




Quantile Regression Support Vector Machine (QRSVM) Model for Time Series Data Analysis

Dharmendra Patel^(✉) 

Smt. Chandaben Mohanbhai Patel Institute of Computer Applications,
CHARUSAT, Changa, Gujarat, India
dharmendrapatel.mca@charusat.ac.in

Abstract. Analysis of time series information is very interesting as it can be used to understand the past and to forecast the future. Mainly, the data models of the time series are based on the normal least square regression (LSR). For handle the outliers, the least square regression is not efficient. Data from the time series contains outliers in a notable quantity that may affect the results of the prediction. The proposed solution will use statistical techniques of quantile regression that robustly gives insights based on different dimensions as well as treats outliers. The advantage of quantile regression is to discover more useful predictive relationships in situations where there is a poor relationship between independent variables. The paper described the statistics of QRSVM model. The paper dealt experiments based on time series data and proved that QRSVM model is superior than LSR model in insights generations and for outlier handling.

Keywords: Least square regression · Quantile regression · Support vector machine · Time series analysis · Quantile regression support vector machine model

1 Introduction

Time series data is very vital for many applications such as economics, medicine, education, social sciences, epidemiology, weather forecasting, physical sciences etc. to derive meaningful insights at different points in time. Conventional statistics methods have several limitations to deal with time series data so specialized methods known as time series analysis requires predominantly in such cases. The simplest and most popular method is linear least square method. Least square method gives the trend line to best fit to a time series data. It exhibits several advantages:

- It is very simple method to understand and derive the prediction
- It is to be applicable for all most all applications
- It gives maximum likelihood solutions if correlate with Markov Conditions.

However, it suffers from several critical limitations:

- Sensitive towards outliers.
- Data needs to be normally distributed for better results.
- It exhibits tendency of outfit data.

Quantile Regressing method by utilizing support vector machine approach is an idyllic approach to deal with the limitations of least square regression methods. It has advantages over least square regression. Table 1 describes the comparison between least square and quantile regression.

Table 1. Comparison of least and quantile regression

Parameter	Least square regression	Quantile regression
Prediction	Conditional mean	Condition quantiles
Size of data	Best suit for small data	Requires sufficient data
Distribution of data	Needs normal distributed data	It does not require any assumption in distribution of data. If data is unclear then also it performs well
Preservation	Conditional mean does not preserve under transformation	It preserves under transformation
Computation of data	It does not require rigorous computation so it is cheap	It is computationally rigorous
Response assumption	Constant variance for the response	No constant variance of the response is required

Support vector machine in correlation with quantile regression may produce excellent outcomes for time series analysis. The support vector machine has an ability to solve nonlinear regression estimate problems so it is the prominent candidate for time series data analysis. One more significant feature of SVM is that the learning here is analogous to resolve a problem of linear quadratic optimization. Thus, unlike the other traditional stochastic or neural network methods, the solution obtained by applying the SVM method is always unique and globally optimal.

The Sect. 2 of paper will deal with related work in this field. Section 3 will describe QRSVM model in details. Section 4 will discuss about Experiments and Results of the model. At Last, Paper provides the conclusions of research work carried out.

2 Related Work

Statistical Methods predominantly used for time series data analysis. Autoregressive Integrated Moving Average (ARIMA) model is the most prevalent and commonly used for time series data analysis [6]. Notwithstanding, these sort of models depend on the hypothesis that take into an account that time series must be linear and follows a normal distribution of the data. C. Hamzacebi in 2008 [1] proposed a distinction of ARIMA model called as Seasonal ARIMA (SARIMA). The prototypical produced good results for seasonal time series data, however it required to undertake linear form of associated time series data. The limitations of the linear models could be overcome by non-linear stochastic models [5, 19]. However, the implementation of these kind of models is very complex.

