Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Stock Price Prediction Using Machine Learning and Swarm Intelligence

*I. Behravan, S.M. Razavi**

Department of Electrical Engineering, Faculty of Engineering, University of Birjand, Birjand, Iran.

Article Info

Article History:
Received 17 February 2019
Revised 27 June 2019
Accepted 08 December 2019

Keywords:
Tehran Stock Exchange market
Automatic clustering
Feature selection
Particle Swarm Optimization
Support Vector Regression

*Corresponding Author's Email Address:
smrazavi@birjand.ac.ir

Extended Abstract

Background and Objectives: Stock price prediction has become one of the interesting and also challenging topics for researchers in the past few years. Due to the non-linear nature of the time-series data of the stock prices, mathematical modeling approaches usually fail to yield acceptable results. Therefore, machine learning methods can be a promising solution to this problem.

Methods: In this paper, a novel machine learning approach, which works in two phases, is introduced to predict the price of a stock in the next day based on the information extracted from the past 26 days. In the first phase of the method, an automatic clustering algorithm clusters the data points into different clusters, and in the second phase a hybrid regression model, which is a combination of particle swarm optimization and support vector regression, is trained for each cluster. In this hybrid method, particle swarm optimization algorithm is used for parameter tuning and feature selection.

Results: The accuracy of the proposed method has been measured by 5 companies' datasets, which are active in the Tehran Stock Exchange market, through 5 different metrics. On average, the proposed method has shown 82.6% accuracy in predicting stock price in 1-day ahead.

Conclusion: The achieved results demonstrate the capability of the method in detecting the sudden jumps in the price of a stock.

Introduction

Nowadays stock markets play a significant role in the countries' economies. Using the stock market, the personal savings and administering funds can be guided through important industries which in turn result in growing the country's economy and also make the investors profit. Investing in a profitable company is critical for investors. So reducing the risk of investing is of great importance. Therefore, the stock market prediction has become one of the interesting and also challenging tasks for researchers and investors. Professional traders use fundamental and technical analysis to predict the future of stocks. In fundamental approaches, the parameters of the company are used while in technical analysis, based on Dow theory [1] the stock price history is important. In the last three decades, several mathematical approaches have been suggested for stock price prediction but the results were not satisfying because of the non-linearity and non-stationary nature of the financial time series data [2]. In recent years, many data scientists have concentrated on developing stock price prediction models based on non-linear and complex machine learning methods such as artificial neural networks (ANN) and regression methods. Several types of research have been conducted in the field of using different types of neural networks for stock price prediction in the last years. Some of these researches have shown that ANN has few drawbacks and

it is not suitable for this purpose due to the complexity of the data and enormous noise. Probably the main reasons are unsuccessful training strategy for multilayer networks and also tend to fall into a local optimum solution which can cause overfitting [3]. Thus using powerful methods to analyze the complex time series data of stock prices is necessary. Metaheuristics are powerful optimization algorithms designed to solve complex and hard optimization problems. In this research, a prediction model based on support vector regression (SVR) [4] and particle swarm optimization (PSO) [5], [6] is introduced which is evaluated on 5 companies' datasets in the Tehran Stock Exchange market. The proposed method works in two-phases. In the first phase, an automatic clustering algorithm, which is based on PSO, is used to cluster the data points and in the second phase a hybrid regression model, which is a combination of PSO and SVR, is trained for each cluster. In both phases, PSO is selected due to its low complexity in compare to other metaheuristics and also effectiveness in solving hard and complex optimization problems, which is also addressed in several literatures [7]-[9].

To estimate the target value of an unknown sample, first, it should be assigned to the closest centroid then, the target value is estimated using the regression model, trained by the samples of the corresponding cluster. Clustering the data points in the first phase, and training independent regression models for different clusters, in the second phase, can remarkably reduce the noise effect, which is shown in our experiments. On the other hand, data complexity and scale are the two main challenges in stock price prediction which can reduce the accuracy of the machine learning method. Dividing the data points into different clusters using an automatic clustering algorithm, can solve this problem. Learning and extracting the hidden patterns and also recognizing the trends hidden in a small cluster of similar samples, are easier tasks compared to learning a huge dataset. Thus, clustering the data points in the first phase and then training different regression models for different clusters, is an effective method for analyzing massive time-series data. Feature selection is another important factor that can improve the accuracy of the proposed method. Eliminating redundant features reduces the ambiguity and complexity of the data points in each cluster.

The main novelties of this research are:

1. Using a novel automatic clustering algorithm that can find the number of clusters in the first phase.

2. Training optimal regression method for each cluster in the second phase. In this optimal regression method, called PSO-SVR, PSO is used for feature selection and parameter optimization for SVR.

The results achieved for 5 symbols in the Tehran Stock Exchange market demonstrate the fabulous performance and effectiveness of the proposed method. The rest of the paper is organized in the following manner: in section 2 the similar works are reviewed and analyzed. In section 3, the raw data and the extracted features are completely described. Sections 4 and 5 are devoted to the proposed method and experimental results respectively.

