Importance of Data Standardization Methods on Stock Indices Prediction Accuracy

Binita Kumari and Tripti Swarnkar

Abstract Stock market indices prediction has drawn huge attention due to its impact on economic stability. Accurate stock market indices prediction is highly essential to reduce the risk associated with it so as to decide good investment strategies. To acknowledge exact prediction, different strategies have been attempted, amid which the machine learning techniques have pinched consideration and been refined achieving extraordinary results in applying machine learning approaches. In our study, we have adopted Support Vector Machine (SVM) for stock market forecasting due to its capacity to deal with risk. SVM in forecasting requires some preliminary works on the data and one of them is standardization. In this study, we analyze four normalization techniques and their influence on the forecasting results. The investigation demonstrates high affectability of the regularly utilized strategies to input information standardization calculations and shows the requirement for a wary way to deal with the outcomes achieved by them.

Keywords Input data standardization · Support vector machines · Stock market indices

1 Introduction

In the data processing field, a very speedily developing technology is data mining. It has been connected to different disciplines, for example, military, engineering, administration, science, and also the business. Within the money-related space,

B. Kumari (✉)
Department of Computer Science & Engineering, Siksha 'O' Anusandhan Deemed to be University, Bhubaneswar, Odisha, India
e-mail: binitakumari@soa.ac.in

T. Swarnkar
Department of Computer Application, Siksha 'O' Anusandhan Deemed to be University, Bhubaneswar, Odisha, India
e-mail: triptiswarnakar@soa.ac.in

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 1082,
https://doi.org/10.1007/978-981-15-1081-6_26

data mining may be utilized to help with the expectation of stock costs, financial assessments, etc.

Stock market indices forecast is viewed as a demanding assignment for the prediction process of financial time-series data as the budgetary market is an intricate, developmental, and nonlinear powerful framework [1]. In the most recent decade, numerous investigations have been led in mining financial time-series information, along with traditional statistical methodologies and data mining procedures. In the territory of financial stock market foreseeing (forecasting), numerous investigations were concentrated on the use of support vector machines [2–5]. As of late, the Support Vector Machine (SVM) strategy that was first proposed by Vapnik in 1995 has been utilized as a part of a scope of uses, including stock market forecast in [2–6]. The SVM procedure is broadly viewed as great classifier and authors in [6, 7] show that SVM forecast ways are better than neural network ways. At first developed for taking care of classification issues, SVM systems can be effectively connected to regression problems. A stock forecast process comprises numerous parts like information gathering, creating integrated information, normalizing information, and classification/prediction.

A piece of the stock market indices forecast process is delivering the criterions (parameters) that depict the outcome imparted in diverse units and scales to a typical and equivalent numeric range. This movement, considered standardization, may have basic effect on the estimation's outcome. In this paper, we will take an insight at the impact of data standardization on stock market prediction.

2 Related Work

Anticipating the stock's trends and critical patterns are appallingly eye-catching to the stock exchange's scientists and any individual who wants to settle on the appropriate stock or potentially the best possible time to look for or offer the stocks [8]. Be that as it may, the right expectation is amazingly troublesome in view of the creaky nature and non-static stock expenses. A few full-scale monetary elements like political occasions, organization's approach, general financial conditions, item esteem lists, interest rates and stocks, desire for speculators, and psychological variables affect the stock expenses [9]. Additionally, government arrangement and administrative measures considerably affect the development of the stocks showcase in general. As per the authors in [10], soft computing procedures are generally utilized for stock market issues and are useful devices for foreseeing the nonlinear behavior. Artificial Neural Network (ANN) and SVM have been used by many researchers for stock foreseeing [11]. But even after constructing so many dynamic models, artificial neural network includes few hindrances inside the learning technique which influences the outcome as shown in [7]. Therefore, a few analysts like picking advanced methodologies based on powerful statistical basis like SVM [12]. As of late, the SVM procedure which is a supervised learning methodology has utilizations in classification and regression problems. SVM shows high performance by minimizing the structural risk as

shown by authors in [13]. Given the over improvements, after all SVM was presented upheld Vapnik's statistical learning hypothesis, a few investigations have practical experience in the theory and its utilizations. Numerous studies utilize the SVM to foresee the time-series data [3, 13] The SVM is a machine learning system which has been created by Vapnik in 1995 and as a result of its eye-catching choices and superb execution in different issues, it has been used for nonlinear predictions. Tai and Cao in [3] attempt to utilize this sort of neural system to anticipate measurement found the SVM to be better than multilayer neural network system with regard to prediction of monetary time series.

