ISSN 2616-5775 Vol. 6, Issue 13: 29-34, DOI: 10.25236/AJCIS.2023.061305

Research on Tobacco Industry Sales Forecast Based on Cloud Computing and SSA-SVR

Jie Gao[1,a,*], Lu Zhang[1], Bo Lin[2], Xunbin Hu[1]

[1]*Xi'an Company of Shaanxi Tobacco Company, Xi'an, China*
[2]*Shaanxi Company of China Tobacco Company, Xi'an, China*
[a]*3116851297@qq.com*
**Corresponding author*

Abstract: *The accuracy of supply forecast and sales forecast of tobacco industry is directly related to the production and development of tobacco industry. Supply and sales forecast is the premise and basis of production and sales plan, which can improve the scientific nature of production and sales plan, so as to create more economic benefits for the tobacco industry and promote social and economic development. The application of cloud computing technology and SSA-SVR hybrid model can effectively solve the problem of tobacco industry and improve the sales prediction accuracy of tobacco industry. In this regard, this paper combines the literature data method, case analysis method, statistical methods, comparative experiment and other methods to deeply study the tobacco industry sales forecast based on cloud computing and SSA-SVR.*

Keywords: *Tobacco industry, Cloud computing technology, Singular spectrum analysis (SSA), Support vector regression (SVR) model, Drosophila optimization algorithm, SSA-SVR hybrid model, Sales forecast*

1. Introduction

The application of cloud computing technology, singular spectrum analysis (SSA) methods, and support vector regression (SVR) models can enable the rapid establishment of rational and mature sales forecasting systems [1]. Portes et al. pointed out that the SSA system is not only able to distinguish uncertain signals, but also to separate and synthesize signals[2], so that the time-series number sets are extracted with high accuracy, and then reconstructed into multi-column component number sets, including long-term trends, periodic signals, etc., and finally multi-domain data are predicted and the sequences are extended in dimensionality[3]. Krishnan et al. showed that the SVR incorporates the Gaussian radial basis kernel and polynomial kernel and other kernel functions on the SVM architecture to achieve nonlinear regression and solve various dimensional problems[4].

2. Building a Tobacco Sales Forecasting System based on Cloud Computing and SSA-SVR

The integration of the SSAmodel and the SVR model under FOA is mainly reflected in the following four aspects: "trajectory matrix building-sampling training integration", "singular value decomposition-regression fittingintegration", "grouping-predicted output integration", and "diagonal averaging-kernel function and parameter selection integration"[3]. The specific integrations are as follows:

2.1. "Trajectory Matrix Building - Sampling Training" Integration

Trajectory matrix [5] building: By setting the original tobacco sales time series as $T=(t_1,t_2,t_3,\ldots,t_n)$, and the spatial dimension of the sales time trajectory matrix as L (window length), the sales trajectory matrix is $Y_i=(t_i,t_{i+1},\ldots,t_{i+L-1}),\mathrm{w}herei=1,\ldots,K,K=N-L+1.$

The trajectory matrix[6] is expressed as follows:

ISSN 2616-5775 Vol. 6, Issue 13: 29-34, DOI: 10.25236/AJCIS.2023.061305

$$Y=(Y_1,Y_2,\ldots,Y_K)=\begin{pmatrix} t_1 & t_2 & \cdots & t_K \\ t_2 & t_3 & \cdots & t_{K+1} \\ \cdots & \cdots & \cdots & \cdots \\ t_L & t_{L+1} & \cdots & t_N \end{pmatrix} \tag{1}$$

In Equation (1), $2 \le L \le N$, $K = N - L + 1$, the original time series of tobacco sales can be embedded in the L dimension of the matrix "Y"[7].

