A Hybrid Prediction Model of Kernel Principal Component Analysis, Support Vector Regression and Teaching Learning Based Optimization Techniques

Mohammed Siddique¹, Tumbanath Samantara¹ and Siba Prasad Mishra²*

¹Department of Mathematics, Centurion University of Technology and Management, Odisha, India.
²Department of Civil Engineering, Centurion University of Technology and Management, Odisha, India.

Authors’ contributions

This work was carried out in collaboration among all authors. All authors read and approved the final manuscript.

ABSTRACT

Forecasting of stock market is considered as one of the most decisive and critical tasks for the data scientists in financial domain. Stock market is one of exciting and demanding monetary activities for individual investors, and financial analysts. The stock market is an inter-connected important economic international business. Prediction of stock price has become a crucial issue for stock investors and brokers. The stock market is able to influence the day to day life of the common people. The stock price is based on the state of market stability. As the dormant high noises in the data impair the performance, reducing the noise would be competent while constructing the forecasting model. To achieve this task, integration of kernel principal component analysis, support vector machine with teaching learning based optimization algorithm is proposed in this research work. Kernel principal component analysis is able to remove the unnecessary and unrelated factors, and reduces the dimension of input variables and time complexity. The feasibility and efficiency of this proposed hybrid model has been applied to forecast the daily open prices of stock.

*Corresponding author: E-mail: 2sibamishra@gmail.com;
This paper is aiming to anticipate the upcoming stock price using machine learning based optimization techniques. The evolution of computing power, database technology, and machine learning algorithms helps to predicts stock market index more accurately. However, high volatility in stock prices makes it difficult to predict the stock market movements. Though many specialized techniques of machine learning such as neural networks, support vector machine, genetic algorithms are already established, there are still scope for developing innovative models or systems which can cater the rising needs of investors [1,2]. The dimension reduction technique Kernel Principal Component Analysis (KPCA) was implemented with Support Vector Regression (SVR) and Teaching Learning Based Optimization (TLBO) to predict the stock price. Here KPCA extracted the relevant features from the data sets which improve the prediction accuracy [3,4].

A supervised machine learning task undergoes two main steps, i.e., training and testing phase. During training phase, the model is constructed and same is tested during the testing phase. Then the selected data are being divided into training (2654) and testing (889) dataset. Now, before entering into the training phase, the dataset undergoes a data pre-processing phase of feature extraction. Then the learning algorithm is selected and its parameters are initialized. In order to end the training process, termination criteria are set. Finally, the training starts and enters an iterative process of parameter optimization and a post processing phase. During this post processing phase, in order to evaluate the model the testing dataset is applied. On application of new unseen and unlabeled data, prediction result is generated [5,6]. In the proposed model, we used SVM for regression, KPCA is used for reducing the dimension and TLBO is used to optimize the values of its free parameters (C and γ) for better forecasting. In this paper, we proposed a hybrid regression model consisting of kernel principal component analysis, support vector regression and particle swarm optimization. This hybrid model, referred as KPCA-SVR-TLBO.

The rest of this paper is organized as follows. Literature review is highlighted in Section-2 and a brief description of KPCA, SVR and TLBO are given in Section 3. Then in Section-4, the implementation of the hybrid model under study, i.e., KPCA-SVR-TLBO, is explained. In Section-5, experimental result analysis and discussion is presented and finally, the paper is concluded in Section-6.

2. LITERATURE REVIEW

From the literature survey we analyzed that the impact of hybrid concepts has improved the prediction accuracy of financial market. Not a particular combination of methods gives good result always. The literature survey speaks about both approaches of machine learning i.e traditional approaches and hybridization approach in the field of financial stock market prediction. In the traditional approaches mostly one technique were used to address the forecasting the stock market for example only ANN and only BP were used by many authors. But with the hybridization of machine learning techniques leads the scope in improvement in accuracy. Traditional regression methods gives optimal result to some extent but a new hybridization method may give more accurate towards optimal solution but it does not mean that always that 1st method has drawback [7,8]. As second method gives more accurate towards result, so that method should be adopted in a particular category of problems. In the forthcoming chapters, hybridizations of different methods were adopted to solve optimal forecasting problems of stock market. Since we are adopting different non traditional methods meant for handling large data set, the method consists of two parts. In first part, the re-organization of huge data set is required where
as in second part; a suitable optimization technique is used [9,10].