Normalization is an integral part of any method wherever data processing techniques are applied. Thus, the result analysis of applying normalization techniques on different domains has been done recently. A large portion of the research work preprocesses the data while not paying any worry to the data complexity. Inquiries have been raised by authors in [14–16] on the requirement of preprocessing based on the data complexity. A preprocessing system called SMOTE ENN for oversampling the unbalanced datasets has been utilized in [17] so as to evaluate the various interims, wherever the usage of oversampling is helpful for the unbalanced datasets. As discussed by authors in [16, 15], the execution of any classification process is also touched with the companionship of noise inside the dataset.Han and Men [18] try and value the impact of normalization on RNA-seq sickness identification. In another paper, Sukirty [19] have evaluated 14 standard learning approaches for constructing a dynamic selection model so as to choose the best normalization process.

Thus, from the literature, it is clear that the normalization technique chosen for performing any data mining functionality may affect the output accuracy. In our paper, we will have a closer look into the importance of normalization for stock prediction.

3 Methods and Materials

3.1 Datasets Used

In order to verify the influence of input data standardization on forecasting performance, this study chooses the NASDAQ and S & P 500 as experimental datasets. The study chooses the data from 4/1/2010 to 30/4/2013. The gathered information comprises every day high, open, closing, and low costs. They are utilized just as informational indexes. The data has been collected from Yahoo finance (https://in.finance.yahoo.com/).

In this paper, the investigation is to foresee the direction of every day stock value record. A major problem in any stock dataset is that it does not contain any class label for up/down. Thus, we use an attribute Δc which indicates change in closing price as described in [20]. Δc has been used as a class label. "1" and "−1" mean the following

day's index is higher or lower than the present day's index, respectively. Forecast miniature is fabricated and the performance is utilized to assess the efficiency.

3.2 Normalization

Normalization is a scaling procedure or a mapping strategy or a pre-handling stage, where we scale input information to fall inside a little indicated range. Basically, normalization of the information is required when managing attributes of various units and scales with the end goal to merge for better outcomes. Unless normalized at preprocessing, variables with disparate ranges or varying precision acquire different driving values. Stronger drivers may obfuscate meaningful variables.

On the other hand, if the mining algorithm has a random sampling component, then normalizing for sample size may help ensuring that all sources are treated equally, and that data-availability bias (and its corresponding misrepresentation of the data universe) is reduced. Normalization of input data plays an important role in the stock prediction process.

We have used the following four standardization methods to examine their influence on stock prediction—Euclidean formula, Manhattan formula, Linear formula, and Weitendorf's linear formula. Jüttler–Korth linear standardization was not used since for positive data values, it is similar to linear formula. The standardization formulas for the four methods used in our paper are listed in Table 1, where A_i represents the *i*th element of a given dataset and n is the total number of records.

As indicated by the authors in [8, 10, 21] and literature study, we found that the standardization methods listed in Table 1 are the widely used standardization methods in various domains like medicine, business, finance, etc.

Based on literature survey, we utilize 70% of the data points (closing cost) as the training information. The rest 30% outstanding data points are utilized as the test information. With the end goal to boost the foreseeing capacity of the miniature, we generated a synthesized dataset which is a dataset consisting of general stock data features along with the technical indicators mentioned in Table 3**. It also consists of Δc as mentioned in** [20] **along with the class label (1/−1)**.

