Drugs Sale Forecasting Based on SVR Integrated Promotion Factors

Yang Liu, Xiao Yang*, Cuiling Zhu and Jie Meng
School of Management Science and Engineering, Shandong University of Finance and Economics, Jinan, China
Email: yangxiao@sdufe.edu.cn

Abstract. In order to improve the inventory management level of pharmaceutical chain enterprise, drug sales forecasting is needed. Using machine learning SVR model, we forecast the drugs sale amount with a relatively high accuracy. Suppose the sales amount is influenced by promotion strategy, we integrated the promotion factors into the SVR model. Experimental result of drug sales prediction on the real sales data of a large chain drug company S shows that the SVR algorithm integrated with promotion factors get the accuracy rate of 91% and the algorithm can greatly improve the drug sales prediction results compared with traditional time series model.

1. Introduction
According to data from the national bureau of statistics (NBS), in the third quarter of 2019, the original drugs of chemical drugs increased by 3.1% compared with the same period last year. Market order in addition, the Ministry of Commerce department documents issued by the national medicine circulation industry development planning (2016-2020), in order to better meet the demand of national in 2020, to drug retail enterprises with annual sales accounting for more than 40% of the total amount of the drug retail market overall goal, hospital pharmacy outsourcing development trend also makes the chain pharmacy sales increase. In addition, trends such as outsourcing of hospital pharmacies have led to increased sales at chain pharmacies. However, many drug chain sales organizations still request goods through their personal experience at present. With the sales volume of chain drugstores increasing significantly, the use of such manual experience leads to a large backlog of drugs, and some drugs cannot be sold even within the period of validity. resulting in a great waste of warehouses and their funds. If we can predict the sales amount of the drugs, we can arrange the stock more reasonably considering the sales amount.

Among many prediction models, SVR (Support Vector Regression) algorithm is favored by many researchers because of its advantages in algorithm effect. For example, Xu S J et al. mixed SVR and SARIMA to predict the statistical indicators of the aviation industry, making the capacity management and planning of the future aviation industry more economical and scientific [1]. Guo Q et al selected six main influencing factors, used SVR to predict the future power demand and improved the accuracy of the prediction through parameter optimization [2]. Hong W C et al. combined SVR with chaotic sequence and evolutionary algorithm for power load prediction, effectively improving the premature convergence of genetic algorithm [3]. Yuan jinming combined SVR with singular spectrum analysis and ARMA to predict the future state of stocks, and the prediction results showed that the effective noise extracted by singular spectrum analysis would improve the accuracy of the SVR prediction model of fusion singular spectrum analysis and ARMA in the prediction of stock indexes [4].
For drug chains, sales amount is affected not only by seasonal factors, but also by outbreaks of infectious diseases, the promotion strategy and even by policies. In this paper, SVR algorithm was used to predict the sales volume of drugs based on S company's sales data for many years. The promotion factors of S company are taken into account when making the forecast.

2. Drug sales Forecasting Models

2.1. SARIMA

SARIMA model refers to the model established by converting non-stationary time series into stationary time series, and taking seasonal factors into consideration, regression of the dependent variable only to the value of its hysteresis judgment and the present value of the random error term. SARIMA (p, d, q) is called differential autoregressive moving average model. AR is autoregressive and p is autoregressive. MA is the moving average, q is the number of moving average terms, and d is the number of differences made when the time series becomes stationary. The basic idea of SARIMA model is to treat the data sequence formed by the predicted object over time as a random sequence and use a certain mathematical model to approximate the sequence. Once the model is identified, it can predict future values from past and present values of the time series.