Neural Network based time series models have grown as of late and pulled in expanding considerations [8, 9]. The astounding element of ANNs is their inherent capability of non-linear modeling with no presupposition about the statistical distribution monitored by the annotations. The incredible highlights about ANN based models are self-versatile in nature [28]. There is assortment of ANN models exist in the literature. The Multi-Layer Perceptron (MLP) is the most famous and basic model dependent on ANN [2, 4, 13, 22]. MLPs contain different layers of computational components, unified in a feed-forward way [18]. MLPs utilize a variety of learning techniques, the conspicuous is back-propagation [16, 20, 29] where the output esteems are related with the exact response to compute the value of some foreordained error-function. The error is then served back through the network. Utilizing this data, the algorithm controls the degree of each linkage so as to decrease the estimation of the error function by some insignificant quantity. An overall strategy for non-linear optimization called gradient descent [21, 23] is applied to regulate the degrees. Time Lagged Neural Network (TLNN) is another variation of Feed Forward way [15, 26]. In TLNN, the input nodes are the time series values at some specific lags. Likewise, there is a constant input term, which may be expediently taken as 1 and this is linked to every neuron in the hidden and output layer. The presentation of this constant input unit circumvents the need of separately introducing a bias term. In 2007, Pang et al. [17] introduced one model dependent on neural network and efficaciously applied to the simulation in the rainfall. Li et al. [14], In 2008, presented hybrid model based on AR * and generalized regression neural network model (GRNN) and that gave respectable results in the setting to the time series data. Chen and Chang in 2009 [3] came out with an Evolutionary Artificial Neural Network model (EANN) to build automatically the architecture and the connections of the weights of the neural network. Khashei and Bijari in 2010 [11] introduced a new hybrid ANN model, utilizing an ARIMA model to discover predictions more precise than the model of neural networks. Wu and Shahidehpour [27] proposed a fusion model based on an adaptive Wavelet Neural Network (AWNN) and time series models, such as the ARMAX and GARCH, to predict the day by day estimation of electricity in the market. In [7] researchers proposed a regression neural network model to anticipate widespread time series, which is a fusion of diverse algorithms for machine learning. Artificial Neural Network based algorithms are overwhelming for time series data analysis however they show various constraints such as: appropriate network structure is attained through trial and error, sometimes mysterious performance of the network, usually require more data to train the model fittingly, computationally complex and affluent. Support Vector Machine [12, 24, 25] is the vigorous machine learning technique for the pattern generation and classification.

The proposed model will use SVM as it isn't just intended for decent classification yet additionally expected for an improved speculation of the training data. Solutions obtained by SVM is always unique as it depends on linearly constrained quadratic optimization. The model will use the fusion methodology of SVM and Quantile Regression [10]. Quantile Regression methodology permits for comprehension relationships between variables outside of the average of the data, making it valuable in understanding outcomes that are non-normally dispersed and that have nonlinear relationships with predictor variables.

3 Quantile Regression Support Vector Machine (QRSVM) Model

The least square regression model is representing by the Eq. (1).

$$Y = a + b X + \varepsilon \quad (1)$$

Where Y is Dependent Variable whose value is going to be predicted

a is the intercept of Y

b is the slope of line

ε represents an error and s identically, independently, and normally distributed with mean zero and unknown variance σ^2 .

Least square regression model attempt to define conditional distribution by utilizing the average of a distribution. Another thing is, it assumes that the error term is same across all values of X in which conditional variable (Y/X) to be assumed a constant variance σ^2 . When this assumption fail, we must change the LSR algorithm to accommodate conditional mean and scale. The new equation based on conditional scale is:

$$Y = a + b X + e^r \varepsilon \quad (2)$$

Where r is the unknown parameter

$$\text{Var}(Y/X) = \sigma^2 e^r \quad (3)$$

In this also, conditional scale for dependent variable y is not vary with independent variable X. In order to realize covariate properties in context to dependent variable Quantile Regressing concept is required.

$$Y = a^{(p)} + b^{(p)} X + \varepsilon^{(p)} \quad (4)$$

Where p is the probability and it ranges between 0 and 1.