Table 1 List of normalization techniques used for comparison

Sl. No.	Normalization technique	Formula
1	Euclidean	$A_i = \frac{A_{i,j}}{\sqrt{\sum_{i=1}^{n}(A_i)^2}}$
2	Manhattan	$A_i = \frac{A_i}{\sum_{i=1}^{n}\lvert A_i\rvert}$
3	Linear	$A_i = \frac{A_i}{\max A_i}$
4	Weitendorf's linear	$A_i = \frac{A_i - \min A_i}{\max A_i - \min A_i}$

Table 2 List of some commonly used technical indicators

Technical indicators
20-day bias
Rate of change
Stochastic indicator
Relative index
10-day moving average
Moving average convergence/divergence (MACD)
Commodity channel index (CCI)
Buying/selling willingness indicator
Moving average oscillators (MAO)
Buying/selling momentum indicator
Psychological line
Relative strength index (RSI)
Rate of change (ROC)
Stochastic slow
Disparity 5
Momentum
Disparity 10

3.3 Technical Indicators

The input features which are typically utilized for stock market indices are opening value, closing cost, lowest cost, highest cost, and total volume. It has appeared in numerous articles that the technical indicators are useful for stock forecasting [21–23]. Thus, beneath completely extraordinary conditions, a few imperative technical indicators sketched out in Yongtao Vietnamese money-related unit, 2017 has been taken into thought alongside the daily cost and trading volume of the particular stocks. The technical indicators are determined by implementing an equation to the opening value, the lowest value, the highest cost, and trading volume information. Some of the widely used technical indicators are listed in Table 2.

3.4 Support Vector Machines

As shown by authors in [13, 14], Support Vector Machines (SVMs) are administered learning miniatures that examine information and find out the patterns, utilized for regression analyses and classification. It works by developing hyperplanes in a multidimensional space that isolates instances of various class labels. It can

deal with multiple continuous and categorical variables. They are powerful in high-dimensional spaces, notwithstanding when the sum of dimensions is more than the sample numbers. They are memory proficient and flexible.

When applying SVM to monetary prediction, the vital factor that must be thought about is the selection of kernel function. Since the elements of financial time series are powerfully nonlinear, it is naturally considered that the nonlinear kernel functions will deliver higher achievement in comparison to the linear kernel. Several analysts have mentioned the selection of kernel functions [24] *in financial forecasting. In this paper, we have used the Gaussian kernel function due to their flexible nature.*

At the point when the kernel function is picked, two vital parameters (C, γ) should be settled. Parameter C is the expense of C-SVM and parameter γ is the estimation of gamma in kernel function. The estimation of C and γ can clearly influence the execution of SVM. In our test, we have picked $C = 35$ and $\gamma = 0.6$ after trial and error method.

4 Results and Discussion

The data was collected from Yahoo Finance for two datasets, namely, NASDAQ and S & P 500. In this paper, the test is to foresee the heading of every day stock value record as "1" or "−1" indicating a rise or fall in the closing price.

Along with the opening cost, closing cost, lowest cost, highest cost, the total trading volume, five fitting technical indicators have been treated as starting feature pool. As per the authors in [25, 26], the technical indicators are viable apparatuses to portray the genuine market circumstance in financial time-series forecast. They can be more instructive than utilizing pure prices [26]. In light of the audit of domain specialists and literary works, the chosen five technical indicators are Momentum (MTM), Exponential Moving Average (EMA), Relative Strength Index (RSI), Moving Average Convergence/Divergence (MACD), and Moving Average (MA). In Table 3, the formulae for the technical indicators used in our study are given. The details about the formulae can be referred from [20].

Based on literature study, we utilize 70% of the data points (closing cost) as the training information. The rest 30% outstanding data points are utilized as the test

Table 3 Used technical indicators formulae

Technical indicator	Formulae
MA	$\mathrm{MA(N)} = \frac{1}{N}\sum_{i=1}^{n} A_{i,\mathrm{close}}$
EMA	$\mathrm{EMA(N)} = A_{1,\mathrm{close}}$
MACD	$\mathrm{MACD} = \mathrm{EMA}_{12,i} - \mathrm{EMA}_{26,i}$
RSI	$\mathrm{RSI(N)} = 100 - \frac{100}{1+\mathrm{EMA(N)}_{\mathrm{up}}/\mathrm{EMA(N)}_{\mathrm{down}}}$
MTM	$\mathrm{MTM(i, N)} = A_{i,\mathrm{close}} - A_{i-N,\mathrm{close}}$

Table 4 Accuracy results for NASDAQ between Euclidean + SVM, Manhattan + SVM, Linear + SVM, and Weitendorf's linear + SVM

Method	Prediction accuracy
Euclidean + SVM	87
Manhattan + SVM	89
Linear + SVM	88
Weitendorf's linear + SVM	88

information. With the end goal to improve the forecasting capacity of the model, we generated a synthesized dataset.