In 2016, the authors S. P. Das, N. S. Achary & S. Padhy developed a hybrid model of DR-SVM-TLBO and applied three dimension reduction techniques PCA, KPCA and ICA to reduce the features of input data. In 2017 the authors Xiao Zhong & David Enke analyzed the performance of three dimension reduction techniques PCA, FRPCA and KPCA on financial data. In 2017, the authors Sabyasachi Mohanty and Sudarshan Padhy proposed hybrid OFS–TLBO–SVR model to predicting gross value added at factor cost. Das and Padhy [11] projected he experimental results using the dataset of everyday’s last prices of the COMDEX commodity futures index and it observed that their planned model was very good as well good as camper to SVM and hybrid model of SVM and particle swarm optimization (PSO). Wei-Chiang Hong [12] presented a forecasting model which combines the seasonal recurrent SVR with chaotic ABC algorithm and investigated electric load forecasting of Northeast China. The study employed here for SVR model to solve the non-linear forecasting problem and the messy behaviour of honey bees, to determine suitable values of the three free parameters of SVR, i.e., C, ε, and γ. The performance results of the investigation indicated that the proposed model (namely SRSVRCABC) gives better prediction results than ARIMA and TF-ε-SVR-SA models.

Hong et al. [13] proposed a hybrid model of support vector regression and CGA to forecast the tourism demands. In the proposed model named as SVR-CGA, CGA was employed in overcoming premature local optimum in determining three free parameters of SVR (i.e. α, ε, and C). The empirical results which were evaluated using on MAPE, MAE, and RMSE, demonstrated that the proposed SVR-CGA model outperformed other competing approaches on the data of tourist arrivals in Barbados. Jiang et al. [14] studied the application of KPCA and SVR for reconstruction of cardiac transmembrane potentials. In the hybrid model, SVR addressed the prediction mechanism, PCA and KPCA were used for feature extraction, and GA and simplex optimization method was used to determine the parameters of the SVR. It was found from the analysis that the SVR with feature extraction performed good as compare to that of without feature extraction. Kazem et al. [15] proposed a forecasting model using SVR with chaos-based firefly algorithm for prediction of stock index. The model had three steps in which a delay coordinate embedding method was employed, followed by, a chaotic firefly algorithm was applied for getting optimal free parameters of SVR, then lastly, the optimal SVR was invoked to predict stock market price. The performance of the proposed model, named as SVR-CFA, was also compared with its competing models such as GA-based SVR, CGA-based SVR, firefly-based SVR, ANNs, and ANFIS based on MSE and MAPE. The results demonstrated over its competing models with MSE and MAPE.

3. METHODOLOGY

3.1 Kernel Principal Component Analysis (KPCA)

KPCA is a statistical technique to remove principle inconsistency of sample. The real meaning of this technique is to expose the character of data by bipartition the major factors. KPCA has been primarily applied for reduction of dimension of the sample. The computation of KPCA is to construct a projection from the most important components of high-dimensional sample, onto a lower one. KPCA is very familiar and well known data analysis technique and used to diminish the dimension of a large sample set of variable to a smaller one. New variables are able to be produced by transformation of originals. By this way the number of variables is reduced but almost all of the required information is reserved. These new variables are termed as “principal components”. Extraction of a subset of variables from a larger data set depends upon the highest correlations of the principal component with the original variables. KPCA has several application areas in science and engineering. Consider a non linear transformation \( \phi(x) \) from the original d-dimensional feature space to a D-dimensional feature space, where usually D >> d, then each data point \( x_i \) is projected to a point \( \phi(x_i) \). Vapnic-chervononkis stated that kernel mapping provides greater classification power by transferring the dimension of input space into a higher dimensional space. KPCA extends conventional PCA to a high dimensional feature space using kernel trick and extracts finite number of non linear principal components. KPCA is useful when input data lie on a low dimensional nonlinear hyperplane.

3.2 Support Vector Regression (SVR)

SVR is a supervised machine learning method developed by Vapnik and Cortes (1995). SVR
makes a decision boundary by which the greater part of the data points of the relevant kind falls on the same side of the boundary. Let us consider the data points of an n-dimensional feature vector space \( X = (x_1, x_2, \ldots, x_n) \), from which we construct a hyper plane \( \alpha_0 = \sum_{j=1}^{n} \alpha_j x_j = 0 \), where the boundary of the optimal hyperplane can be obtained by the maximizing the distance from any point to the plane. The maximum margin hyperplane (MMH) separates the similar types of data points. The necessary feature is that only neighboring points to the boundary of the hyperplane are participated in selection keeping all other points as it is. These points are well-known as the support vectors, and support vectors are separated in respective class by a hyperplane, which is called the Support Vector Classifier (SVC). The inner products of support vector classifier are weighted by their labels, and it helps to maximize the distance from support vectors to the hyperplane.