Related Works

In [10], Gozalpour and Teshnehlab have proposed a prediction method based on deep neural networks. Also, they have used principal component analysis (PCA) and an autoencoder network for dimension reduction. In their research, the data of the past 30 days has been used to predict the closing price of the next day. Their data includes closing price, opening price, highest price, lowest price, and volume of the stock transactions. The method is tested on three NASDAQ symbols. Despite using a complex method, the main drawback of this research is ignoring technical indicators of the stock data. In [11], a method called (2D)2PCA+Deep NN is introduced for stock multimedia prediction. Although, the method has shown improvement in stock multimedia prediction, in case of large window size it has shown poor performance. Besides that, using PCA, as a preprocessing unit, and training a deep neural network is a time-consuming procedure which is a drawback. In another research, Ramezanian et al. have proposed a complicated method which is an integrated framework including genetic network programming (GNP), multilayer perceptron (MLP), and time-series models for stock return forecasting and rule extraction [12]. The performance of the method has been tested on 9 symbols active in the Tehran Stock Exchange market. In their research, the GNP model along with reinforcement learning and multi-layer perceptron is applied to classify data and also time-series models to forecast the stock return. Moreover, the rules of accumulation, based on the GNP model's results are utilized to forecast the return. In addition to the high complexity of the proposed method, which may result in slow training, the major drawback is the high possibility of being trapped in the local optimum point in the training phase of the MLP. In [13] a hybrid method called BOA-SVR is introduced which is a combination of the butterfly optimization algorithm (BOA) and support vector regression machine (SVR). In this research first, the data is prepared by phase space reconstruction. Then the model is trained on the data. In the model, the BOA algorithm is used to tune the parameters of the SVR since its performance is highly dependent on the values of the parameters. In the data preparation phase, technical indicators are not considered while they can

improve the performance of the method. Also, BOA can be used to search the solution space for the best feature subset which is ignored in this research. Zahedi and Rounaghi have researched the Tehran Stock Exchange market [14]. They adopted a three-layer feedforward MLP to predict the stock price on the Tehran Stock Exchange. Also, PCA is used in their model for dimension reduction. Again, the main drawback of their method is neglecting the possibility of falling into a local optimum point during the training phase of the MLP network. This problem becomes more intense when dealing with massive and complex datasets. In another research, Akita and his colleagues presented a deep learning method for stock prediction [15]. The proposed model is tested on 10 companies in the Tokyo stock market. In their paper, they have proposed an approach that converts newspaper articles into their distributed representations via a method, called Paragraph Vector, and models the temporal effects of past events on opening prices about multiple companies with LSTM. Even though the news published in newspapers affects the stock prices, gathering financial news from different newspapers and converting them to numerical information is a very hard task which makes this method less applicable than other methods. In a similar but more comprehensive research, conducted by Khan et al., the related tweets in Twitter and the financial news from Business Insider are extracted and converted into numerical format [16]. Also, the historical data of the stocks are used to create a structured dataset. Several classifiers are used and evaluated in this research. like the previous research, this method suffers from gathering non-uniform and complex data that should be interpreted using sentiment analysis. Rustam and Kindantani have proposed a method to predict the closing price of stocks in the Indonesian stock market [17]. In this method first, a dataset is created by calculating 14 indicators from raw data. After dimensionality reduction using PSO, SVR is used to estimate the closing price of the next day. PSO is used to search the solution space for the best feature subset with the minimum value of MSE function. This means that SVR should be trained, for each search agent, on a massive dataset in each iteration to calculate MSE value. Although SVR is a strong regression method, in the case of processing big datasets it fails to perform well. In our method, which is introduced in section 4, to overcome this problem the dataset was divided into different clusters by using an automatic clustering algorithm, then SVR was trained for each cluster. In [18] Omidi and his colleagues used a simple ANN model to predict the future price of stocks. They have evaluated their method capability on the Tractor manufacturing company in the Tehran stock market. Excluding technical indicators and the high possibility of being trapped in local optimum point in the training process of ANN, are the main disadvantages of their method. In an interesting research, a hybrid model is introduced for stock price prediction by Gocken et al [19]. In this research, the Harmony Search (HS) algorithm is used to find the best architecture of the Jordan Recurrent Neural Network (JRNN) and also the best subset of input variables. HS searches the solution space to find the optimal number of hidden neurons, the best technical indicators as input variables, and the best transfer function. Thus, HS should explore and exploit a very high dimensional feature space which increases the probability of being trapped in the local optimum point. On the other hand, training JRNN for each search agent in each iteration makes their method very slow especially, in the training phase. In another research conducted recently, Nikou and her colleagues investigated the performance of deep learning on stock price prediction [20]. They have used the LSTM network to predict the close price of iShares MSCI United Kingdom. Also, they have compared the achieved results to three data mining techniques (Artificial neural network, SVR, and Random Forest). In their research, the effects of technical indicators are not analyzed and the close price has been just considered as the input value. on the other hand, despite SVR has shown better performance than neural network and Random Forest, its parameters have not been tuned with a powerful optimization algorithm. In 2020, Chandar proposed a hybrid method, called GWO-ENN, for stock price prediction which is evaluated on NYSE and NASDAQ stock data [21]. In this method, GWO is used to optimize the Elman Neural Network. The proposed method uses 10 technical indicators to forecast the close price in 1 day ahead. Unlike the previously mentioned researches, only technical indicators are used as input variables and the historical data of the past days are excluded. Gandhmal and his colleagues have provided a review paper on stock market prediction techniques which is very helpful in analyzing different methods proposed for stock price prediction up to now [22].

Data Preparation

The historical data of the stocks are freely accessible on the website of the Tehran Stock Exchange Market[1]. This historical data contains open price, close price, highest price, lowest price, and volume of the transactions for each working day. Besides that, building an accurate model, it is necessary to calculate technical indicators. In this research the following indicators are used for data preparation: MACD, Bollinger bands (upper band, the middle band, and lower band), fast stochastic,

[1] www.tsetmc.com

SI, and William index. Therefore, a structured dataset containing the historical data and the indicators is created with the window size of 26 [23]. For example, the first object of the dataset contains the historical data and the indicators calculated for the period of 26 days, from the first day to the 26th day, which are listed in a vector form. The second object of the dataset contains similar information from the second to the 27th working day. Other objects of the dataset are calculated using the same scheme. As mentioned before, for each day, the historical data contains 5 items including open price, close price, highest price, lowest price, and volume of the transactions. So for the whole period, each object has 130 items. Also, 7 more items are showing the values of the indicators for the corresponding period. Therefore, each object of the dataset is a vector with a size of 137. The target value of the object is the close price of the first day after the 26 days. For instance, for the first object the close price of day 27, is the target value and similarly, the close price of day 28 is the target value of the second object. This means that the main goal of this research is to estimate the stock price of tomorrow. In other words, to estimate the price of a stock in tomorrow using this model, the information of the past 26 days is needed. In the next section, the proposed method is completely explained.