Sampling training [8]: The original tobacco sales number set is set as $Z^{original}$, which can be mapped to some high dimensional space and transformed into a number set Z^{map}. And then linear regression is used to analyze the number set Z^{map}. The original tobacco sales training number set can be first set as $Xrain^{original}$, which is expressed as follows:

$$Xrain^{original}=\left\{\left\{Z_1^{original},P_1^{original}\right\},\left\{Z_2^{original},P_2^{original}\right\},\cdots\left\{Z_m^{original},P_m^{irginal}\right\}\right\} \tag{2}$$

In Equation (2):

$Z^{original}$ -- a set of independent variables (quantity of tobacco sold); $P^{original}$ -- a set of dependent variables (time of sale, market share, production and inventory, etc.); “m”--a sample data point in the training set $Xrain^{original}$.

2.2. Integration of "Singular Value Decomposition - Regression Fitting"

Singular value decomposition: decompose the singular value (SVD) of the matrix Y. Let the singular value of tobacco sales be $H=YY^T$, $H_{\text{eigenvalue}}=\left\{\lambda_1,\lambda_{,2},\ldots,\lambda_d;\ d\le L\right\}$and then the matrix after SVD processing is expressed as follows:

$$Z=Z_1+Z_2+\cdots Z_d \tag{3}$$

$$Z_i=\sqrt{\lambda_i}U_iV_i^T \tag{4}$$

In equations (3) and (4), $i=1,\cdots,d$; λ_i -- the eigenvalue of H; U_i --the left singular vector; V_i --the right singular vector; $\sqrt{\lambda_i}$ -- the matrix spectrum.

Regression fitting: The purpose is to derive the best value of vector b, set as "b^*", fit and regress to the training set $Xrain^{original}$, and then build the system, as follows:

$$\min\frac{1}{2}\|\omega\|+c\sum_{i=1}^{m}(\breve{\vartheta}_i+\hat{\vartheta}_i) \tag{5}$$

$$s.t.\begin{cases} -\varepsilon-\breve{\vartheta}_i\le P_i^{original}-\omega^T\cdot\varphi(Z_i^{original})-b\le\varepsilon+\hat{\vartheta}_i \\ \breve{\vartheta}_i\ge 0 \\ \hat{\vartheta}_i\ge 0 \end{cases} \quad (i=1,2,\cdots,m) \tag{6}$$

$\breve{\vartheta}_i$、$\hat{\vartheta}_i$ -- the soft boundary relaxation factor, which minimizes the fitting error.

Equation (6) can be transformed by the Lagrange function as follows:

$$\min_{\omega,b,\vartheta_i}\ \max_{a_i\ge 0,\mu_i\ge 0}\ L(\omega,b,\breve{a}_i,\hat{a}_i,\breve{\vartheta}_i,\hat{\vartheta}_i,\breve{\mu}_i,\hat{\mu}_i) \tag{7}$$

ISSN 2616-5775 Vol. 6, Issue 13: 29-34, DOI: 10.25236/AJCIS.2023.061305

$$\begin{aligned} &L(\omega,b,\breve{a}_i,\hat{a}_i,\breve{\vartheta}_i,\hat{\vartheta}_i,\breve{\mu}_i,\hat{\mu}_i) \\ &=\frac{1}{2}\|\omega\|+c\sum_{i=1}^{m}(\breve{\vartheta}_i+\hat{\vartheta}_i)+\sum_{i=1}^{m}\breve{a}_i(-\varepsilon-\breve{\vartheta}_i-P_i^{original}+\omega^T\cdot(Z_i^{original})+b) \\ &+\sum_{i=1}^{m}\hat{a}_i(P_i^{original}-\omega^T\cdot(Z_i^{original})-b-\varepsilon-\hat{\vartheta}_i)-\sum_{i=1}^{m}\breve{\mu}_i\breve{\vartheta}_i-\sum_{i=1}^{m}\hat{\mu}_i\hat{\vartheta}_i \end{aligned} \tag{8}$$

In Equations (7) and (8), $\breve{a}_i,\hat{a}_i,\breve{\mu}_i,\hat{\mu}_i$ are no less than 0, all of which are Lagrange coefficients.

