8th International Seminar of Research Month 2023
Volume 2024
http://dx.doi.org/10.11594/nstp.2024.4105

Conference Paper

Intermittent Data Forecasting using Kernel Support Vector Regression

Amri Muhaimin[1]*, Endah Setyowati[2], Kartika Maulida H[1], Allan Ruhui Fatma Sari[1]

[1]Department of Data Science, Faculty of Computer Science, Universitas Pembangunan Nasional "Veteran" Jawa Timur, Surabaya 60294, Indonesia
[2]Faculty of Economy, Institut Agama Islam Negeri Ponorogo, Indonesia

Corresponding author:
E-mail:
amri.muhaimin.stat@upnjatim.ac.id

ABSTRACT

Forecasting involves making future estimates. Forecasting methods are commonly employed to predict stock prices, monetary distribution, and weather conditions. To generate accurate forecasts, it is crucial that the data used is consistent, comprehensive, and unchanging. Some data can be readily predicted, while some poses a considerable challenge. An illustration of this is found in discontinuous data, which is notably hard to forecast. Discontinuous data is marked by frequent instances of zero values due to sporadic events. For instance, when tracking the sales of aircraft or other products, sales do not transpire daily, causing recorded data to often register as zero. Various techniques have been explored to handle this kind of data. In this particular study, the chosen method is support vector regression. This method is capable of predicting discontinuous data with a quality level of 1.004, which is lower than traditional approaches like exponential smoothing.

Keywords: Discontinuous, forecasting, neural networks, support vector regression, time series

Introduction

Time Series Analysis is one way that can be used to predict the future. Different types of data can be predicted using different approaches. Several types of data are difficult to predict, especially if the data contains a lot of zero values. Data that has these characteristics is sales data. Such as selling motorbikes, cars, and the like. Sales of these goods do not always occur every day, so when no sales occur in a certain time period, the data is recorded with a value of zero (Croston, 1970). Time Series methods such as Autoregressive Integrated Moving Average (ARIMA) are not able to predict data with these characteristics because there are many zero values that cause the ARIMA model to produce negative forecast values. Such data is called intermittent data, which means intermittent. Intermittent data itself has 4 types, namely intermittent, lumpy, erratic, and smooth. Lumpy and intermittent types contain more zero values compared to erratic and smooth. Therefore, this forecasting data is quite difficult to predict and a solution is needed to overcome this.

This research aims to overcome this problem. By using a mixed combination of Neural Network and Time Series Regression, it can overcome the large number of zero values and produce positive forecasts. The model evaluation also uses the Root Mean Square Scaled Error (RMSSE) score (Kourentzes, 2013). This value is very robust if applied to data with a majority of zeros. If you use Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE), it will produce an undefined value because the formula for these two scores is not robust to many zero values.

Research on discontinuous data forecasting has been widely carried out. The use of methods also varies greatly. As was done by (Kourentzes, 2013) who only used the Neural Network method. The

How to cite:
Amri Muhaimin, A., Setyowati, E., Maulida H. K., & Sari, A. R. F. (2024). Intermittent data forecasting using kernel support vector regression. *8th International Seminar of Research Month 2023*. NST Proceedings. pages 23-26. doi: 10.11594/ nstp.2024.4105

results are not good enough and sometimes tend to stagnate. Because the data has too many zero values, the forecasting results do not seem to move and do not change. Furthermore (Rožanec et al., 2022) uses a Two-Fold approach. This approach aims to increase the goodness of the resulting model. Combined with the Exponential Smoothing method they succeeded in getting the best results but not robust for all types of intermittent data. Then (Muhaimin et al., 2021) used a Deep Learning approach directly with the Recurrent Neural Network model. This approach is quite expensive for modeling intermittent data. The results obtained are not significantly different from the exponential smoothing method or the Croston method. So, it is not efficient enough to be used on intermittent data. Next, the Neural Network model is used using modified error criteria (Liu, 2020). The error criterion is named sMDL, this error criterion is able to optimize the NN method calculation process. But sometimes there are still problems, such as the model being underfitting or overfitting.