Proposed Method

A two-phase method is presented in this paper for stock price prediction. In the first phase, an automatic clustering algorithm, called APSO-Clustering, is used to partition the dataset into different clusters. APSO-Clustering, which is designed and developed in our previous project, searches the solution space to find the proper number of clusters and the position of the centroids simultaneously. The effectiveness and the power of this automatic clustering method are proved in our last papers [24], [25]. In the second phase, an optimized regression method called PSO-SVR is trained for each cluster. In this method, PSO is used to find the best feature subset from 137 features of the dataset and to find the optimal kernel function's value of SVR. In the next subsections, support vector regression, APSO-Clustering method, and PSO-SVR are completely clarified.

A. Support vector regression

Support vector regression is the promoted version of the support vector machine [26]. SVR uses a new loss function called ε-insensitive loss function which is used to penalize data greater than ε [27]. SVR is a non-linear kernel-based regression method that provides the best regression hyperplane considering the smallest risk minimization principle in high-dimensional feature space [28]. Assume that $D = \{(x_i . y_i)\}_i^n$ is the training dataset that the regression model is supposed to be built on. In this dataset x_i is the ith data point, y_i is its target value, d is the dimension, and n is the number of data points in the dataset. The following formula shows the function of SVR:

$$y = f(x) = W^T \emptyset(x) + b \tag{1}$$

Where $\emptyset$ is a non-linear mapping function which maps the data points from input space to feature space, W is the vector of weight coefficients and b is the bias term. In the training phase the goal is to find W and b through solving the following optimization problem:

$$\min \frac{1}{2} \|W\|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*)$$
$$\text{s.t.:} \begin{cases} y_i - W^T \phi(x_i) - b \le \varepsilon + \xi_i \\ y_i - W^T \emptyset(x_i) - b \ge -\varepsilon - \xi_i^* \\ \xi_i . \xi_i^* \ge 0 . i = 1 . \cdots . n \end{cases} \tag{2}$$

Where C is the penalty factor, ε is insensitive loss function and ξ_i and ξ_i^* are slack variables that measure the difference between the model's output and the target value beyond ε. After solving the optimization problem through Lagrangian multipliers and conditions for optimality, the following equation can be represented as the dual form of equation (1):

$$f(x) = \sum_{i=1}^{n} (\beta_i - \beta_i^*) \cdot K(x_i . x) + b \tag{3}$$

In this equation, β_i and β_i^* are nonzero Lagrangian multipliers and $K(x_i . x)$ is the kernel function. In this paper, Gaussian kernel function is used which is shown in equation (4):

$$K(x_i . x_j) = exp\left(-\gamma \|x_i - x_j\|^2\right) \tag{4}$$

According to equation (4) defining the value of γ is necessary. In other words, the value of γ has an important effect on the final accuracy of the regression model. Therefore, finding the best value for γ is of great importance which has been done by PSO in this research.

B. APSO-Clustering method

APSO-Clustering is an automatic clustering method that can detect the proper number of clusters in addition to the position of the centroids. Inability in finding the number of clusters is one of the drawbacks of common and popular clustering methods such as K-means and fuzzy C-means which makes them inefficient in clustering big datasets. although many methods have been proposed to estimate the number of clusters such as [29], [30] but in the case of big data clustering, complexity and huge amount of data make them inaccurate. In 2000, a promoted version of K-means, called X-means [31], is introduced which can find the number of clusters, but the main advantage of APSO-Clustering, over X-means, is its high accuracy in

partitioning complex and massive datasets. APSO-Clustering clusters the data points in two stages. In the first stage, detecting the appropriate number of clusters and in the second stage, finding the centroids are the main goals. Both stages are based on a non-automatic clustering method called PSO-Clustering. In other words, the PSO-Clustering method is the basic block in both stages. In the first stage, it is used to find the number of clusters and in the second stage, it is tuned to find the position of the k centroids where k is the number of clusters detected in the first stage. Therefore, PSO-Clustering should be explained completely first.

In PSO-Clustering, PSO is used to find the position of the centroids. As mentioned before, it is a non-automatic clustering method which means that k (number of clusters) should be predetermined by the user. So particles, which are potential solutions to the clustering problem, contain $k \times p$ cells where p is the number of features in the dataset, and k is the number of clusters. For fitness evaluation, Calinski-Harabasz index [32] is used which measures the quality of a clustering. PSO-Clustering and k-means perform the same task in different ways. The main downside of k-means is the high probability of finding the local optimum point instead of global point. This results in poor performance, especially when dealing with massive and complex datasets [33]. In PSO-Clustering this problem is covered by using a metaheuristic optimization algorithm. PSO has shown great performance in solving hard and complex optimization problems [34]. High capability in escaping from local optimum points and finding the global point is the main property of PSO. In the first stage of the APSO-Clustering method, PSO-Clustering should be run sequentially with different values of k. In other words, in a sequence, containing 10 steps, PSO-Clustering is run in each step with a specified value of k which is defined using the following equation:

$$k_{new} = k_{old} \pm \alpha \tag{5}$$

where α is a random integer number, k_{new} is the value of k in the current step and k_{old} is the value of k in the previous step. At the end of the first stage, among the solutions found in different steps of the sequence, the solution with the best fitness value determines the best value of k.

