Review of the evolution of statistical models and deep learning models: Based on the perspective of ARIMA and SVM

Yimeng Wang

University of Minnesota, Twin Cities, Minneapolis, 55455, United state

wang9739@umn.edu

Abstract. The importance of the evolution of statistical models and deep learning models, such as autoregressive integrated moving average model (ARIMA) and support vector machine (SVM), has drawn much attention from scholars. This paper investigates the evolution of statistical models and deep learning models based on the perspective of ARIMA and SVM. Then, this paper finds that there are some limitations of the literature on ARIMA and SVM, including less attention to causal inference, the scarce application of the future option, and the deep exploration of parameter optimization.

Keywords: ARIMA; SVM; causal inference; parameter optimization.

1. Introduction

Generally, we can draw conclusions through data, but more importantly, we can understand the distribution of variables and the relationship between variables through data. Using visualization methods such as charts and graphs to summarize the distribution and relationships between variables is a good idea. To visualize the data, a module can be used to make the data more three-dimensional and better observed. However, finding a more suitable model is a critical problem. In this paper, I will discuss the systematic combination and review of statistical models and deep learning models. At the same time, I will focus on combining and studying classical ARIMA, CARCH, LSTM, and SVR models.

ARIMA should pay more attention when a prediction time series includes a trend or seasonal figure. To date, as for predicting commodity prices, many scholars have widely adopted the ARIMA model. For instance, Areepong proposed methods to forecast the product and farm price of 5 vegetables and fruits (Rangsan Nochai and Titida Nochai, 2006). Backpropagation - Long and Short Term Memory (BP-LSTM) algorithm has been found to be suitable for the travel time prediction and energy consumption prediction scenarios to realize intelligent navigation and energy saving of new energy vehicles (Chen Mingnuo and Zhao Wanzhong, 2019). Through this article, we know that the LSTM model has unique advantages in addressing the long-term forecasting problem, and it is used to discuss the problem of long-term analysis.

The remainder of this paper concludes the section literature review, including the review of statistical and deep learning models. Then, this paper concludes the section on conclusions and insights.

2. Literature Review

2.1 Review of statistical models: Based on ARIMA

ARIMA models are a classical algorithm of time series prediction and analysis. The models are characterized by three main components: AR and MA. Several different types of statistical models are based on ARIMA. Some of the most commonly used ones include the following:

(1) ARMA (Autoregressive Moving Average) model:

This is a simpler version of ARIMA that only includes AR and MA components. It is employed in the prediction of time series that exhibit prominent stationary behavior.

Few studies have explored the parameter setting of the ARIMA model, including the distance measure (Piccolo, 1990; Kalpakis et al., 2001). There have been few studies exploring the parameter setting of the ARIMA model. Piccolo (1990) proposed a parametric approach. Kalpakis et al. (2001) considered the problem of clustering ARIMA using the Euclidean distance. Then, we mainly introduce the evolution of the VARIMA literature stream.

(2) VARIMA model:

ARIMA model is a multivariate time series model that extends ARIMA to multiple time series. It is used when multiple time series are related to each other and need to be analyzed together. The VARIMA model is used when multiple time series are believed to be interdependent and influence each other. Instead of modeling each time series separately, a VARIMA model models the relationships between the time series by including lagged values of each variable as predictors of the other variables. The VARIMA model is useful when the data are believed to have multiple interrelated components, such as macroeconomic indicators. It can be used to make forecasts for each of the variables in the model and to understand how changes in one variable may affect the others.

(3) GARCH (Generalized Autoregressive Conditional Heteroskedasticity) model:

This model is used to analyze time series that exhibit volatility clustering, where periods of high volatility are followed by periods of low volatility. It is used in finance and economics to model stock prices, exchange rates, and other financial variables.

These are just a few examples of the different types of statistical models that are based on ARIMA. Many other variations and extensions of the basic ARIMA model are used in different fields and applications.

2.2 Review of deep learning model: Based on SVR

Several statistical models based on SVR have been proposed in the literature. Here are some examples:

(1) Least-Squares Support Vector Regression (LS-SVR):

This extension of the standard SVR algorithm uses a least-squares approach to estimate the model parameters. LS-SVR has been shown to be computationally efficient and can handle large datasets.LS-SVR has several advantages over the standard SVR algorithm. LS-SVR can be used to estimate the probability of a predicted output, which can be useful in specific applications.

One of the limitations of LS-SVR is that it may perform poorly in noisy data or the number of training examples is small. In such cases, it may be necessary to use other variants of SVR or other regression algorithms. Therefore, we should filter the noise before using LS-SVR. For example, based on the comparative perspective, some scholars have pointed out that the prediction performance of LS-SVR is better than PLS and other conventional models. Thus, LS-SVR is an excellent model that could successfully apply in various domains, such as finance, engineering, and bioinformatics. It has been used for time series prediction, function approximation, and classification tasks. LS-SVR is a powerful algorithm for regression problems and can be an excellent alternative to traditional statistical models.

(2) Twin Support Vector Regression (TSVR):

TSVR is a modification of the standard SVR algorithm that uses two parallel hyperplanes to approximate the data. TSVR aims to improve the accuracy of the regression model by reducing the number of support vectors needed. The hyperplanes of TSVR are constructed to minimize the distance between the data points and the hyperplanes. As for the parameter optimization, some scholars find that the prediction performance of TSVR can be improved by the wavelet transform method. However, TSVR may also be computationally expensive, especially for large datasets, and may require careful tuning of its parameters to achieve optimal performance (Li, Lei, Wang, Zhang, and Wang,2020).