The optimal values $\breve{a}_i value,\hat{a}_i value,i.e.,value\,\breve{a}_i^*\,and value\,\hat{a}_i^*$ are calculated according to the equation, and finally the optimal value of the vector b^* is derived.

2.3. "Grouping - Predicted Output" Integration

Grouping[9]: By grouping the matrix Y, it can be divided into a number of discrete subsets expressed as $K=\{K_1,\cdots,K_m\}$, where "m" denotes the number of groups. In this study, tobacco sales in different months are taken as the number of groups of discrete subsets, and the matrix Y after grouping can be expressed as follows:

$$Y=Y_{K_1}+Y_{K_2}+\cdots+Y_{K_m} \tag{9}$$

In Equation ⑨, $K_1,\cdots,K_m$ -- divided into m discrete subsets; "m" -- model parameter.

Predicted output: Let the feature data of the test set be "T^{test}", input the data into the SVR model that has completed training, and output the predicted value $Z^{predict}$. The predicted output of the system is expressed as follows:

$$Z^{predict}=\sum_{i-1}^{m}(\hat{a}_i^*-\breve{a}_i^*)\cdot k(T^{test},T_i^{original})+b^* \tag{10}$$

In Equation ⑩, $k(T^{test},T_i^{original})=\varphi(T^{test},T_i^{original}),i=1,2,\cdots,m$.

3. Prediction Process of Tobacco Industry Sales based on Cloud Computing and SSA-SVR

In this study, the SSA-SVR hybrid sales prediction system was analyzed online usingFOA based on the current sales situation of a tobacco company. The tobacco sales volume was predicted from the singular spectrum analysis model and the support vector regression model, respectively, and finally the most reasonable sales prediction result was synthesized.

3.1. Data Sources

In order to verify the effect of the SSA-SVR hybrid system for tobacco sales volume prediction, the annual data of cigarette sales volume of a tobacco company from January 2016 to December 2020 were selected as the experimental data information in this paper. The data revealed that the company had the highest sales volume of about 428,000 cases in 2020 and the lowest sales volume of about 417,200 cases in 2016, with an average annual sales volume of about 422,600 cases,

3.2. Error Evaluation Indexes

These include mean absolute percentage error, mean absolute error, mean square error and root mean square error [13], which are mainly used to calculate the accuracy of tobacco sales volume prediction in the SSA-SVR hybrid system, as follows:

Mean absolute percentage error: Abbreviated as "MAPE", its principle is to use a percentage (%) to represent the average value of the proportion of sales data points in the test set to the actual sales value of the test set. The lower the MAPE value, the smaller the error and the higher the accuracy of the model prediction. It is calculated by the equation below:

ISSN 2616-5775 Vol. 6, Issue 13: 29-34, DOI: 10.25236/AJCIS.2023.061305

$$MAPE = \frac{1}{Q}\sum_{t=1}^{Q}\left|\frac{y_t^{value} - y_t^{predict}}{y_t^{value}}\right| \tag{11}$$

In Equation (11), Q -- number of tobacco sales predictions; y_t^{value} -- actual tobacco sales values at time point t; $y_t^{predict}$ -- predicted values of tobacco sales for the model at time point in table 1.