2.2. SVR (Support Vector Regression) Model

SVR (Support Vector Regression) is an important branch of support vector machine (SVM). The standard SVM algorithm is also known as Support Vector Classification (SVC). The hyperplane decision boundary in SVC is the regression model of SVR:

\[ f(X) = w^T X + b \]  

(1)

If the sample points are close enough to the regression model, that is, they fall into the interval boundary of the regression model, then the loss of the sample is not calculated and the corresponding loss function is called - insensitive loss function: 

\[ L(z) = \max(0, |z| - \varepsilon) \], 

where \( \varepsilon \) is the super parameter that determines the width of the interval boundary. It can be seen that the insensitive loss function is similar to the hinge loss function used by SVC, and the value of the part near the origin is fixed to 0.

3. The Data Preprocessing and Feature Selection

We use the real drug sales data from S enterprise as experimental data. the enterprise take 8\textsuperscript{th}, 18\textsuperscript{th}, 28\textsuperscript{th} of each month as the promotion day, and 11 May as the annual member promotional day.

3.1. Data Description and Preprocessing

We take the data from January 2017 to December 2019 as the train and test data. Sales data include the number of drugs purchased by consumers, the place of origin of the drugs, whether the drugs are prescription drugs, whether the drugs are covered by medical insurance, whether the drugs contain specific drugs and other characteristics. After deleting duplicate data, filling missing data, revising the invalid data in the original data and unifying the types and prices, the data are normalized within the interval of 0 ~ 1. The data from October 1 to December 31, 2019 are used as the test set, and the remaining data are used as the training set.

3.2. Feature Selection

The extracted features should be able to fully describe the effective sales information of the drug. We selected six features as the features of the model, as shown in figure.1. That are prescription and non-prescription drugs, health care medicine and health care medicine, with specific drugs and drugs do not contain specific sales on average, consumers are for drug prescription attention compared with whether to lower health care medicine and whether they contain specific, consumers to buy more for health care drugs also contain specific drugs.
4. Drug Sales Forecast

4.1. Total Sales Analysis
As shown in figure 2, the sales volume in the first half of 2019 has a cyclical change and a slow rising trend. The peak point in the figure is the holiday and promotion day. There will hold the promotion activities in the 8th, 18th and 28th of each month, and promotion on May 18 each year reaching the annual sales highest point for each year. The annual sales slump lowest is at the beginning of the Spring Festival period.

Figure 2. Total Sales from 2017 to 2019.

Figure 3 and figure 4 show the sales preference of consumers in this enterprise. Figure 3 shows that the number of consumers to purchase prescription and over-the-counter drugs were similar, whether for drugs for prescription drug properties awareness is not high. Figure 4 and figure 5 show that health care drugs and drugs containing no specific purchases is far greater than the health care drugs and drugs containing specific. The fluctuation of no health care sales and the average sales of specific was close to zero. People will buy bulk drugs in the promotion day. Thus, it can be seen that the extracted data can clearly describe the effective the consumer's purchase demand.
4.2. Prediction Comparison of Models

The basic idea of SVR is to minimize the structural risk in the case of limited sample data, seeking a compromise between the accuracy of the given data approximation and the complexity of the approximation function to establish a model. With nearly tens of thousands of processed data sets, the trained model was used for prediction and compared with the results of single models such as exponential smoothing method, SARIMA, SVR and etc. The results were obtained as shown in figure 6. The sales prediction result of each model is shown in figure 6. We can see that the exponential smoothing method is a linear model based on time series. Therefore, it is not enough to take into account the influencing factors in the model calculation, resulting in a low confidence of the prediction results. The prediction at each peak point, i.e. on promotion day, is greatly different from the real value, so the fitting is poor.
Although SARIMA is also based on time series, the fitting effect of its prediction curve is slightly better than that of the exponential smoothing method considering seasonal factors, while the SVR takes into account various characteristics affecting sales into the model, and the confidence degree is better than other single models with full consideration. Compared with the results predicted by exponential smoothing method, SARIMA and SVR, it has the best fitting effect with the real value.