In the second stage, the PSO-Clustering method searches the solution space to find the exact position of k centroids, while k is the output of the first stage. In the second stage, the number of the particles and the iteration number are set to 50 and 600 respectively. These numbers are 5 and 150 in each step of the sequence in the first stage. In Fig. 1, the pseudo-code of the APSO-Clustering method is demonstrated.

C. PSO-SVR

PSO-SVR is a hybrid regression method which has been introduced by Xiong and Xu in 2006 [35]. Also, it has been used in different researches in the last years. For example, in 2015, Hu and his colleagues have used PSO-SVR for short-term traffic flow forecasting [36]. In their research, PSO is used for tuning the parameters of SVR. The performance of SVR is heavily dependent on the kernel function and the value of its parameter. Gaussian kernel (or RBF) is one of the most popular kernel functions which usually yields high accuracy [37]. In our method, PSO is used to find the proper value of the Gaussian function's parameter (γ). Besides that, PSO searches the solution space to find the best feature subset among the 137 features of the dataset. So in the proposed method, PSO is used for feature selection and parameter tuning while in [33], parameter tuning is the only task of PSO.

```
Stage 1:
Inputs: number of sequences, length of sequence, number of population
For i=1 to number of sequences do:
        1.k = Generate a random integer number
        2. beginning population = Generate a random population
        3.Current state.fitness = inf
        4. Current state.k=Ø
        for j=1 to length of sequence do:
            if (j==1) do:
                current state.k=k
            end
            best fitness = PSO-Clustering(k, beginning population)
            if best fitness < current state.fitness
                current state.fitness=best fitness
                current state.k=k
            end
            k = current state.k ± ε
        end
Best[i] = current state.k
End
Output: find the best solution which has the best fitness amount and its corresponding k
and its corresponding beginning population
Stage 2:
Final output = PSO-Clustering(best k, beginning population)
```

Fig. 1: pseudo-code of APSO-Clustering [25].

Each particle contains 137 cells for feature selection, which are encoded in binary form, and one cell for the value of γ. Mean squared error (MSE) function is used for fitness evaluation which is calculated using a 3-fold cross-validation method. In the second phase of the proposed method, for each cluster, a PSO-SVR model is trained. In fact, for each cluster, an SVR is trained with the optimal value of γ and the best feature subset, found by PSO. To estimate the target value of an unknown sample, first, the cluster of the sample is found based on the closeness of the sample to the centroids. Then the corresponding regression model estimates the target value of the unknown sample. In Fig. 2, the flow chart of PSO-SVR is shown. In the next section, the final results of the proposed method for 5 symbols of the Tehran Stock Exchange market are presented.

Experimental results

Five symbols of the Tehran Stock Exchange market are selected for performance evaluation including *Chekaren*, *Shekabir*, *Tepompy*, *Sharak, and Hekhazar*. The accuracy is measured using MSE, MAE, MAPE, RMSE, and R^2 metrics. To compare the performance of PSO with another metaheuristic, the results of GWO-SVR is also measured. This means that the second phase of the proposed method is done by the two metaheuristics to reach a fair comparison between the performance of PSO and GWO. Also, the results of simple SVR are shown for each dataset to analyze the effect of clustering and feature selection. In all of the experiments, 70% of the datasets are used for training and 30% for testing. All of the results, written in the following tables, are for testing data.

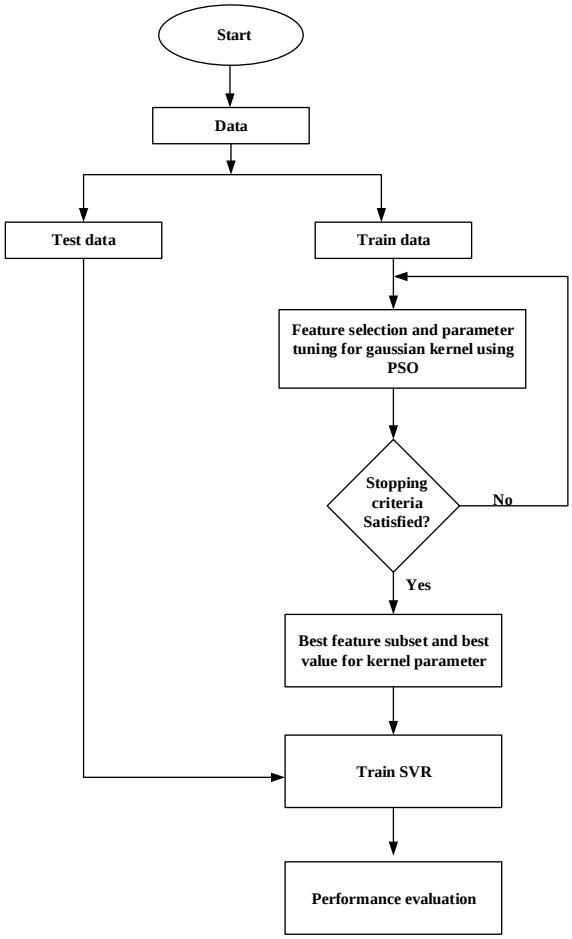

Fig. 2: Flow chart of PSO-SVR.

In the second phase of the proposed method, when PSO-SVR and GWO-SVR are training, 20% of the training data is used for validation which means that for each search agent, after training SVM with 80% of the training data, the fitness value is measured by calculating MSE on the validation data. In PSO-SVR and GWO-SVR, the number of search agents and the maximum number of iterations are set to 15 and 150 respectively which are found after several experiments. For each dataset, the number of detected clusters, the number of selected features, and the optimal values of γ are written separately.