Table 1: Evaluation criteria of prediction effect corresponding to different MAPE values

MAPE value (%)	Effect evaluation
MAPE≤10%	High accuracy of tobacco sales prediction
10%<MAPE≤20%	Good effect of tobacco sales prediction
20%<MAPE≤50%	Results of tobacco sales prediction within a reasonable range
MAPE>50%	Inaccurate tobacco sales prediction

Mean absolute error: Abbreviated as "MAE", it refers to the average difference between the data points of the sales volume sets in different months of the test tobacco, which reflects the actual sales volume; the smaller the MAE value, the more accurate the prediction. It is calculated by the equation below:

$$\mathrm{MAE} = \frac{\sum_{t=1}^{Q}\left|y_t^{value} - y_t^{predict}\right|}{Q} \tag{12}$$

Mean square error: Abbreviated as "MSE", it has a similar principle to MAE, but this index amplifies the impact of a single difference in the test set of the tobacco sales volume. For example, in January and February, it can be judged by the MSE that it has a greater impact on the sales volume of tobacco companies in a year, and the MSE can be used to evaluate the stability of the prediction model. It is calculated by the equation below:

$$\mathrm{MSE} = \frac{\sum_{t=1}^{Q}\left(y_t^{value} - y_t^{predict}\right)\text{^}2}{Q} \tag{13}$$

Root mean square error: Abbreviated as "RMSE", it is one of the methods for extracting a root based on the MSE index to assess the actual numerical size of tobacco sales. It is calculated by the equation below:

$$\mathrm{RMSE} = \left(\frac{\sum_{t}^{Q}\left(y_t^{value} - y_t^{predict}\right)^2}{Q}\right)\text{^}0.5 \tag{14}$$

4. Prediction Results and Comparative Analysis

4.1. Prediction Results of the SSA-SVR Hybrid Model

It is known that the tobacco sales volumes of this company in 2016, 2017, 2018, 2019 and 2020 were 417.2 thousand cases[10], 420.1 thousand cases, 420.5 thousand cases 427.2 thousand cases and 428.1 thousand cases, respectively. The prediction results of the SSA-SVR model are shown in Tables 2 and 3.

Table 2: The SSA-SVR mixed model prediction results

Year	2015	2016	2017	2018	2019
Prediction result	41.62(99.76%)	42.01(100%)	42(99.88%)	42.68(99.90%)	42.75(99.86%)

The prediction errors of the SSA-SVR model, including MAPE, MAE, MSE and RMSE, were calculated according to Equations ⑭, ⑮, ⑯, ⑰ and expressed as P_{MAPE}, P_{MAE}, P_{MSE} and P_{RMSE} pairs, respectively, as shown in Table 3.

ISSN 2616-5775 Vol. 6, Issue 13: 29-34, DOI: 10.25236/AJCIS.2023.061305

Table 3: Comparison of prediction error indexes between SSA-SVR model and other models

Index	P_{MAPE}	P_{MAE}	P_{MSE}	P_{RMSE}
Calculated result	23.12%	24.08%	41.65%	22.18%

4.2. Comparison of the Prediction Effects of the SSA-SVR Hybrid Model with Several Single Models

The hybrid SSA-SVR system was compared with the separate LSTM system[11], SVR system, PM system, and ARIMA system to measure the forecasting performance of the hybrid SSA-SVR tobacco sales prediction system in a comprehensive manner[12]. The results of the comparison are shown in Table 4.

Table 4: Prediction effects of SSA-SVR hybrid model with several single models

Year / Model	2015	2016	2017	2018	2019
LSTM model	40(95.88)	40.23(95.76%)	40.32(95.88%)	41.02(96.02%)	41.23(96.30%)
SVR model	40.58(97.23%)	40.18(95.64%)	40.35(95.95%)	40.23(94.17%)	40.52(94.65%)
PM model	40.05(95.99%)	39.68(94.45%)	39.85(94.77%)	40.58(94.99%)	40.48(94.55%)
ARIMA model	39.15(93.84%)	39.56(94.17%)	39.62(94.62%)	40.46(94.71%)	40.22(93.95%)
FOA-SVR model	40.11(96.14%)	40.05(95.33%)	40.84(97.12%)	41.12(96.25%)	41.36(96.61%)
SSA-SVR model	41.62(99.76%)	42.01(100%)	42(99.88%)	42.68(99.90%)	42.75(99.86%)