For given a sample data-set \( S = (x_i; y_i); (x_2; y_2); \ldots; (x_l; y_l) \) representing \( l \) input-output pairs, where each \( x_i \in X \subset \mathbb{R}^n \), where \( X \) represents the n-dimensional input sample space and matching target values \( y_i \in Y \subset \mathbb{R}^n \) for \( i = 1, 2, \ldots, l \), where \( l \) is the size of the training data. The purpose of this regression problem is to construct a function \( f : \mathbb{R}^n \rightarrow \mathbb{R} \), to approximate the value of \( y \) for unseen data \( x \), which was not participated in training sample. By taking a nonlinear function \( \phi \), the input data is mapped from \( \mathbb{R}^n \) to a high dimensional space \( \mathbb{R}^m \), \( m > n \), and consequently the estimation function \( f \) is defined as

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

where \( w \in \mathbb{R}^m \) is the regression coefficient vector, \( b \in \mathbb{R} \), is the bias or threshold value. The main intention of the support vector regression is to build a function \( f \) which has the most \( \epsilon \)-deviation from the target \( y_i \). We need to find \( w \) and \( b \) for which the value of \( f(x) \) can be obtained by minimizing the risk.

\[
R_{reg}(w) = \frac{1}{2} ||w||^2 + C \sum_{i=1}^{l} L_\epsilon(y_i, f(x_i))
\]

where \( C \) is the user-defined penalty factor, which determines the trade-off between the training error and the penalizing term \( ||w||^2 \) and \( L_\epsilon(y_i, f(x_i)) \) is the \( \epsilon \)-intensive loss function originally proposed by Vapnik et al., which is defined as

\[
L_\epsilon(y_i, f(x_i)) = \begin{cases} 0, & |y_i - f(x_i)| \leq \epsilon \\ |y_i - f(x_i)| - \epsilon, & |y_i - f(x_i)| > \epsilon \end{cases}
\]

The minimum risk functional equation (2) can be reformulated by introducing non-negative slack variables \( y_i \) and \( \xi_i \) as

\[
R_{reg} (w, y_j, \xi_j) = \text{Minimize} \quad \frac{1}{2} ||w||^2 + C \sum_{j=1}^{l} (y_j + \xi_j)
\]

subject to constraints

\[
\begin{align*}
\begin{cases} y_j - W^T x_j - b \leq \epsilon + y_j \\
W^T x_j + b - y_j \leq \epsilon + \xi_j \\
y_j, \xi_j \geq 0
\end{cases}
\end{align*}
\]

where \( \frac{1}{2} ||w||^2 \) is the regularization term preventing over learning \( (y_j + \xi_j) \) is the realistic risk and \( C > 0 \) is the regularization constant, that controls the trade-off between the empirical risk and regularization term.

By introducing Lagrange multipliers \( \alpha_j, \beta_j, \mu_j \) and \( \eta_j \) the quadratic optimization problem (4) and (5) can be formulated as

\[
L = \frac{1}{2} ||w||^2 + C \sum_{j=1}^{l} (y_j + \xi_j) - \sum_{j=1}^{l} \alpha_j (y_j + \xi_j - \epsilon) - \sum_{j=1}^{l} \beta_j (y_j - \epsilon)
\]

\[
+ \sum_{j=1}^{l} \mu_j (y_j + \xi_j - \epsilon) + \sum_{j=1}^{l} \eta_j (y_j - \epsilon)
\]

The dual of the corresponding optimization problem (4) and (5) is represented as

\[
\text{Maximize} \quad -\frac{1}{2} ||w||^2 + C \sum_{j,k=1}^{l} (\alpha_j - \beta_j)(\alpha_k - \beta_k) (x_j)^T x_k - \epsilon \sum_{j=1}^{l} (\alpha_j + \beta_j) + \sum_{j=1}^{l} (\alpha_j - \beta_j)
\]

Subject to constraints

\[
\begin{cases}
\sum_{j=1}^{l} (\alpha_j - \beta_j) = 0 \\
\alpha_j, \beta_j \epsilon [0, C]
\end{cases}
\]

By changing the equation \( w = \sum_{j=1}^{l} (\alpha_j - \beta_j) x_j \), the function \( f(x) \) can be written as

\[
f(x) = \sum_{j=1}^{l} [(\alpha_j - \beta_j)x_j]^T \phi(x) + b
\]

accordingly by applying Lagrange theory and Karush-Kuhn-Tucker condition, the general
3.4 Implementation of the Hybrid Model

The proposed model is built using three components, i.e., KPCA, SVR, and TLBO. In this process, KPCA plays a key role to extract the most significant attributes from the dataset, SVR is used to address the forecasting mechanism and TLBO optimizes the hyper parameters of support vector regression. Proper selection of kernel type, regularization parameter, and the \( \epsilon \)-insensitive loss of SVR greatly influences the efficiency of the prediction model. The support vector regression function can be expressed as

\[
f(x) = \sum_{j=1}^{n}(\alpha_j - \beta_j)K(x, x_j) + b \quad (8)
\]

where \( K(x, x_k) \) known as Kernel function.