D. Chekaren

Daily information for this symbol (from 2001/3/26 to 2019/11/17) is extracted from the website of the Tehran Stock Exchange market. After creating a dataset using the raw data and technical indicators, in the first phase of the proposed method, APSO-Clustering has divided the data points into 2 clusters. Table 1 and Table 2, presents the details and the accuracy of the results achieved by PSO-SVR and GWO-SVR in the second phase.

Table 1: The details of the results achieved by PSO-SVR and GWO-SVR for Chekaren.

Method	Number of clusters	Number of selected Features	Values of γ
PSO-SVR	2	65, 71	1, 4
GWO-SVR	2	131, 121	1,2

Table 2: The accuracy achieved by the three methods for Chekaren

Method	MSE	MAE	MAPE	RMSE	R^2
PSO-SVR	6.83×10^{6}	1.62×10^{3}	32.62	2.62×10^{3}	0.87
GWO-SVR	8.34×10^{6}	1.88×10^{3}	32.21	2.88×10^{3}	0.84
SVR	5.46×10^{7}	3.64×10^{3}	93.23	7.39×10^{3}	0.0024

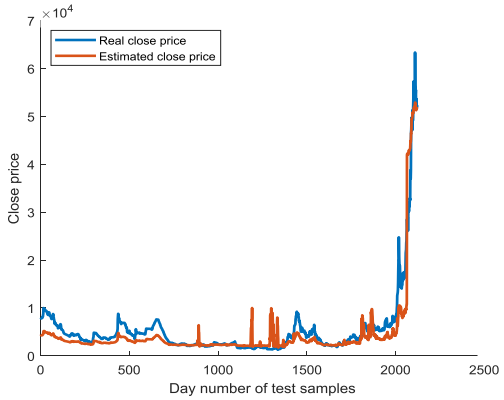

Fig. 3: Real and estimated close price for Chekaren found by PSO-SVR.

According to Table 1, in the second phase, PSO has found 65 and 71 features for the first and second clusters respectively while GWO has detected 131 and 121. Also, the best values found for γ by PSO are 1, 4 for the first and second clusters respectively. These values are 1 and 2 for GWO-SVR. The superiority of the proposed method over the simple SVR with the Gaussian kernel is shown in Table 2. Also in this table, the superiority of PSO over GWO is demonstrated while PSO

has selected fewer features for both clusters in comparison to GWO. In Fig. 3, the blue curve shows the close prices of the test samples and the red curve shows the estimated values which is the output of PSO-SVR. It can be seen, that the red curve approximates the blue curve very well which is confirmed by the amount of R^2. R^2 index for this symbol is 0.87 (for PSO-SVR) which means that the similarity of the two curves is approximately 87 percent while this index is 0.0024 for SVR. According to Fig. 3 and Table 2, the performance of the proposed method, especially in detecting the fluctuations of the stock price, is great. In another point of view, comparing the performance of the proposed method (both PSO and GWO in the second phase) to SVR, we can see the positive effect of clustering and feature selection.

E. Shekabir

For this symbol, which the raw data is recorded from 2011/6/7 to 2019/11/17, 4 clusters are found by APSO-Clustering in the first phase. In the second phase, 4 optimal SVR models are trained for each cluster by PSO and GWO. The number of selected features and the γ values are shown in Table 3. In Table 4, the accuracy measurements achieved by the three methods are shown. According to this table, the accuracy of the proposed method is approximately 87% which means that we can predict the stock price of the next day with an 87% probability of success. Fig. 4 shows the real and estimated price curves. This figure indicates the high capability of the proposed method in detecting the sudden rise and fall in the stock price which is very important for the investors.

Table 3: The details of the results achieved by PSO-SVR and GWO-SVR for Shekabir.

Method	Number of clusters	Number of selected Features	Values of γ
PSO-SVR	4	64, 59, 64, 76	1, 4, 1, 1
GWO-SVR	4	32, 42, 38, 29	2, 4, 1, 1

Table 4: The accuracy achieved by the three methods for Shekabir

Method	MSE	MAE	MAPE	RMSE	R^2
PSO-SVR	1.48×10^6	969.12	14.88	1.21×10^3	0.87
GWO-SVR	1.85×10^6	1.09×10^3	16.25	1.36×10^3	0.84
SVR	4.53×10^6	1.28×10^3	28	2.12×10^3	0.62

According to Table 4, PSO has a slight superiority over GWO while it has selected more features.

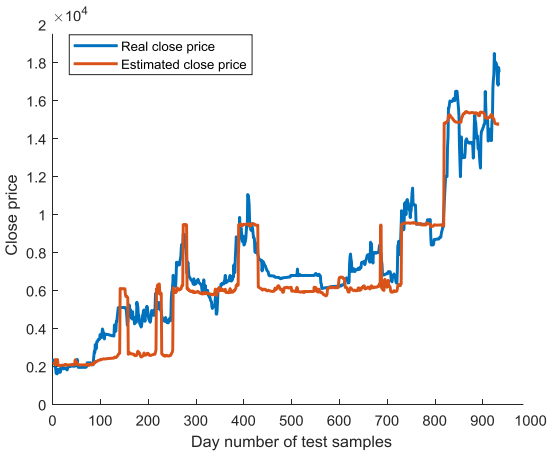

Fig. 4: Real and estimated close price for Shekabir found by PSO-SVR.