As can be seen in Table 4, the accuracy of the SSA-SVR model was significantly higher than all other models, and the accuracy of the SSA-SVR model ranged from 99.76% to 100% (with an error rate of 0%-0.24%), with an average accuracy of 99.88% and an average error of 0.12%. Thus, it is confirmed that the SSA-SVR model built in this paper possesses a more successful prediction mechanism and can increase the accuracy of tobacco sales prediction to more than 99.59%.[13]

5. Conclusion

Cloud computing is a product of the era of information network, which is a technical model integrating a variety of information technologies, network technologies, and digital technologies, and has been applied in the tobacco industry. Both SSA and SVR belong to big data analytics technology, which is evolved based on cloud computing. So a variety of cloud computing methods are applied in the process of building the SSA analysis system and SVR analysis system.

The SSA-SVR hybrid system has the advantages of simple operation, few parameters, easy and fast calculation and high accuracy. The application of the SSA-SVR model based on cloud computing in the sales prediction of the tobacco industry can effectively improve the accuracy of sales prediction in the tobacco industry. Moreover, it can also be applied to retail sales prediction, high-accuracy prediction, supply prediction of branded products and other fields.

References

[1] Wu Mingshan, Wang Bing, et al Research on cigarette sales volume combination prediction model [J]. Chinese Journal of tobacco, 2019, 25 (3): 84-91

[2] Zhao Xinbo. Research on Big Data and Cloud Computing Technology. Management & Technology of SME, 2021, 14(8), 186-187.

[3] He Xuefeng. Exploration of job costing method based on artificial intelligence, big data and cloud computing-- A case study of tobacco companies in China. Finance and Accounting Monthly, 2018, 17(2): 69-72.

[4] Wang Haifei, He Lili. Application of MQT based on Hadoop cloud computing in tobacco marketing decision analysis [J]. Industrial control computer, 2012,25 (12): 101-103

[5] Zhu Feng, Gao Lin. Research on cigarette market demand prediction based on combination model [J]. Cooperative economy and technology, 2017 (1): 62-64

[6] Krishnan, Polak. Short-term traffic prediction under normal and incident conditions using singular spectrum analysis and the k-nearest neighbour method[C]. // IET and ITS Conference on Road Transport Information and Control: Curran Associates, Inc., 2012:1-6.

[7] Chang Bingguo, Zang Hongying, Liao Chunlei, et al. Retail sales forecast based on selective integrated ARMA combination model [J]. Computer measurement and control, 2018,26 (5): 132-135

ISSN 2616-5775 Vol. 6, Issue 13: 29-34, DOI: 10.25236/AJCIS.2023.061305

[8] Wang Shihao, Zhang Xiaoni, Zhang Yun, et al. Integrated prediction of cigarette demand in Tongchuan City [J]. Chinese Journal of tobacco, 2019,25 (6): 105-109
[9] Portes, Leonardo, Aguirre,et al. Matrix formulation and singular-value decomposition algorithm for structured varimax rotation in multivariate singular spectrum analysis[J]. Physical review, E, 2016,93(5 Pt.A).
[10] Zhong K, Wang Y, Pei J, et al. Super efficiency SBM-DEA and neural network for performance evaluation[J]. Information Processing & Management, 2021, 58(6): 102728.
[11] Jan N, Gwak J, Pei J, et al. Analysis of networks and digital systems by using the novel technique based on complex fuzzy soft information[J]. IEEE Transactions on Consumer Electronics, 2022, 69(2): 183-193.
[12] Yu Z, Pei J, Zhu M, et al. Multi-attribute adaptive aggregation transformer for vehicle re-identification[J]. Information Processing & Management, 2022, 59(2): 102868.
[13] Li J, Li S, Cheng L, et al. BSAS: A Blockchain-Based Trustworthy and Privacy-Preserving Speed Advisory System[J]. IEEE Transactions on Vehicular Technology, 2022, 71(11): 11421-11430