

Stock market closing price prediction based on sentiment of news by using Artificial/Evolutionary Neural Network

Habiba abbasi

Muhammad Younus

Rabbia Anwar

Department of Computer Science

ABSTRACT. Stock market prediction is one of the current red-hot topic in the sector of finance. It's always been difficult task to predict daily prices. A great many researches have been conducted to predict stock market index movement directly. In prediction great many factors are involved, which have a direct influence on index of stock market such as political events, trader expectations, general economic conditions. In this experimental study, our main focus is to design and implement model, which could help stock market investors in predicting the daily closing price of market shares. The novelty of our practice is that, we use a combination of optimized artificial neural network (ANN) along with sentiment of real time news which have a greater impact on stock market values. We have scrapped real-time news from the official DAWN news site. For predictions we use two types of neural network model Artificial Neural Network (ANN) and Evolutionary Neural Network (ENN). We have collected news of the last 8 years along with same period stock market investment data of the HBL from KSE index official website. The input to the system is sentiment of news text having compound, neutral, negative, positive on the basis of that we predict close price for HBL stock shares. The training set is to be determined about 80% of the data set and the rest of 20 % will be used for testing. Training and testing are performed by using two architectures. The architecture of our ANN model is considered with one input layer, two hidden layer and one output layer and for error checking we use Cos function of the mean absolute error.

Keywords— stock price trends, prediction model, Stock market prediction, Machine Learning, sentiment analysis, market trends, news analytics Online news, text data, artificial neural networks.

1. INTRODUCTION:

Stock market price prediction is always an important issue and also, one of the most trending topics in the sector of finance, in which many researchers have played their part in order to obtain more accurate results. As, it has great importance in financial gain, therefore it becomes a subject of interest for great many investors, financial analyst, academic and business side people. Predicting the best time of buying or selling is one of the most difficult tasks. Accurate prediction can help investors to acquire more opportunities of gaining profit in the stock exchange. Hence, precise prediction of the trends of the stock price index can be extremely advantageous for investors [1]. Leung, Daouk [2] holds the view that trading could be made profitable by an accurate prediction of the direction of movement of the stock index. Their work suggested that financial forecasters and traders should focus on accuracy predicting the direction of movement so as to minimize the estimates' deviations from the actual observed values. Mostafa [3] also believes that accurate predictions of the direction of stock price indices are very important for investors. However, the behavior of stock markets is based on a great many factors such as previous open, close, volume, low, high and some political events, trader expectations, general economic conditions.

As, there seems to be a rapid progress in digital data acquisition, which helps in accessing news easily by means of online news services which have been sprouting on the World Wide Web. By the availability of twenty-four hours of digital, new and the combination of both, the latest developments in Natural Language Processing (NLP) and the availability of faster computers lead to the solution of extracting more information from news articles and predict their effects on stock market data.

. The Financial analysts usually aren't aware of the current stock market movements. They are usually puzzled as to which stocks share they should buy and which to sell, to gain more profits. All these users know that the progress depends on numerous relevant news and because of that they should have to deal daily with a vast amount of information. They have to make their-self up-to date about all the news that appears in newspapers, magazines and other textual resources. Doing analysis on such great amount of news and articles is beyond the human capabilities. For the sake of human easiness, the text mining is one of the techniques through which useful knowledge can be automatically extracted.

For the prediction of stock market trend, there are two standard measures, namely, technical analysis and fundamental analysis. Technical analysis is a trading tool employed to evaluate securities and identify trading opportunities by analyzing statistics gathered from trading activity, such as price movement and volume [4]. Whereas, Fundamental analysis is the examination of the underlying forces that affect the well-being of the economy, industry groups, and companies [5].

2. OUR APPROACH:

The novelty of our experimental study is that, we use the essence of both kinds of standard measures of stock market trends.