The value of kernel function is equal to the inner product of \( x_j \) and \( x_k \) in the feature space \( \phi(x_j) \) and \( \phi(x_k) \) such that \( K(x_j, x_k) = \phi(x_j) \cdot \phi(x_k) \) \( (9) \)

3.3 Teaching Learning Based Optimization (TLBO)

TLBO is metaheuristic population based optimization model, which imitate the transmission of knowledge in a classroom atmosphere. It perform through two stage one is learner phase another is teacher phase. In the teacher phase a teacher transform the knowledge between the students, and during learner phase the students are acquire the knowledge by interaction among themselves. A set of student’s represents the initial population and known as learners, and then the best student among all is recognized and leveled as the new teacher. This two phased process of acquire knowledge repeated until we get the desired accuracy. The flowchart in Fig. 1, describes the SVR-TLBO working principles.

Initially, recognized the decision variables and set as the features of individual’s population. Let us consider “n” number of learners with “m” number of decision variables. Every decision variables were assigned in a range with upper limit \( U \) and lower limit \( L \) and the features for each learner represented as \( Y[j] \), which initializing randomly in the specified ranges.

\[
Y[j][k] = L[k] + RAND() \times (U[k] - L[k]), \quad \text{where, } j = 1, 2, 3, \ldots, m; \ k = 1, 2, 3, \ldots, n; \ \text{RAND()} \text{ produces number randomly in (0, 1).}
\]

3.4 Implementation of the Hybrid Model

The hybrid model is used to predict the opening stock value and prediction of training and testing data sets to the proposed model to measure the prediction efficiency of both phases. KPCA is the first component that receives the dataset. The KPCA is utilized considering the regression aspect of the problem. Once the task of feature extraction is over, the obtained extracted features are forwarded to the SVR, which is optimized by TLBO.

After splitting the data sets for training and testing phase with a predefined proportion of 80% and 20%, we extracted the relevant data by using KPCA then we initialized the parameters of TLBO and SVR and this filtered data were applied for building the training model. The optimized values of the “C” and “\( \gamma \)” are obtained through TLBO and the termination criteria. Finally to evaluate the efficiency of the proposed KPCA-SVR-TLBO hybrid model, the same is applied to the testing dataset.

3. RESULTS AND DISCUSSION

For the performance measure of the presented hybrid model, we have used three standard statistical metrics. They MAE, RMSE and MAPE and their details have been described in Table 1. Our main intention is to minimize the forecasting error to obtained better accuracy in the proposed hybrid model.

Projected KPCA-SVR-TLBO model is designed with KPCA for feature extraction, SVR as core prediction mechanism and its hyper parameters are optimized by TLBO. Through the training data we built the hybrid model and after completion of training phase, we applied the testing data sets to the proposed model to measure the prediction efficiency of both phases. The hybrid model is used to predict the opening share price. Errors evaluated with MAE, RMSE, and MAPE in training phase are 0.28, 1.08, and 3.12\% (approx) respectively and the errors in testing phase are 0.32, 1.16 and 3.28\% (approx.) respectively.

The Figs. 3 to 5 shows the comparison of the actual stock value and prediction of training and testing phase error graph.
Fig. 1. Working model of SVR-TLBO

Table 1. Performance measure principle for the hybrid model

| SI | Metric | Definition |
|----|--------|------------|
| 1  | MAE    | \( \frac{1}{l} \sum_{i=1}^{l} |y_i - d_i| \) |
| 2  | RMSE   | \( \sqrt{\frac{1}{l} \sum_{i=1}^{l} (y_i - d_i)^2} \) |
| 3  | MAPE   | \( \frac{1}{l} \left( \frac{\sum_{i=1}^{l} |y_i - d_i|}{d_i} \right) \times 100 \) |

Here, \( l \) is the total records for evaluation, \( d_i \) is original value, and \( y_i \) is the estimated value obtained by prediction mechanism.