We specify the pth conditional quantile given X with

$$Q^{(p)}\left(\frac{Y}{X}\right) = a^{(p)} + b^{(p)} X \quad (5)$$

Least square regression having only one conditional mean while Quantile Regression contains numerous conditional quantiles. In the nonlinear quantile regression, the quantile of the dependent variable Y for a given independent attribute X is assumed to be nonlinearly related to the input vector $X_i \in \mathbb{R}^d$ and represented by nonlinear mapping function $\phi(\dots)$. The new version related to nonlinearity characteristic of quantile function is represented as:

$$QX = W\theta \phi(X) \quad (6)$$

Where $\theta \in (0, 1)$,

W_θ is θ^{th} regression quantile.

Absolute deviation loss will occur in quantile regression so SVM with quantile regression plays a vital role. The equation of quantile regression with SVM is represented as:

$$\text{Minimize } \frac{1}{2} \|w\theta\|^2 + C \sum_i \rho(Y_i - W\theta \phi(X_i)) \tag{7}$$

for any $\theta \in (0, 1)$

Equation (7) is considered as QRSVM Model.

4 Experiments and Results

Experiments of proposed study is carried out by considering weather data of Anand District of Gujarat State, India. The sample data is depicted in the Table 1.

Table 2. Sample weather data of Anand district

Temperature(in Celsius)/Month	January	February	March	April	May	June
Avg. temperature	20.5	22.9	27.2	31.1	33.4	32.2
Min. temperature	12	14.1	18.5	22.9	26.4	27.2
Max. temperature	29	31.7	35.9	39.4	40.5	37.3
Rain fall(mm)	1	0	1	0	2	92

Experiment simulation is carried out using R programming language (Figs. 1 and 2).