2.1 Sentiment Analysis of Real time News Feed:

Real time news is generally easily available for the users all the time, from various news sites without visiting them individually. In order to saved time and extract daily news related to market, we have written a scraper in Python by using two libraries, namely, **Request** for data fetching and the other is **Requests-html** (This library intends to make parsing HTML (e.g. scraping the web) as simple and intuitive as possible). that is used for parsing the required news content tag from DAWN, official news website [6]. Through, this scrapper we have separated needed news information from unwanted information. For sentiment analyses with **NLTK** (Natural Language Tool Kit Library)

we use **VADER** library (Valence Aware Dictionary for sEntiment Reasoning is a model used for text sentiment analysis that is sensitive to both polarity (positive/negative) and intensity (strength) of emotion.) [7]. This library polarized news into four categories, namely compound, positive, negative and neutral. The Positive polarization means, it has a positive effect on stock markets. On the other hand, if the negative polarization means, this tends to have a negative impact on stock markets and makes the stock market values go down.

2.2 Stock Market:

A stock market is also known as the equity market. It is a facility where traders can buy and sell securities such as shares, bonds and other financial instrument. In simple words stock market trade is the transfer of money of stock or security from a seller to a buyer or vice versa. This requires two parties to confess on a same price. Equities (stocks or shares) consult is an ownership interest in a particular company. Participants in the stock market range can be from small individual stock investors to larger trader's investors, anywhere in the world. Can perform trading through bank accounts, insurance companies or pension funds, or hedge funds. Their buy or sell orders may be executed on their behalf by a stock exchange trader. Its movement prediction is one of the most difficult task which is influenced by many external factors of social, political and economical. So, the main objective of this experimental study is to improve the prediction accuracy of stock closing price by using the ANN and ENN model. For that, we focus on effective indicators that can be used to predict the output variable. We correlate the variable of sentiment analysis with closing price variables of HBL stock market shares which we get from the daily stock market index. We use a genetic algorithm (GA) ENN model which is employed to improve the prediction accuracy.

3. RELATED WORK:

Previous studies have applied various models in predicting the direction of the stock market index movement. Huang, Nakamori [8] forecasted stock market movement using support vector machines (SVM) and concluded that the model was good at predicting the direction. Kara, Boyacioglu [9] applied Artificial Neural Network (ANN) and SVM in predicting the direction of the Istanbul stock exchange. Their study proves that the two different models are both useful prediction tools, and ANN is significantly better than the SVM model. Şenol and Özturan [10] applied seven different prediction system models for predicting the direction of the stock market index in Turkey, concluding that ANN could be one of the most robust techniques for forecasting. The ANN model has been popularly claimed to be a useful technique for stock index prediction because of its ability to capture subtle functional relationships among the empirical data even though the underlying relationships are unknown or hard to describe [11, 12]. Application of ANN has become the most popular machine learning method, and it has been proven that such an approach can outperform most conventional methods [13–19]. In this study, we attempt to apply an ANN model and ENN model to predict the direction of the HBL stock market shares closing price.

4. PREDICTION MODEL:

4.1 Artificial neural network (ANN) model:

Machine learning tactics are always fascinating artificial intelligence. As, it is based on the principle of learning to train and experience model, in order to make the model more intact that can evaluate itself with rapid change. Connectionist model, such as ANNs, is fitted for machine learning where connection weights are adjusted in order to improve the performance of a network. An ANN is basically a network of nodes that are connected with directed arcs, each have a numerical weight that specify the strength of the connection. These weights indicate the influence of previous node, on to the next node, where positive weights represent reinforcement; negative weights represent inhibition [20]. Generally, the initial connection weights are randomly selected.

In our neural network the data flows forward to the output continuously without any feedback. We have used a typical four-layer neural network model for predicting the closing price of HBL stock shares. The input nodes consist of technical variable, while the output layer provides the predicted result based on sentiments variable and true value (actual closing price).