Table 2. Performance of KPCA-SVR-TLBO models on training and testing datasets

|          | KPCA-SVR-TLBO |
|----------|---------------|
| **Training** |               |
| MAE      | 0.28          |
| RMSE     | 1.08          |
| MAPE     | 3.12 %        |
| **Testing** |              |
| MAE      | 0.32          |
| RMSE     | 1.16          |
| MAPE     | 3.28 %        |
Fig. 2. Flowchart of KPCA-SVR-TLBO model

Fig. 3. Actual vs predicted stock value in training phase
5. CONCLUSION

Our KPCA-SVR-TLBO hybrid model which is comprising of three leading techniques, is applied to the forecasting problem of next day stock price and the results indicate that the model is acceptable not only for research but also from application view point. The testing results obtained from the empirical study demonstrated 0.32 mean absolute error (MAE). Such remarkable performance is achieved due to the application of KPCA on the lagged time-series dataset and use of TLBO to optimize the hyper parameters of support vector regression (SVR). Based on the outcome of this piece of research work, we propose to use our proposed KPCA-SVR-TLBO hybrid model for the future applications of regression based forecasting tasks. This proposed model can be used for taking better decision and more accurate predictions for financial investors on daily stock market forecasting. From the application point of view, we are quite hopeful that our proposed model (KPCA-SVR-TLBO) will be of great help to forecast not only the stock price but also every aspect in the financial domain.

DISCLAIMER

The products used for this research are commonly and predominantly use products in our area of research and country. There is absolutely no conflict of interest between the authors and producers of the products because we do not intend to use these products as an avenue for any litigation but for the advancement of
knowledge. Also, the research was not funded by the producing company rather it was funded by personal efforts of the authors.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Siddique M, Panda D. Prediction of stock index of tata steel using hybrid machine learning based optimization techniques, International Journal of Recent Technology and Engineering. 2019;8(2):3186-3193.
2. Siddique M, Panda D. A hybrid forecasting model for prediction of stock index of Tata Motors using principal component analysis, support vector regression and particle swarm optimization, International Journal of Engineering and Advanced Technology. 2019;1(9):3032-3037.
3. Singh J, Yadav P. A Study on the factors influencing investors decision in investing in equity shares in Jaipur and Moradabad with Special Reference to Gender. Amity Journal of Finance. 2016;1(1):117-130.
4. Xiao Zhong, David Enke. Forecasting daily stock market return using dimensionality reduction, Expert Systems with Applications. 2017;67:126-139.
5. Nelson DMQ, Pereira ACM, De Oliveira RA. Stock market's price movement prediction with LSTM neural networks. Proceedings of the International Joint Conference on Neural Networks. 2017;1419-26.
6. Sabyasachi Mohanty, Sudarsan Padhy. A novel OFS–TLBO–SVR hybrid model for optimal, Journal of Management Analytics. 2017;5(10):1-22.
7. Pesaran MH, Timmermann A. The Robustness and Economic Significance of Predictability of Stock Returns. Journal of Finance. 1995;50:1201-1228.
8. Budget allocation of government schemes to maximize GVA at factor cost, Journal of Management Analytics.
9. Savsani M, Rathod U. Comparative Risk Return Analysis of Bombay Stock Market With Selected Banking Stocks in India. International Journal of Scientific Research in Science and Technology. 2018;4(5):908-916.
10. Shom Prasad Das N, Sangita Achary, Sudarsan Padhy. Novel hybrid SVM-TLBO forecasting model incorporating dimensionality reduction techniques, Appl Intell. 2016;45:1148–1165.
11. Das SP, Padhy S. Support vector machines for prediction of futures prices in Indian stock market, International Journal of Computer Applications. 2015;41(3):22-26.
12. Hong WC. Electric load forecasting by seasonal recurrent SVR (support vector regression) with chaotic artificial bee colony algorithm. Energy. 2011;36(9):5568–5578.
13. Hong WC, Dong Y, Chen LY, Wei SY. SVR with hybrid chaotic genetic algorithms for tourism demand forecasting, Applied Soft Computing. 2011;11(2):1881–1890.
14. Jiang M, Zhu L, Wang Y, Xia L, Shou G, Liu F, et a. Application of kernel principal component analysis and support vector regression for reconstruction of cardiac transmembrane potentials, Physics in Medicine and Biology. 2011;56(6):1727,2011.
15. Kazem A, Sharifi E, Hussain FK, Saberi M, Hussain OK. Support vector regression with chaos-based firefly algorithm for stock market price forecasting, Applied Soft Computing. 2013;13(2):947–958.

Peer-review history:
The peer review history for this paper can be accessed here: https://www.sdiarticle4.com/review-history/72779