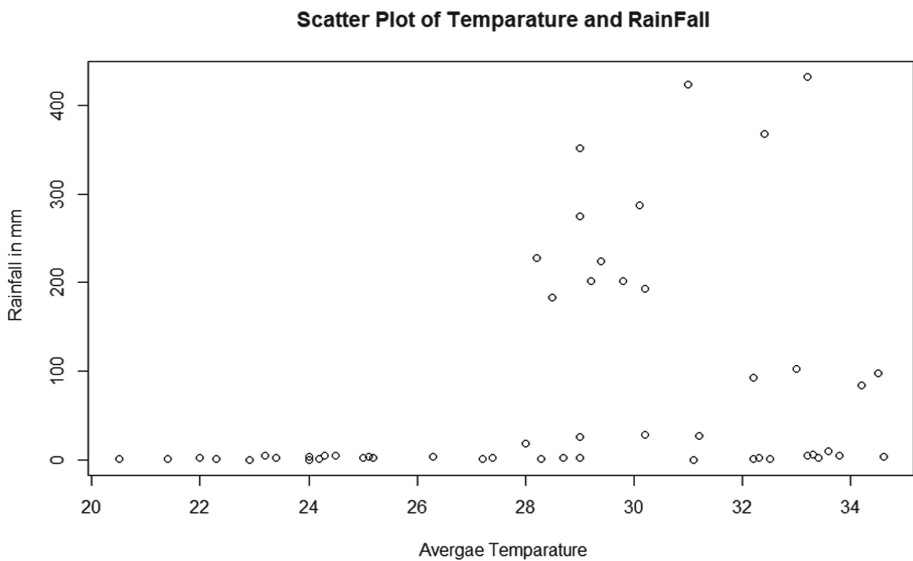


Fig. 1. Scatter plot temperature vs. rainfall

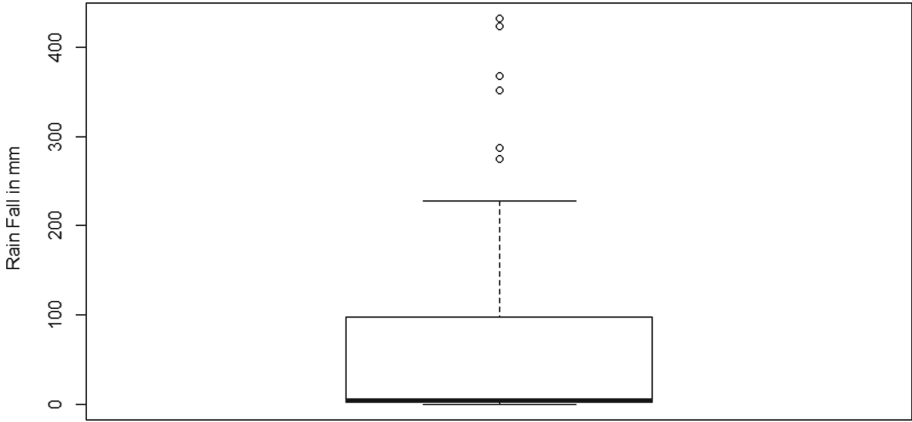


Fig. 2. Box plot for rainfall data

The box plot of rainfall data indicates that several values are outliers. For the accurate prediction of the time series data, outlier values also play an important role.

According to least square regression model, residuals values and Coefficient statistics, are depicted in Table 2 and Table 3 respectively. They having one value based on central tendency value.

Table 3. Residual values of least square regression

Min	First quartile	Median	Third quartile	Max
-7.1849	-2.6955	-0.4849	3.4257	6.8939

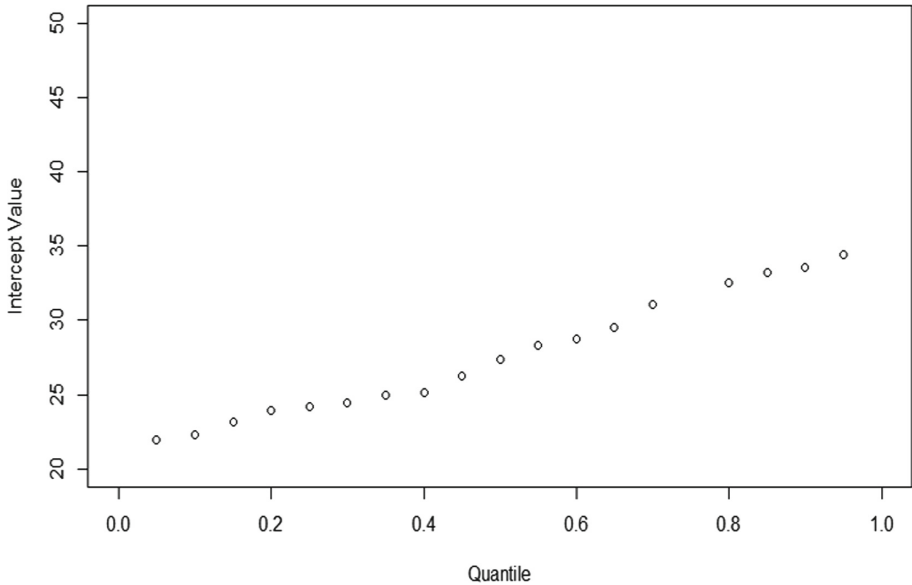
Table 4. Coefficient values of least square regression

	Estimate	Std. error	t-value
Intercept	27.674335	0.620280	44.616
X	0.010579	0.004356	2.429

Quantile Regression SVM model generate coefficient based on quantile value. For the similar data, the coefficient of the QRSVM Model is depicted in Table 4 (Table 5, Figs. 3 and 4).

Table 5. Coefficient values of QRSVM model

Quantile to be estimated	Intercept	X
0.05	21.96	0.02
0.1	22.279384	0.020616
0.15	23.125537	0.018616
0.2	23.950000	0.016666
0.25	24.182481	0.017518
0.3	24.433580	0.016605
0.35	24.964210	0.017894
0.4	25.165614	0.017193
0.45	26.259859	0.013380
0.5	27.3810526	0.0094737
0.55	28.2936019	0.0063981
0.6	28.7135135	0.0054054
0.65	29.5286956	0.0034782
0.7	31.1000000	0.0035326
0.75	3.219946e+0	5.449591e-04
0.8	32.50027248	-0.00027248
0.85	3.320000e+01	-1.431147e-17
0.9	33.60851063	-0.00094562
0.95	34.4413793	-0.00287356

Relationship between Quantile and Intercept based on QRSVM Model**Fig. 3.** Scatter plot quantile vs. intercept value in QRSVM