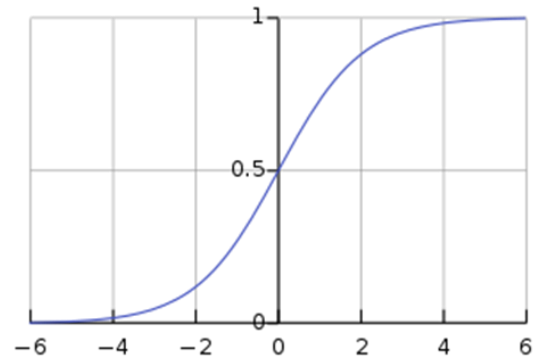
). Hidden nodes with appropriate nonlinear transfer functions are used to process the information received by the input nodes. For calculation equation can be written as Eq. (1).

$$I_j = \sum_i w_{ij} O_i + \theta_j,$$

Where, i is the number of neuron nodes, j is the number of hidden layer neurons, w is weights, and O is input neurons, θ is the bias value of hidden nodes.

Sigmoid Function is used as activation function to calculate the output of the neuron of our hidden layer or output layer. From available sigmoid function namely, logistic and hyperbolic tangent, Rectifier Linear Unit (ReLU). We use logistic function, as it takes the value between 0 and 1 can be written as Eq. (2).

$$f(x) = \frac{1}{1 + \exp(-x)}$$

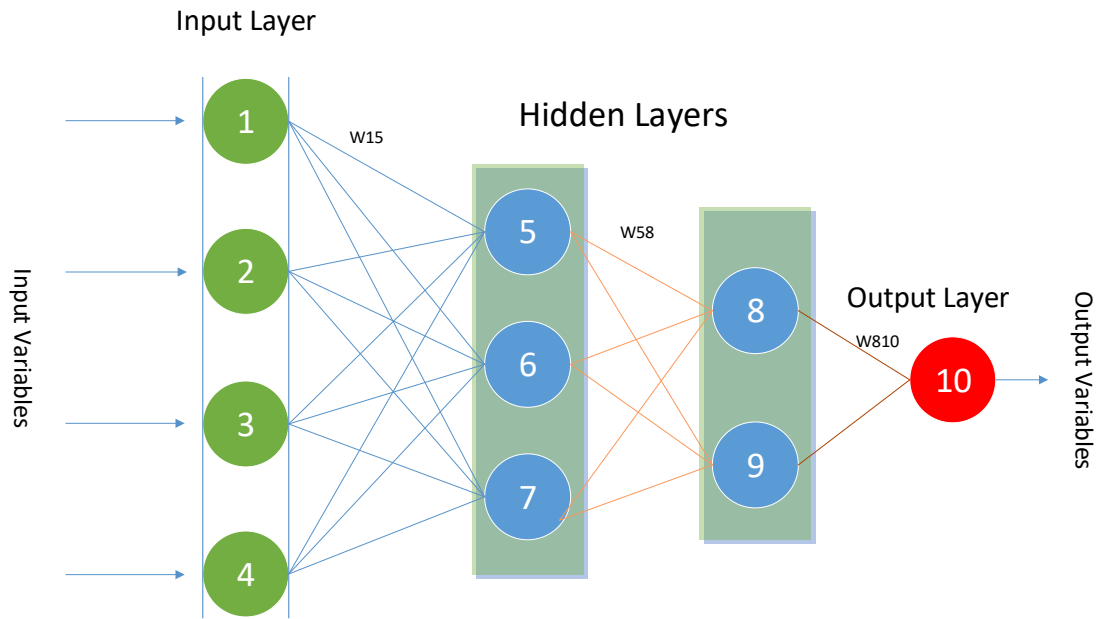


The weights and biases are updated to reflect the propagation errors. Weights are updated by the following equation, written as Eq (3)

$$\Delta w_{ij} = (l) Err_j O_j$$

$$w_{ij} = w_{ij} + \Delta w_{ij}$$

The variable l is the **learning rate**,



ANN model Diagram Figure –1

4.2 Genetic Algorithm (GA):

GA is inspired by biological systems' in which fitness is improved through evolution [21]. In this each individual in the population needs to be described in a chromosome representation. A chromosome should be composed of a sequence of genes from a certain alphabet. An alphabet could be contained binary digits, continues values, integers, symbols, matrices, etc. The representation method determines how the problem is structured in the GA.