F. Tepompy

For this company, the stock price data is recorded from 2001/3/26 to 2019/11/13. In the first phase of the proposed method, the data points are divided into 2 clusters and in the second phase, two SVR models are trained for each cluster using PSO and GWO. Table 5 and Table 6 include the details and accuracy of the three methods. According to Table 6, PSO has shown a better performance than GWO. Also, the superiority of PSO-SVR and GWO-SVR over simple SVR shows the effectiveness of the proposed method. Fig. 5, demonstrates the real and the estimated curves of the close price.

Table 5: The details of the results achieved by PSO-SVR and GWO-SVR for Tepompy.

Method	Number of clusters	Number of selected Features	Values of γ
PSO-SVR	2	67, 68	3, 1
GWO-SVR	2	29, 19	2, 1

Table 6: The accuracy achieved by the three methods for Tepompy.

Method	MSE	MAE	MAPE	RMSE	R^2
PSO-SVR	9.3×10^5	816.76	22.97	966.06	0.63
GWO-SVR	1.06×10^6	906.74	26.27	1.03×10^3	0.57
SVR	2.78×10^6	1.45×10^3	52.09	1.66×10^3	0.10

G. Sharak and Hekhazar

In Tables 7 to 10, the details and the accuracy of the results achieved for these two symbols are presented. Also, the corresponding curves are shown in Fig. 6 and Fig. 7.

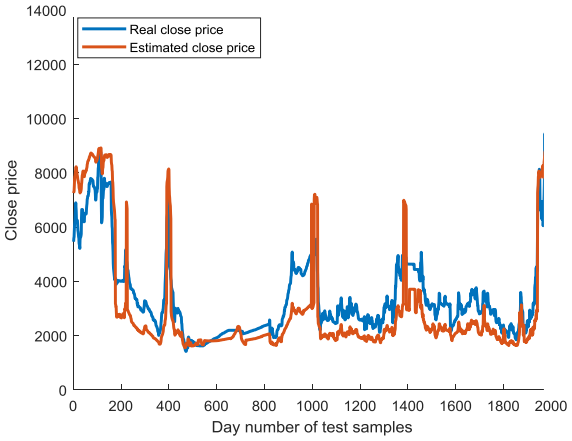

Fig. 5: Real and estimated close price for Tepompy found by PSO-SVR.

Table 7: the details of the results achieved by PSO-SVR and GWO-SVR for Sharak

Method	Number of clusters	Number of selected Features	Values of γ
PSO-SVR	4	73, 74, 66, 68	1, 1, 1, 3
GWO-SVR	4	38, 30, 34, 35	2, 3, 2, 3

Table 8: the details of the results achieved by PSO-SVR and GWO-SVR for Hekhazar.

Method	Number of clusters	Number of selected Features	Values of γ
PSO-SVR	3	65, 65, 66	1, 4, 4
GWO-SVR	3	31, 30, 52	3, 4, 5

Table 9: The accuracy achieved by the two methods for Sharak.

Method	MSE	MAE	MAPE	RMSE	R^2
PSO-SVR	2.60×10^{6}	1.16×10^{3}	17.81	1.61×10^{3}	0.89
GWO-SVR	2.54×10^{6}	1.18×10^{3}	20.26	1.59×10^{3}	0.90
SVR	2.33×10^{7}	3.67×10^{3}	74.84	4.83×10^{3}	0.09

Table 10: The accuracy achieved by the two methods for Hekhazar.

Method	MSE	MAE	MAPE	RMSE	R^2
PSO-SVR	1.61×10^{6}	956.90	19.4	1.27×10^{3}	0.87
GWO-SVR	1.69×10^{6}	980.99	19.67	1.30×10^{3}	0.86
SVR	1.01×10^{7}	2.85×10^{3}	72.89	3.18×10^{3}	0.21

In all of the experiments conducted for these 5 symbols, the proposed method has shown a better performance than SVR. Actually, the tables show the dramatic effect of using clustering and feature selection. Figures 3 to 7 indicate that although in some cases, the proposed method has failed to predict the exact price, in all experiments it has demonstrated a great capability in detecting rise and falls of the stock price.

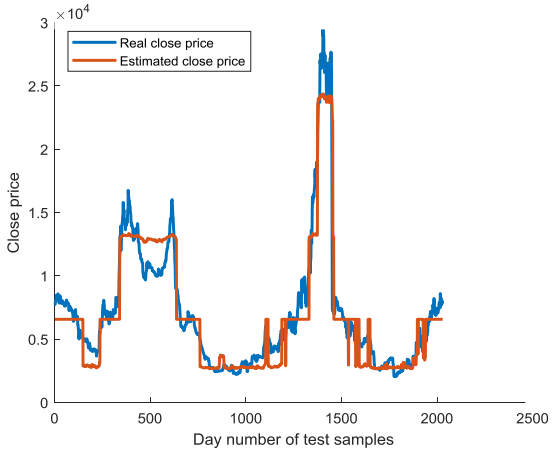

Fig. 6: Real and estimated close price for Sharak found by PSO-SVR.

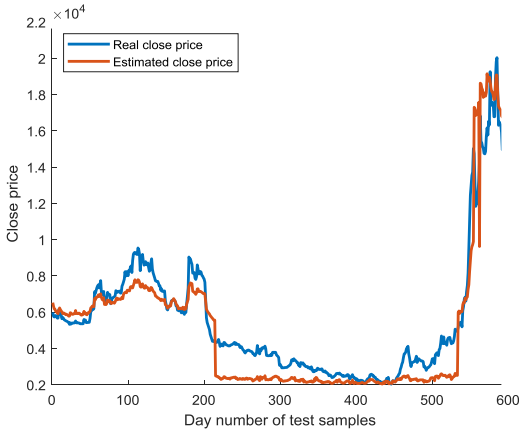

Fig. 7: Real and estimated close price for Hekhazar found by PSO-SVR.