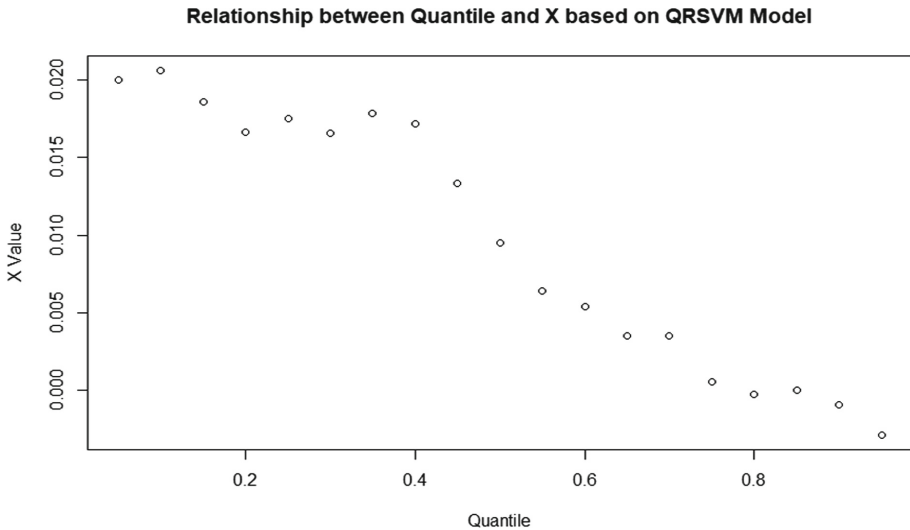


Fig. 4. Scatter plot quantile vs. X in QRSVM

From the results we can see that Least Square regression distributes Intercept and X values on central tendency whereas QRSVM model distributes them with multiple values based on the values of percentile so we can understand and explore the insights with multiple dimensions. The other thing is if outlier exist in the data then central tendency value might be compromised whereas this situation will not affect in QRSVM Model.

5 Conclusions

In this proposed work, it is concluded that Least Square Regression Model exhibits several limitations such as it attempts to define conditional distribution by utilizing only the average of a distribution. Another thing is, it assumes that the error term is same across all values of X in which conditional variable (Y/X) to be assumed a constant variance σ^2 . In order to realize covariate properties in context to dependent variable, Quantile Regressing concept is required. Based on the experiments of time series data of weather, the paper concluded that QRSVM model distributes Intercept and X values in multiple values in order to understand and interpret them effectively. Paper also concluded that the results of LSR model might be compromised to deal to with outliers whereas this situation does not exist in QRSVM model.

References

1. Hamzacebi, C.: Improving artificial neural networks' performance in seasonal time series forecasting. *Inf. Sci.* **178**, 4550–4559 (2008)
2. Calcagno, G., Antonino, S.: A multilayer neural network-based approach for the identification of responsiveness to interferon therapy in multiple sclerosis patients. *Inf. Sci.* **180**(21), 4153–4163 (2010)