(3) Fuzzy Support Vector Regression (FSVR):

FSVR is a variant of SVR that incorporates fuzzy logic to deal with uncertain and imprecise data. FSVR has been applied successfully in various domains, such as traffic flow prediction, stock price prediction, and medical diagnosis.

The key idea behind FSVR is to use fuzzy membership functions to represent the uncertainty in the input data. The membership functions assign each input data point a degree of membership to each of the fuzzy sets. These fuzzy sets are then used to construct a set of support vectors that are used to define the regression function.

FSVR works by finding a hyperplane that maximizes the margin between the predicted output values and the actual output values. This hyperplane is defined by the set of support vectors that are closest to the hyperplane. The fuzzy membership functions determine the margin, which helps to handle the uncertainty in the data. For example, the FSVR has been successfully adopted by some authors for the prediction of business cycles under uncertain circumstances.

(4) SMulti-Task Twin Support Vector Regression (MTTSVR):

MTTSVR is an extension of TSVR that is designed to handle multiple regression tasks simultaneously. MTTSVR has been used in various applications, such as protein-ligand binding affinity prediction and drug design.

MT-TSVR works by jointly solving multiple regression problems using a shared feature space. The algorithm tries to find multiple hyperplanes that can predict the output of each task while also maintaining the structural similarity between tasks. By sharing information between tasks, MT-TSVR can achieve better generalization performance and reduce overfitting, especially when limited samples are available for each task.

MT-TSVR may also require careful tuning of its parameters and can be computationally expensive, especially when there are many tasks and large datasets.

3. Conclusion and Insight

Different statistical models are used in many other areas depending on their strengths. I discuss two statistical models, ARIMA and SVR, in-depth. The choice of model depends on the specific data and task at hand, and it is essential to carefully evaluate and compare different models before making a final decision. Statistical models such as ARIMA are beneficial for modeling time series data with clear trends or seasonal patterns. In contrast, machine learning models such as SVR can handle linear and nonlinear relationships between variables.

By observing many other papers, I found that, first, as for the research objects, the current research objects are primarily focused on the prices of traditional financial products, such as stocks and commodities. There needs to be more research on price predictions, such as financial derivatives. The price composition of financial results and its influencing factors are relatively complex, so in-depth exploration is needed. Second, the research methods need to be more profound. Algorithms are used directly in fields such as finance, and too few forecasts are made through graphics and model creation, while much focus is on optimizing algorithm parameters. I need to pay more attention to the model's validity, help understand the model in depth, and know when to use it but only calculate it. The third type of research question focuses more on prediction and classification problems, with no deep causal inference research. This is because it is more difficult for causal inference to consider all the current macro factors. At the same time, the effectiveness of the current macro-financial policy is seriously threatened. Many people need to learn how to make better predictions. More importantly, the cause-and-effect problem must be found first; many people need help. Finally, the exploration of factor validity needs to be more profound. More factors can be found in the future to enhance the robustness of predictions. The results can be more accurate only when all aspects are considered in any problem. At the same time, the emergence of more factors can change the previous cognition, and the result may also be altered.

References

[1] Livieris, Ioannis E., Emmanuel Pintelas, and Panagiotis Pintelas. "A CNN–LSTM model for gold price time-series forecasting." Neural computing and applications 32 (2020): 17351-17360.

[2] Piccolo D. A distance measure for classifying ARIMA models[J]. Journal of time series analysis, 1990, 11(2): 153-164.

[3] Kalpakis K, Gada D, Puttagunta V. Distance measures for effective clustering of ARIMA time-series[C]//Proceedings 2001 IEEE international conference on data mining. IEEE, 2001: 273-280.

[4] Lindberger, N. A. (1973). Stochastic identification of computer-regulated linear plants in noisy environments. International Journal of Control, 17(1), 65-80.

[5] Martinez, E. Z., Silva, E. A. S. D., & Fabbro, A. L. D. (2011). A SARIMA forecasting model to predict the number of cases of dengue in Campinas, State of São Paulo, Brazil. Revista da Sociedade Brasileira de Medicina Tropical, 44, 436-440.

[6] Chadsuthi, S., Modchang, C., Lenbury, Y., Iamsirithaworn, S., & Triampo, W. (2012). Modeling seasonal leptospirosis transmission and its association with rainfall and temperature in Thailand using time–series and ARIMAX analyses. Asian Pacific journal of tropical medicine, 5(7), 539-546.

[7] Li, Y., Shao, X., & Cai, W. (2007). A consensus least squares support vector regression (LS-SVR) for analysis of near-infrared spectra of plant samples. Talanta, 72(1), 217-222.

[8] Li, S., Lei, W., Zhang, W., Wang, X., & Wang, L. (2020). Weighted TSVR based nonlinear channel frequency response estimation for MIMO-OFDM system. IEEE Access, 8, 224283-224291.

[9] Lin, K. P., & Pai, P. F. (2010). A fuzzy support vector regression model for business cycle predictions. Expert Systems with Applications, 37(7), 5430-5435.

[10] Yong Yu, Xiaosheng Si, Changhua Hu, Jianxun Zhang; A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. *Neural Comput* 2019; 31 (7): 1235–1270. doi: https://doi.org/10.1162/neco_a_01199