In Table 11, the average accuracy measurements of PSO-SVR and GWO-SVR are written to reach a general comparison between the performance of GWO and PSO. According to this table, although the performances of PSO and GWO are very close to each other, there is a slight difference between their accuracy which shows the superiority of PSO. The most important result which can be extracted from Tables 1 to 11 is the effectiveness of the proposed idea based on using clustering and hybrid regression method for stock price prediction.

Table 11: The average accuracy of PSO-SVR and GWO-SVR

Method	MSE	MAE	MAPE	RMSE	R^2
PSO-SVR	2.69×10^{6}	1.10×10^{3}	21.53	1.53×10^{3}	0.82
GWO-SVR	3.09×10^{6}	1.20×10^{3}	22.93	1.63×10^{3}	0.80

Results and Discussions

Stock price prediction is one of the challenging tasks because of the chaotic and non-linear trend of stock

prices. Due to the availability of the historical data of the stocks in different markets, predicting the future price of stocks has become one of the interesting topics for data scientists. Therefore, different linear or non-linear methods have been introduced in the past years. In this paper, a new machine learning method is introduced to tackle this challenge. The performance of the proposed method is tested on 5 companies in the Tehran stock market. The results indicate that this method has great power in detecting the jumps in the stock price. For future works, we can improve this method to predict the stock price for more than 1 day ahead, which can help us in producing buy and sell signals.

Conclusion

The results indicate that it is possible to predict the stock price trends using swarm intelligence methods with acceptable accuracy. In other point of view, the results demonstrate the power and effectiveness of the swarm intelligence methods in solving hard and complex optimization problems.

Authors contributions

The proposed method is designed and implemented by Iman Behravan and the results are interpreted by Dr. Seyed Mohammad Razavi.

Acknowledgment

The authors acknowledge Dr. Seyed Hamid Zahiri for his help in improving the proposed idea.

Conflict of Interest

The author declares that there is no conflict of interests regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancy have been completely observed by the authors.

Abbreviations

PSO	Particle Swarm Optimization
SVR	Support Vector Regression
ANN	Artificial Neural Network
APSO-Clustering	Automatic PSO-Clustering
PCA	Principle Component Analysis
GNP	Genetic Network Programming
MLP	Multi-Layer Perceptron
BOA	Butterfly Optimization Algorithm
LSTM	Long-Short Term Memory
JRNN	Jordan Recurrent Neural Network
GWO	Grey Wolf Optimizer
MACD	Moving Average Convergence Divergence

References

[1] J. J. Murphy, Technical analysis of the financial markets: A comprehensive guide to trading methods and applications: Penguin, 1999.

[2] J. Patel, S. Shah, P. Thakkar, K. Kotecha, "Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques," Expert systems with applications, 42(1): 259-268, 2015.

[3] H. Larochelle, Y. Bengio, J. Louradour, P. Lamblin, "Exploring strategies for training deep neural networks," Journal of machine learning research, 10(1): 1-40, 2009.

[4] H. Drucker, C. J. Burges, L. Kaufman, A. J. Smola, and V. Vapnik, "Support vector regression machines," in Advances in neural information processing systems, 1997): 155-161.

[5] J. Kennedy and R. Eberhart, "Particle swarm optimization," in Proceedings of ICNN'95-International Conference on Neural Networks,: 1942-1948, 1995.

[6] M. Hasanluo, F. Soleimanian Gharehchopogh, "Software Cost Estimation by a New Hybrid Model of Particle Swarm Optimization and K-Nearest Neighbor Algorithms," Journal of Electrical and Computer Engineering Innovations (JECEI), 4(1): 49-55, 2016.

[7] M. Khosravy, N. Gupta, N. Patel, T. Senjyu, C. A. Duque, "Particle swarm optimization of morphological filters for electrocardiogram baseline drift estimation," in Applied nature-inspired computing: algorithms and case studies, ed: Springer, 2020 : 1-21, 2020

[8] P. Zhang, Z.-Y. Yin, Y.-F. Jin, T. H. Chan, "A novel hybrid surrogate intelligent model for creep index prediction based on particle swarm optimization and random forest," Engineering Geology, 265(1): 105328, 2020.

[9] J. Liang, S. Ge, B. Qu, K. Yu, F. Liu, H. Yang, et al., "Classified perturbation mutation based particle swarm optimization algorithm for parameters extraction of photovoltaic models," Energy Conversion and Management, 203(1): 112138, 2020.

[10] N. Gozalpour M. Teshnehlab, "Forecasting Stock Market Price Using Deep Neural Networks," in 2019 7th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS),: 1-4, 2019.

[11] R. Singh, S. Srivastava, "Stock prediction using deep learning," Multimedia Tools and Applications, 76(18): 18569-18584, 2017.

[12] R. Ramezanian, A. Peymanfar, S. B. Ebrahimi, "An integrated framework of genetic network programming and multi-layer perceptron neural network for prediction of daily stock return: An application in Tehran stock exchange market," Applied Soft Computing, 82(1): 105551, 2019.

[13] M. Ghanbari, H. Arian, "Forecasting Stock Market with Support Vector Regression and Butterfly Optimization Algorithm," arXiv preprint arXiv:1905.11462, 2019.

[14] J. Zahedi, M. M. Rounaghi, "Application of artificial neural network models and principal component analysis method in predicting stock prices on Tehran Stock Exchange," Physica A: Statistical Mechanics and its Applications, 438(1): 178-187, 2015.

[15] R. Akita, A. Yoshihara, T. Matsubara, K. Uehara, "Deep learning for stock prediction using numerical and textual information," in 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS),: 1-6, 2016.