4.3. The Error Analysis

The accuracy of the prediction results of each model is mainly evaluated based on the accuracy of the prediction results. We use commonly used measurement: MAE (mean absolute error), RMSE (mean square root error) and MAPE (mean absolute percentage error) as our measurement [5-6]. The smaller the measurement is, the better model fitting and the higher the prediction accuracy is. The calculation formula is shown in equation (2), (3) and (4), where n is the number of samples, $x_i$ is the real value of samples, and $y_i$ is the predicted value of samples.

Average absolute error:

$$MAE(x, y) = \frac{1}{n} \sum_{i=1}^{n} |x_i - y_i|$$  \hspace{1cm} (2)

Square root of mean square error:

$$RMSE(x, y) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - y_i)^2}$$  \hspace{1cm} (3)

Average absolute percentage error:

$$MAPE(x, y) = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{x_i - y_i}{x_i} \right|$$  \hspace{1cm} (4)
Table 1. Error Analysis.

| Method          | Exponential Smoothing | SARIMA     | SVR        | P+SVR      |
|-----------------|-----------------------|------------|------------|------------|
| MAE             | 34242.05              | 14608.75   | 6045.33    | 4062.39    |
| RMSE            | 51946.71              | 22803.29   | 6827.28    | 4835.35    |
| MAPE            | 0.2576                | 0.1034     | 0.0615     | 0.0315     |

As can be seen from table 1, the average absolute error of the exponential smoothing method, SARIMA, SVR and P+SVR is 34242.05, 14608.75, 6045.33 and 4062.39, and the mean square root error is 51946.71, 22803.29, 6827.28 and 4835.35, and the average absolute percentage error is 0.2576, 0.1034, 0.0615 and 0.0315, compared with the average absolute percentage error of the single model. The average absolute percentage error was improved by 7%, and the P+SVR absolute percentage error was improved by 3% compared with the SVR without promotion factors.

5. Conclusion
In order to help the drug chain enterprises to make more accurate sales forecast, as well as to achieve the long-term goal of reducing costs, accurate procurement and zero inventory. This article USES the fusion promotion factors of P + SVR algorithm, with S enterprise's sales data as the research object, through the analysis of original data processing, the extraction of feature modeling after normalization, and exponential smoothing, SARAMA and SVR prediction comparison, find fusion promotion factors of P + SVR algorithm than any other single model to predict the results of the highest accuracy increased by 7%, thus can see fusion promotion factors of P + SVR algorithm in chain type pharmaceutical enterprise sales forecast more accurate prediction results can be obtained. This research can also be used to predict the sales volume of a single category of drugs, which is helpful for enterprises to better carry out rational procurement, inventory management and market development, and has important practical significance for the development of enterprises.

Acknowledgements
This paper is supported by the Key R & D project of Shandong Province(2019GSF108222); A Project of Shandong Province Higher Educational Science and Technology Program (Project No. J15LN56)

References
[1] Xu S, Chan H K and Zhang T 2019 Forecasting the demand of the aviation industry using hybrid time series SARIMA-SVR approach Transportation Research Part E Logistics & Transportation Review 122 169-180
[2] Guo Q, Feng Y and Sun X 2017 Power demand forecasting and application based on SVR Procedia Computer Science 122 269-275
[3] Hong W C, Dong Y and Zhang W Y 2013 Cyclic electric load forecasting by seasonal SVR with chaotic genetic algorithm International Journal of Electrical Power & Energy Systems 44 (1) 604-614
[4] Yuan J M 2019 Application of ARMA-SVR model based on singular spectrum analysis in stock index prediction Journal of Chongqing Institute of Technology (Natural Sciece) 23 (2)
[5] Huo R G 2016 Establishment and evaluation of a predictive model for the incidence of HFMD in liuzhou city Chinese Journal of Women and Children Health
[6] Ramos P, Santos N and Rebelo R 2015 Performance of state space and ARIMA models for consumer retail sales forecasting Robotics & Computer Integrated Manufacturing 34 151-163