3. Chen, Y., Chang, F.-J.: Evolutionary artificial neural networks for hydrological systems forecasting. *J. Hydrol.* **367**, 125–137 (2009)
4. Wang, D., Liu, D., Zhao, D., Huang, Y., Zhang, D.: A neural-network-based iterative GDHP approach for solving a class of nonlinear optimal control problems with control constraints. *Neural Comput. Appl.* **22**(2), 219–227 (2013)
5. Zhang, G.P.: A neural network ensemble method with jittered training data for time series forecasting. *Inf. Sci.* **177**, 5329–5346 (2007)
6. Zhang, G.P.: Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing* **50**, 159–175 (2003)
7. Gheyas, I.A., Smith, L.S.: A novel neural network ensemble architecture for time series forecasting. *Neurocomputing* **74**, 3855–3864 (2011)
8. Kihoro, J.M., Otieno, R.O., Wafula, C.: Seasonal time series forecasting: a comparative study of ARIMA and ANN models. *Afr. J. Sci. Technol. (AJST) Sci. Eng. Ser.* **5**(2), 41–49 (2004)
9. Kamruzzaman, J., Begg, R., Sarker, R.: *Artificial Neural Networks in Finance and Manufacturing*. Idea Group Publishing, USA (2006)
10. Yu, K., Lu, Z., Stander, J.: Quantile regression: applications and current research areas. *J. R. Stat. Soc. Ser. D (The Stat.)* **52**, 331–350 (2003)
11. Khashei, M., Bijari, M.: An artificial neural network (p,d,q) model for timeseries forecasting. *Expert Syst. Appl.* **37**(1), 479–489 (2010). <https://doi.org/10.1016/j.eswa.2009.05.044>
12. Cao, L.J., Tay, F.E.H.: Support vector machine with adaptive parameters in financial time series forecasting. *IEEE Trans. Neural Netw.* **14**(6), 1506–1518 (2003). <https://doi.org/10.1109/TNN.2003.820556>
13. Tawfiq, L.N.M.: Design and training artificial neural networks for solving differential equations. Ph.D. thesis, University of Baghdad, College of Education Ibn-Al-Haitham (2004)
14. Li, W., Luo, Y., Zhu, Q., Liu, J., Le, J.: Applications of AR*-GRNN model for the financial time series forecasting. *Neural Comput. Appl.* **17**, 441–448 (2008)
15. Moseley, N.: Modeling economic time series using a focused time lagged feed forward neural network. In: *Proceedings of Student Research Day, CSIS, Pace University* (2003)
16. Nawi, N.M., Ransing, M.R., Ransing, R.S.: An improved conjugate gradient based learning algorithm for back propagation neural networks. *J. Comput. Intell.* **4**, 46–55 (2007)
17. Pang, B., Guo, S., Xiong, L., Li, C.: A non linear perturbation model based on artificial neural network. *J. Hydrol.* **333**, 504–516 (2007)
18. Prochazka, A.P.: Feed-forward and recurrent neural networks in signal prediction. In: *4th International Conference on Computational Cybernetics*. IEEE (2007)
19. Parrelli, R.: Introduction to ARCH & GARCH models. Optional TA Handouts, Econ 472 Department of Economics, University of Illinois (2001)
20. Rehman, M.Z., Nawi, N.M., Ghazali, M.I.: Noise-induced hearing loss (NIHL) prediction in humans using a modified back propagation neural network. In: *2nd International Conference on Science Engineering and Technology*, pp. 185–189 (2011)
21. Ruder, S.: An overview of gradient descent optimization algorithms. <https://arxiv.org/pdf/1609.04747.pdf>. Accessed 23 Jan 2018
22. Hoda I, S.A., Nagla, H.A.: On neural network methods for mixed boundary value problems. *Int. J. Nonlinear Sci.* **11**(3), 312–316 (2011)
23. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)* (2017). <https://doi.org/10.1109/ICCV.2017.74>

24. Farooq, T., Guergachi, A., Krishnan, S.: Chaotic time series prediction using knowledge based Green's Kernel and least-squares support vector machines. In: *Systems, Man and Cybernetics*, pp. 373–378 (2007)
25. Raicharoen, T., Lursinsap, C., Sanguanbhoki, P.: Application of critical support vector machine to time series prediction. In: *Proceedings of the 2003 International Symposium on Circuits and Systems, ISCAS 2003*, vol. 5, pp. 741–744 (2003)
26. Yolcu, U., Egrioglu, E., Aladag, C.H.: A new linear & nonlinear artificial neural network model for time series forecasting. *Decis. Support Syst.* **54**(3), 1340–1347 (2013)
27. Lei, W., Shahidehpour, M.: A hybrid model for day-ahead price forecasting. *IEEE Trans. Power Syst.* **25**(3), 1519–1530 (2010). <https://doi.org/10.1109/TPWRS.2009.2039948>
28. Wang, X., Meng, M.: A hybrid neural network and ARIMA model for energy consumption forecasting. *J. Comput.* **7**(5), 1184–1190 (2012)
29. Zweiri, Y., Seneviratne, L., Althoefer, K.: Stability analysis of a three-term backpropagation algorithm. *Neural Netw.* **18**(10), 1341–1347 (2005). <https://doi.org/10.1016/j.neunet.2005.04.007>