In our work, a chromosome is represented by the sum of twenty weights and 6 biases as shown in figure 2. In GA, search starts with an initial set of random solutions known as population. Each chromosome of the population is evaluated using some measure of fitness function which represents a measure of the success of the chromosome. Based on the value of the fitness functions, a set of chromosomes is selected for breeding. In order to simulate a new generation, genetic operators such as crossover and mutation are applied. According to the fitness value, parents and offspring are selected, while rejecting some of them in order to keep the population size constant or considering fittest, which could be survive for a longer period for making next generation. The cycle of evaluation– continued until an optimal or a near-optimal solution is found.

The Crossover allows solutions to exchange information in a way similar to the natural organism undergoing reproduction.

Mutation is used to randomly change (flip) the value of single bits within individual strings to keep the diversity of a population and help a genetic algorithm to get out of a local optimum.

4.3 Evolutionary Neural Network (ENN) model:

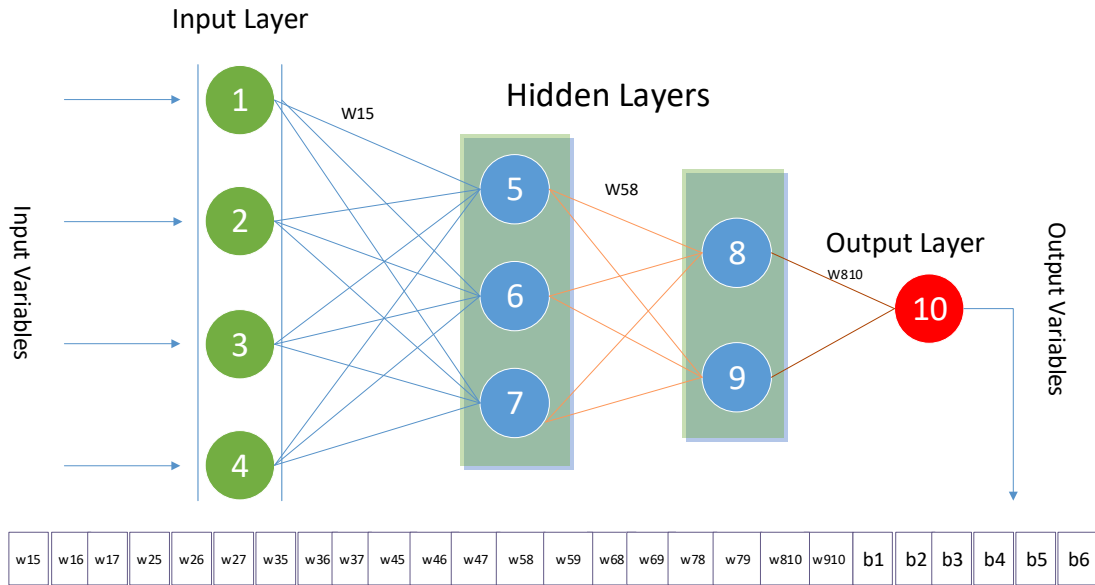
We apply GAs in the form of ENN to evolve the weights between neurons in different layers in the neural network. Steps for evolving is described below:

Step1- Encoding:

Each gene is presented by weight, connected neurons of different layers. Chromosome consists of the combination of weights and biases as shown in figure -2. Our ENN model architecture is based on four neurons in input layer, 5 neurons in two hidden layers and 1 neuron in the output layer. The first gene of chromosome is w_{15} as shown in figure 2. We have used a real number form to represent the connection weights.