The use of mixed methods between Time Series Regression and Neural Networks can provide significant results. By using an activation function that does not produce negative values, Neural Networks can provide maximum results in the model. This mixed concept relies on the Time Series Regression method as the main model. In the regression model, there is an error or residual value which is then modeled again using a Neural Network model. After that, the error forecasting results are returned to the regression model to produce forecast values. The forecast results are evaluated with the RMSSE value which is very suitable for sparse data. Parameter selection in Neural network models uses optimization methods such as grid search and the like. In this way, the forecast results are very robust to even sparse data.

The data used in this research is sales data at a supermarket. The time span in this data is five years, recorded daily as 1941 days. This data is data for the M5 competition on the Kaggle website. Researchers only chose one data used in this research. The variables used in the research are the running average value of the data and the lag of the data.

Details are explained in the following sections: Section 2 Research Methods, Section 3 Results and Discussion, and Section 4 conclusions. Section 2 discusses the methods used in this research, section 3 discusses the results of the proposed method, and section 4 discusses the findings during the research and conclusions.

Material and Methods

Support vector regression

Support Vector Regression (SVR) is a variant of Support Vector Machine (SVM) which performs regression by predicting continuous or numeric values in the data. SVR is basically a point closer to a hyperplane generated in an n-dimensional feature space that clearly separates data points about that hyperplane (Parbat & Chakraborty, 2020) [A Python-based support vector regression model for prediction of COVID-19 cases in India]. The goal of the SVR algorithm is to find a hyperplane that minimizes the overall deviation value by ensuring that most of the training data is within the deviation limit or tube ϵ. SVR uses a portion of the given data set to create a function estimator as follows:

$$f(x) = w^T \varphi(x) + b \qquad (1)$$

Where $\varphi(x)$ is a nonlinear transformation function that transfers the nonlinear properties into linear problems through high-dimensional space. At the same time, w and b are the weight and constant coefficients, respectively, which are estimated by minimizing the arranged risk function. Therefore, we should minimize error from the function, that is the distance between the predicted and the desired outputs. The equation is:

$$\frac{1}{2} \|w\|^2 + C \sum_{i=1}^{N} \square\, \xi_i + \xi_i^* \qquad (2)$$

subject to $\{y_i - (, x_i) - b \ \leq \ \varepsilon + \xi_i \ (, x_i) + b - \ y_i \ \leq \ \varepsilon \ + \ \xi_i^* \ \xi_i, \xi_i^* \ \geq 0 \ \}$ (3)

Where C refers to a penalty factor that is greater than zero, which reflects the degree of attention paid to spatial outliers. Generally, the more massive C is, the more attention will be paid to outliers.

Model Evaluation

Evaluation of the model in this research uses the calculation of the Root Mean Square Scaled Error (RMSSE) value. The RMSSE value is calculated using the following formula, where n is the length of the training data and h is the length of the testing data (Hyndman & Koehler, 2006). Apart from using RMSSE, it also uses mean squared error (MSE) and root mean squared error (RMSE).

$$RMSSE = \sqrt{\frac{\frac{1}{h}\sum_{t=n+1}^{n+h} (Y_t - \widehat{Y_t})^2}{\frac{1}{n-1}\sum_{t=2}^{n} (Y_t - Y_{t-1})^2}} \qquad (4)$$

Results and Discussion

Empirical result

With some experiments that we do in the dataset. We compare our proposed model using a conventional method and a hybrid method. The methods that we used as a benchmark are exponential smoothing and Hybrid Neural Network with Time Series Regression. For the input variable for the data, we use the seven-day lag data. After we lagged the data, model training was applied. The parameters that are used are C and gamma. Also, we use the Radial Basis Function (RBF) as the kernel for the model. The best parameters are obtained through the tuning parameter procedure. We use the grid search for this. Here is the result of the predicted value against the truth value.