[16] W. Khan, M. A. Ghazanfar, M. A. Azam, A. Karami, K. H. Alyoubi, A. S. Alfakeeh, "Stock market prediction using machine learning classifiers and social media, news," Journal of Ambient Intelligence and Humanized Computing, 1(1): 1-24, 2020.

[17] Z. Rustam, P. Kintandani, "Application of Support Vector Regression in Indonesian Stock Price Prediction with Feature Selection Using Particle Swarm Optimisation," Modelling and Simulation in Engineering, 1(1), 2019.

[18] A. Omidi, E. Nourani, M. Jalili, "Forecasting stock prices using financial data mining and Neural Network," in 2011 3rd International Conference on Computer Research and Development, 3(1): 242-246, 2011.

[19] M. Göçken, M. Özçalıcı, A. Boru, A. T. Dosdoğru, "Stock price prediction using hybrid soft computing models incorporating parameter tuning and input variable selection," Neural Computing and Applications, 31(2): 577-592, 2019.

[20] M. Nikou, G. Mansourfar, J. Bagherzadeh, "Stock price prediction using DEEP learning algorithm and its comparison with machine learning algorithms," Intelligent Systems in Accounting, Finance and Management, 26(4): 164-174, 2019.

[21] S. K. Chandar, "Grey Wolf optimization-Elman neural network model for stock price prediction," Soft Computing, 1(1): 1-10, 2020.

[22] D. P. Gandhmal, K. Kumar, "Systematic analysis and review of stock market prediction techniques," Computer Science Review, 34(1): 100190, 2019.

[23] T. T.-L. Chong, W.-K. Ng, "Technical analysis and the London stock exchange: testing the MACD and RSI rules using the FT30," Applied Economics Letters, 15: 1111-1114, 2008.

[24] I. Behravan, S. H. Zahiri, S. M. Razavi, R. Trasarti, "Clustering a Big Mobility Dataset Using an Automatic Swarm Intelligence-Based Clustering Method," Journal of Electrical and Computer Engineering Innovations, 6(2): 243-262, 2018.

[25] I. Behravan, S. H. Zahiri, S. M. Razavi, R. Trasarti, "Finding Roles of Players in Football Using Automatic Particle Swarm Optimization-Clustering Algorithm," Big data, 7(1): 35-56, 2019.

[26] J. A. Suykens, J. Vandewalle, "Least squares support vector machine classifiers," Neural processing letters, 9(3): 293-300, 1999.

[27] V. Vapnik, The nature of statistical learning theory: Springer science & business media, 2013.

[28] C.Y. Yeh, C.W. Huang, S.-J. Lee, "A multiple-kernel support vector regression approach for stock market price forecasting," Expert Systems with Applications, 38(3): 2177-2186, 2011.

[29] R. Tibshirani, G. Walther, T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(2): 411-423, 2001.

[30] A. Cuevas, M. Febrero, R. Fraiman, "Estimating the number of clusters," Canadian Journal of Statistics, 28(2): 367-382, 2000.

[31] D. Pelleg, A. W. Moore, "X-means: Extending k-means with efficient estimation of the number of clusters," in Icml, : 727-734, 2000.

[32] U. Maulik , S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," IEEE Transactions on pattern analysis and machine intelligence, 24(12): 1650-1654, 2002.

[33] A. S. Shirkhorshidi, S. Aghabozorgi, T. Y. Wah, T. Herawan, "Big data clustering: a review," in International conference on computational science and its applications,: 707-720: 2014.

[34] S. Sengupta, S. Basak, R. A. Peters, "Particle Swarm Optimization: A survey of historical and recent developments with hybridization perspectives," Machine Learning and Knowledge Extraction, 1(1): 157-191, 2019.

[35] W.-l. Xiong, B.-g. Xu, "Study on optimization of SVR parameters selection based on PSO," Journal of System Simulation, 18: 2442-2445, 2006.

[36] W. Hu, L. Yan, K. Liu, H. Wang, "Pso-svr: A hybrid short-term traffic flow forecasting method," in 2015 IEEE 21st International Conference on Parallel and Distributed Systems (ICPADS),: 553-561, 2015.

[37] I. Behravan, O. Dehghantanha, S. H. Zahiri, "An optimal SVM with feature selection using multi-objective PSO," in 2016 1st IEEE Conference on Swarm Intelligence and Evolutionary Computation (CSIEC),: 76-81, 2016.

Biographies

Iman Behravan received his B.S.c in electronics engineering from Shahid Bahonar University of Kerman, Kerman, Iran. Also, he received his M.Sc. and Ph.D. degrees from the University of Birjand, Birjand, Iran. Now he is a post-doctoral researcher at the University of Birjand under the supervision of Professor Seyed Mohamad Razavi. His research interests include Big data analytics, pattern recognition, machine learning, swarm intelligence, and soft computing.

Seyyed Mohammad Razavi received the B.Sc. degree in Electrical Engineering from the Amirkabir University of Technology, Tehran, Iran, in 1994 and the M.Sc. degree in Electrical Engineering from the Tarbiat Modares University, Tehran, Iran, in 1996, and the Ph.D. degree in Electrical Engineering from the Tarbiat Modares University, Tehran, Iran, in 2006. Now, he is an Associate Professor in the Department of Electrical and Computer Engineering, the University of Birjand, Birjand, Iran. His research interests include Computer Vision, Pattern Recognition, and Artificial Intelligence Algorithm.

How to cite this paper:

I. Behravan, S.M. Razavi, "Stock price prediction using machine learning and swarm intelligence," Journal of Electrical and Computer Engineering Innovations, 8(1): 31-40, 2020.

DOI: 10.22061/JECEI.2020.6898.346

URL: http://jecei.sru.ac.ir/article_1421.html