Step 2- Generate the initial population:

Initially population is generated randomly. Each of Initial weights are randomly generated between -2 and 2, and biases are between -1 and 1.



Step 3 - Calculating the fitness values:

As regards the fitness function, we have selected the root mean squared error (RMSE) over a training data set. it's the Eq can be written as:

$$RMSE(C_j) = \sqrt{\frac{1}{N} \sum_{i=0}^n (Y_i - P_i)^2}$$

Where Y_i is the actual value and P_i is the output value of i th training data obtained from the neural network using the weights coded in j th chromosome. Where (C_j) and N is the number of training data.

Step 4 – Selection mechanism:

As, in ENN every time we perform crossover and mutation, a new child is created. In order to make predictions more simpler and easy, we find the best individual from population. We use binary tournament selection scheme. In this every time we select two individuals randomly for the selection of one parent that could generate new offspring by using genetic operators. After comparing the fitness of these two selected individuals, only best individual could get selected same process is done for the selection of second parent and by that they are able to make offspring or child that could survive for the next generation.

Step 5 - Genetic operators:

We have performed two-point crossover [26] along with one-point mutation [26] if required.

Step 6 –Replacement:

Replace the current population by the top fittest that could be used for creating further offspring's.

Step 7 - Stopping criteria:

If the number of generations equals to the maximum generation number or optimal criteria, then stop; otherwise go to step 3.

5. PROPOSED WORK

5.1 Data collection:

For our experimental study, we need real time news data that have a variety and reflect the flow of stock market investment data. For this we have collected HBL bank data for past 8 years, From 3rd Jan 2011 to 7th June 2018. Due to its extremely wide range there is need of automated system which should be able to extract a necessary information. For that we have written a news scrapper that scrap, only necessary data required for prediction.

For scrapper code we use a Python library of Request and Requests-html (This library intends to make parsing HTML (e.g., scraping the web) as simple and intuitive as possible).

For Backend we use PyQuery and LXML for preparing and fetching data from the official website of DAWN NEWS and KSE Stock index (official website of stock market). Daily stock prices contain six values as Open, High, Low, Close, Adjusted Close, and Volume. For integrity throughout the project, we considered Adjusted Close price as the stock price.

5.2 Pre-processing:

The dynamics of Stock market price data is not complete understandably as it is closed during weekends and public holidays, when it does not function. We join the real time news scrapped by scrapper with the stock closing price of the same day. We have ignored the gap of missing data, i.e., weekend holidays. For pre-processing of news data, we distinguish news on the bases of tags or token. As the news is split in different tags. We have selected or considered only those tags, that contains a class of theme—Pakistan from the official news site of DAWN after that we read those news that have a tag of the article.

	Symbol	Date	Open	High	Low	Close	Volume
0	HLB	04-Jan-2010	123.40	124.70	123.20	123.79	121296
1	HLB	05-Jan-2010	124.50	128.00	124.31	127.40	866016
2	HLB	06-Jan-2010	128.25	129.99	127.41	128.00	926924
3	HLB	07-Jan-2010	128.49	134.40	127.67	133.61	3268357
4	HLB	08-Jan-2010	133.25	134.00	131.50	133.26	1042705

Figure 5-1

	Date	compund	neg	neu	pos	Close
0	03-Jan-2011	-0.8934	0.172	0.746	0.082	119.49
1	04-Jan-2011	-0.6996	0.133	0.806	0.06	121.87
2	05-Jan-2011	-0.9442	0.319	0.638	0.044	124.93
3	06-Jan-2011	-0.7906	0.179	0.726	0.094	124.98
4	07-Jan-2011	-0.5859	0.14	0.766	0.094	124.99

Figure 5-2

5.3 Input variables:

We get Input variables by performing sentiment analysis on the news. By using an open source python library of VADER [7] we get four factors which are extracted from real time news namely, compound, positive, negative and neutral. We have joined these input variables with stock market closing price inorder to test and train our model of ANN and ENN.