The synthesized dataset is needed to be normalized so as to get good prediction results. The normalization technique used for the intake data greatly influences the output of the machine learning methods. We have analyzed four different normalization techniques for each of the two datasets. In our study, the normalization techniques which have been considered are Euclidean, Manhattan, Linear, and Weitendorf's linear.

We adequately check the forecasting performance and impact of standardization methods between Euclidean + SVM, Manhattan + SVM, Linear + SVM, and Weitendorf's linear + SVM with the same set of training dataset and testing dataset of NASDAQ and S & P 500, respectively. The evaluation of the model has been done using Matthews correlation coefficient (MCC) so as to avoid the accuracy bias due to data skew [20]. MCC is a single summary value including all four cells of a 2X2 confusion matrix. Given a confusion matrix (TP, FN, FP, TN), MCC is given by

$$\mathrm{MCC} = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}$$

Table 4 lists the accuracy results of NASDAQ between Euclidean + SVM, Manhattan + SVM, Linear + SVM, and Weitendorf's linear + SVM for predicting two class labels, namely, up or down for the test dataset.

From Table 4, we can see that the prediction efficiency of SVM varies when different input data standardization techniques are applied. We can also see that the prediction accuracy of SVM based on Manhattan data standardization is better as compared to Euclidean + SVM, Linear + SVM, and Weitendorf's linear + SVM. Thus, we can say that the prediction accuracy is dependent on the normalization technique implemented for the input data along with other parameters like parameter tuning in the machine learning technique used, etc. As we know, normalization is a scaling procedure to scale input information to fall inside a little indicated range. Thus, when variables with disparate ranges or varying precision acquire different driving values, they may influence the final outcome. Thus, applying same normalization technique on different types of datasets along with the same data mining technique may have different outputs. Similarly, application of different types of normalization techniques on a single dataset may also have different outcomes due to the characteristics of the underlying dataset.

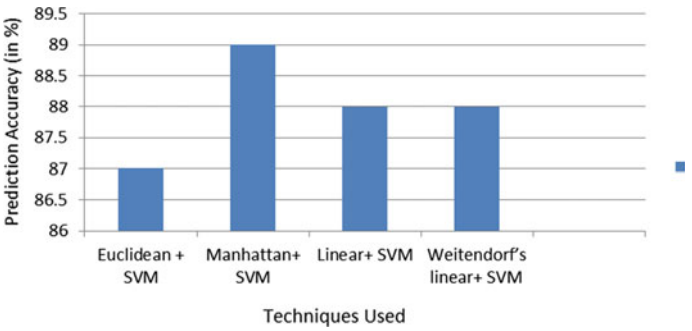

Fig. 1 Comparison results for NASDAQ between Euclidean + SVM, Manhattan + SVM, Linear + SVM, and Weitendorf's linear + SVM

Table 5 Accuracy results for S & P 500 between Euclidean + SVM, Manhattan + SVM, Linear + SVM, and Weitendorf's linear + SVM

Method	Prediction accuracy
Euclidean + SVM	88
Manhattan + SVM	89.1
Linear + SVM	89.8
Weitendorf's linear + SVM	87

Figure 1 shows and compares the results obtained for different techniques in Table 4.

Table 5 lists the accuracy results for S & P 500 between Euclidean + SVM, Manhattan + SVM, Linear + SVM, and Weitendorf's linear + SVM for predicting two class labels, namely, up or down for the test dataset.

