ARIMA and LS-SVM combined models is not as good as the fitting effect, and further research and discussion are needed in the future model prediction process.

Acknowledgments

This work was financially supported by Beijing Municipal Science and Technology Project (No. Z181100009018003).

References

[1] A survey of big data management: Taxonomy and state-of-the-art. Journal of Network & Computer Applications, 2016, 71: 151-166.
[2] Alam F, Mehmood R, Katib I, et al. Data Fusion and IoT for Smart Ubiquitous Environments: A Survey. IEEE Access, 2017, 5 (99): 9533-9554.
[3] Din S, Ahmad A, Paul A, et al. A Cluster-Based Data Fusion Technique to Analyze Big Data in Wireless Multi-Sensor System. IEEE Access, 2017, 5 (99): 5069-5083.
[4] Yu-Ling Z, Ren-Jie Z. Study on application of multi-senor data fusion technology on network security, International Conference on Electronics. IEEE, 2011.
[5] Ma K, Zhang H, Wang R, et al. Target tracking system for multi-sensor data fusion, Technology,

Networking, Electronic & Automation Control Conference. IEEE, 2018.
[6] Liu X, Wang H, Lai J, et al. Multiclass Support Vector Machines Theory and Its Data Fusion Application in Network Security Situation Awareness, International Conference on Wireless Communications. IEEE, 2007.
[7] Bocca E, Viazzo S, Longo F, et al. Developing data fusion systems devoted to security control in port facilities. 2005.
[8] Barry, Dickie, Mazel. Perimeter security at San Francisco International Airport: Leveraging independent, existing systems to form an integrated solution, IEEE Conference on Technologies for Homeland Security. IEEE, 2009.
[9] Rein K, Biermann J. Your high-level information is my low-level data - A new look at terminology for multi-level fusion, International Conference on Information Fusion. IEEE, 2013.
[10] Sutor S, Reda R. Multi sensor technologies augmenting video surveillance: Security and data fusion aspects, International Symposium on Computer & Information Sciences. IEEE, 2008.
[11] Mixia L, Qiuyu Z, Hong Z , et al. Network Security Situation Assessment Based on Data Fusion, International Workshop on Knowledge Discovery & Data Mining. IEEE, 2008.
[12] Giacobe N A. Data fusion in cyber security: first order entity extraction from common cyber data, Proceedings of SPIE - The International Society for Optical Engineering, 2012, 8408: 11.
[13] Wang H, Liu X, Lai J, et al. Network Security Situation Awareness Based on Heterogeneous Multi-sensor Data Fusion and Neural Network, International Multi-symposiums on Computer & Computational Sciences. IEEE Computer Society, 2007.
[14] Newman A R. Confidence, pedigree, and security classification for improved data fusion, International Conference on Information Fusion. IEEE, 2002.
[15] Zhenyu Lu, Jungho Im, Lindi Quackenbush, et al. Population estimation based on multi-sensor data fusion. International Journal of Remote Sensing, 2010, 31 (21): 5587-5604.