5.4 Training process:

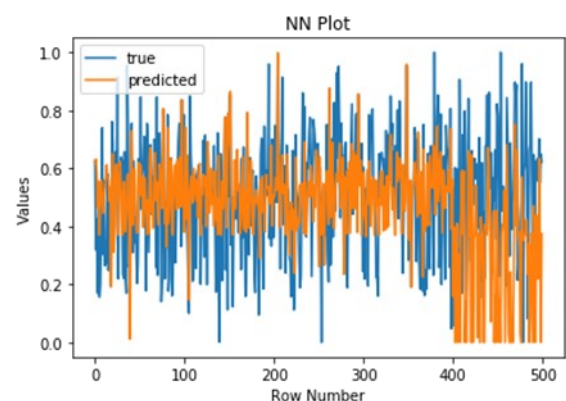
From the above extracted factors that we were extracted using the sentiment analysis are fed to the regression and trained using Artificial neural network (ANN) and Evolutionary Neural network (ENN) model. For training we use 80% of data having approximation of 1.8k plus records. The training of ANN has been done by using the **sklearn python** library [22][23], **pandas** (open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language)[24] and **random** (for generating numbers randomly). Both the textual representations performed well and the results are comparable as shown in fig below.

For the training of ENN model we use python library of pandas [24], **random and numpy** (**numpy** is the core library for scientific computing in Python. It provides a high-performance multidimensional array object, and tools for working with these arrays)[25].

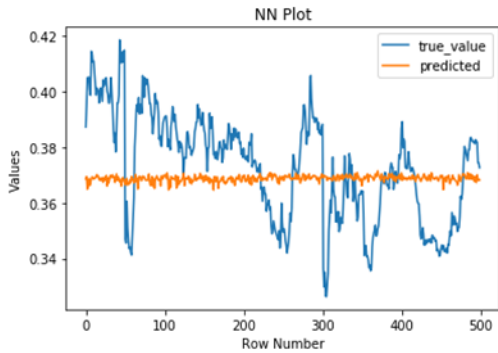
5.5 Testing process:

This gives an overview of accuracy rates of the trained model of ANN and ENN. Following are the graphs.

ANN Accuracy Graph:



ENN Accuracy Graph:



FUTURE WORK:

The future scope of this experimental study would be creating an app in which we add more companies' data and check the prediction accuracy, making a pipeline in which we entertain user with multiple functionalities like user is able to select the desired company after that we scrap news related to that company from the official website of DAWN news and same date data from a stock market official website and train our model to predict the desired selected variable data and shows result on the user screen. For those companies where availability of financial news is a challenge, we would be using twitter data for similar analysis.

CONCLUSION:

Our experimental study concludes that through an artificial neural network (ANN) model and an evolutionary neural network (ENN) model we can predict the variable value, but an ANN model for prediction of the closing price is far better than ENN model. We use python libraries of **pandas**, **psklearn**, **numpy** and for accuracy use **model R2 score** for calculating accuracy of prediction. We have scraper for pre-processing news, data and perform transformation (to improve prediction accuracy through ENN). Our experimental results show that the proposed approach of ANN model is able to cope with closing value of HBL stock shares and it also yields good prediction accuracy in case of closing price prediction as compared to ENN. we would recommend readers of that research not to use ENN model for stock price prediction it can be considered as a less effective approach to predict the stock market.

ACKNOWLEDGEMENTS:

The Authors would like to thank our teacher who supported in the completion of this research project. Appreciating everyone who helped us knowingly or unknowingly in this project.

REFERENCES:

[1]. Gholamiangonabadi D, Mohseni Taheri SD, Mohammadi A, Menhaj MB, editors. Investigating the performance of technical indicators in electrical industry in Tehran's Stock Exchange using hybrid methods of SRA, PCA and Neural Networks. Therm Power Plants IEEE 2014;2014:75–82.