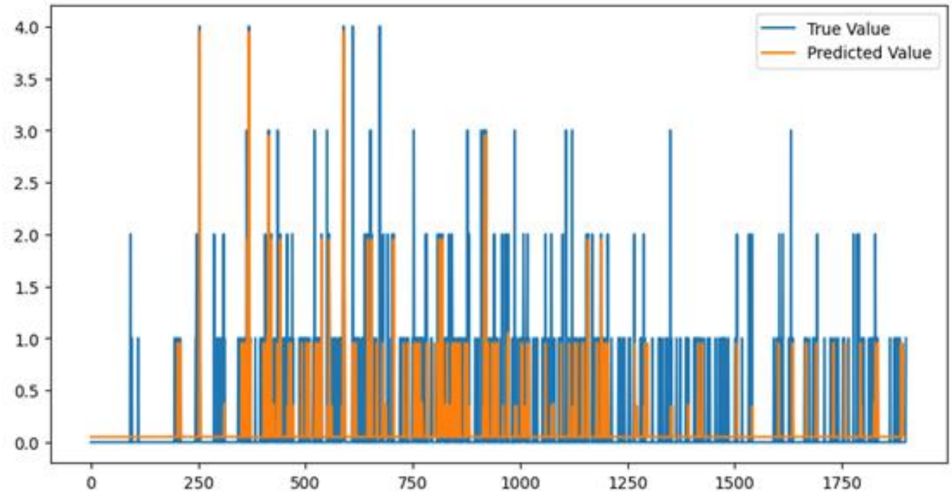

Figure 1. Model result

As the result from Figure 1 and Table 1, we can conclude that our proposed model can predict the sales well. There are no significant differences between the true value and the predicted value. Also for evaluation criterion, our model outperformed the conventional method which is exponential smoothing. But not the hybrid method.

Table 1. Comparison result

Model	Data	MSE	RMSE	RMSSE
Exponential Smoothing	Training	0.370	0.608	-
	Testing	1.015	1.000	1.280
Hybrid NN-TSR	Training	0.354	0.594	-
	Testing	0.723	0.850	0.986
K-SVR (our)	Training	0.267	0.516	-
	Testing	0.750	0.866	1.004

Conclusion

Support vector regression is a semi-parametric machine learning that can do both, classification and regression. In this case, that method is used to forecast the sales data. The data is unusual because it contains a lot of zeros. It is called intermittent demand data. Using the RBF as the kernel function, we produce the result that overcomes the exponential smoothing but is not hybrid. For future works, we will combine the Kernel-SVR method with some parametric methods to produce better forecasting and metric criteria.

Acknowledgment

This work was financially supported by LPPM UPN "Veteran" Jawa Timur, Surabaya.

References

Chaerudin, Dinar, S. A., & Fadillah, S. (2009). *Strategi pencegahan dan ppenegakan hukum tindak pidana korupsi.* Bandung: Refika Aditama.

Croston, J. D. (1970). Forecasting and stock control for intermittent demands. *Operational Research Quarterly, 23*(3), 289-303. https://doi.org/10.2307/3007885

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*(4), 679-688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Kourentzes, N. (2013). Intermittent demand forecasts with neural networks. *International Journal of Production Economics, 143*(1), 198-206. https://doi.org/10.1016/j.ijpe.2013.01.009

Liu, P. (2020). Intermittent demand forecasting for medical consumables with short life cycle using a dynamic neural network during the COVID-19 epidemic. *Health Informatics Journal, 26*(4), 3106-3122. https://doi.org/10.1177/1460458220954730

Muhaimin, A., Prastyo, D. D., & Lu, H. H. (2021). Forecasting with recurrent neural network in intermittent demand data. *Proceedings of the Confluence 2021: 11th International Conference on Cloud Computing, Data Science and Engineering*, 802-209.

Parbat, D., & Chakraborty, M. (2020). A python-based support vector regression model for prediction of COVID-19 cases in India. *ELSEVIER, 138*. https://doi.org/10.1016/j.chaos.2020.109942

Rožanec, J. M., Fortuna, B., & Mladenić, D. (2022). Reframing demand forecasting: A two-fold approach for lumpy and intermittent demand. *Sustainability (Switzerland), 14*(15). https://doi.org/10.3390/su14159295