[View Article](#) [Google Scholar](#)

[2]. Leung MT, Daouk H, Chen A. Forecasting stock indices: a comparison of classification and level estimation models. Int J Forecast. 2000;16(2):173–190.

[View Article](#) [Google Scholar](#)

[3]. Mostafa MM. Forecasting stock exchange movements using neural networks: Empirical evidence from Kuwait. Expert Syst Appl. 2010;37(9):6302–6309.

[View Article](#) [Google Scholar](#)

[4]. <https://www.investopedia.com/terms/t/technicalanalysis.asp>.

[5]. http://stockcharts.com/school/doku.php?id=chart_school:overview:fundamental_analysis.

[6]. <https://www.dawn.com/archive/>

[7] -Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014

[8]. Huang W, Nakamori Y, Wang S. Forecasting stock market movement direction with support vector machine. Comput Oper Res. 2005; 32(10):13–22.

- [9]. Kara Y, Boyacioglu MA, Baykan ÖK. Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange. *Expert Syst Appl.* 2011; 38(5):11–19.
- [10]. Şenol D, Özturan M. Stock price direction prediction using artificial neural network approach: the case of Turkey. *J Artif Intell.* 2008; 1(2):70–77.
- [11]. Vellido A, Lisboa PJ, Vaughan J. Neural networks in business: a survey of applications (1992–1998). *Expert Syst Appl.* 1999; 17(1):51–70.
- [12]. Zhang G, Patuwo BE, Hu MY. Forecasting with artificial neural networks: The state of the art. *Int J Forecast.* 1998; 14(1):35–62.
- [13]. Fernando FR, Christian GM, Simon SR. On the profitability of technical trading rules based on artificial neural networks: Evidence from the Madrid stock market. *Econ Lett.* 2000; 69(1):89–94.
- [14]. Lu C. Integrating independent component analysis-based denoising scheme with neural network for stock price prediction. *Expert Syst Appl.* 2010; 37(10):56–64.
- [15]. Versace M, Bhatt R, Hinds O, Shiffer M. Predicting the exchange traded fund DIA with a combination of genetic algorithms and neural networks. *Expert Syst Appl.* 2004; 27(3):17–25.
- [16]. Specht DF. A general regression neural network. *IEEE Trans Neural Netw.* 1991; 2(6):68–76.
- [17]. Wang Z, Wang L, Szolnoki A, Perc M. Evolutionary games on multilayer networks: a colloquium. *Eur Phys J B.* 2015; 88(5):1–15.
- [18]. Wang Z, Kokubo S, Jusup M, Tanimoto J. Universal scaling for the dilemma strength in evolutionary games. *Phys Life Rev.* 2015; 14:1–30. doi: 10.1016/j.plrev.2015.04.033 PMID: 25979121
- [19]. Wang Z, Zhao DW, Wang L, Sun GQ, Jin Z. Immunity of multiplex networks via acquaintance vaccination. *Europhys Lett.* 2015; 112(4):48002–48007.
- [20] Gallant SI. Neural network learning and expert systems. MIT Press, Cambridge, 1993.
- [21] Holland, J.: Adaptation in natural and artificial systems. University of Michigan Press (1975).
- [22]. -API design for machine learning software: experiences from the scikit-learn project, Buitinck et al., 2013.
- [23] -Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay. Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, 12, 2825-2830 (2011)
- [24] - Wes McKinney. **Data Structures for Statistical Computing in Python**, Proceedings of the 9th Python in Science Conference, 51-56 (2010)
- [25] -Travis E, Oliphant. A guide to NumPy, USA: Trelgol Publishing, (2006).
- [26]. Goldberg, D.: Genetic algorithms in search, optimization and machine learning. AddisonWesley, Reading (1989)