From Table 5, we can see that the prediction efficiency of SVM varies when different input data standardization techniques are applied. We can also see that the prediction accuracy of SVM based on linear data standardization is better compared to Euclidean + SVM, Manhattan + SVM, and Weitendorf's linear + SVM. Thus, as seen from Tables 4 and 5, we can say that application of different types of normalization techniques on a single dataset may have different outcomes due to the characteristics of the underlying dataset. Accordingly, we can say that the prediction accuracy is dependent on the normalization technique implemented for the input data along with other parameters. Figure 2 shows and compares the results obtained for different techniques in Table 5.

From our analysis, we find that application of same normalization technique to different datasets may give different levels of results. Thus, the prediction error evaluation results vary from one dataset to another.

Normalization is used to scale input information to fall inside a little indicated range. There may be an influence on the final output when variables with disparate ranges or varying precision acquire different driving values. Thus, application of same normalization technique on different types of datasets using the same data mining

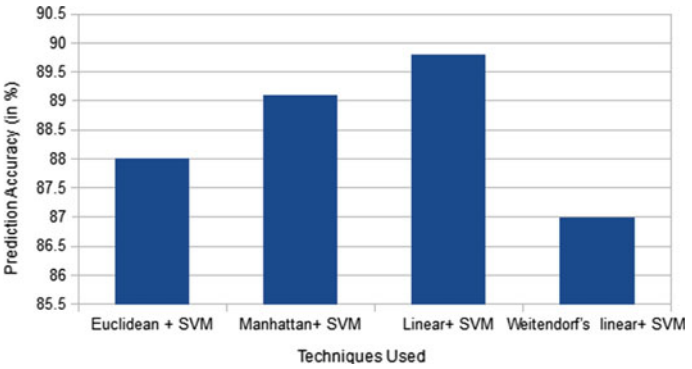

Fig. 2 Comparison results for S & P 500 between Euclidean + SVM, Manhattan + SVM, Linear + SVM, and Weitendorf's linear + SVM

technique may have different outputs. Similarly, application of different types of normalization techniques on a single dataset may also have different outcomes due to the characteristics of the underlying dataset.

Thus, the prediction accuracy results vary from one normalization technique to another. Different normalization techniques may give different prediction accuracy results for the same machine learning algorithm and dataset. Thus, the error accuracy results may also differ for different datasets.

References

1. Abu-Mostafa, Y.S., Atiya, A.F.: Introduction to financial forecasting. Appl. Intell. **6**(3), 205–213 (1996)
2. Huang, W., et al.: Forecasting stock market movement direction with support vector machine. Comput. OR **32**, 2513–2522(2005)
3. Tay, F., Cao, L.: Application of support vector machines in financial time series forecasting. Omega **29**(3), 309–317 (2001)
4. Kim, K.-j.: Financial time series forecasting using support vector machines. Neurocomputing **55**, 307–319 (2003)
5. Cao, L.J., Tay, F.: Support vector machine with adaptive parameters in financial time series forecasting. IEEE Trans. Neural Netw. **14**(3), 1506–1518 (2003)
6. Chen, W.-H., Shih, J.-Y.: Comparison of support-vector machines and back propagation neural networks in forecasting the six major Asian stock markets. Int. J. Electron. Financ. **1**(3), 49–67 (2006)
7. Huang, Z., Chen, H., Hsu, C.J., Chen, W.H., Soushan, W.: Credit rating analysis with support vector machines and neural networks: a market comparative study. Decis. Support Syst. **37**(4), 543–558 (2004)
8. Sahin, U., Ozbayoglu, M.: TN-RSI: Trend-normalized RSI indicator for stock trading systems with evolutionary computation. Procedia Comput. Sci. **36**, 240–245 (2014)
9. Majhi, B., Rout, M., Baghel, V.: On the development and performance evaluation of a multiobjective GA-based RBF adaptive model for the prediction of stock indices. J. King Saud Univ. Comput. Inf. Sci. **32**, 319–331 (2014)

10. Barak, S., Arjmand, A., Ortobelli, S.: Fusion of multiple diverse predictors in stock market. Inf. Fusion **36**, 90–102 (2017)
11. Anbalagan, T., Maheswari, S.U.: Classification and prediction of stock market index based on fuzzy metagraph. Procedia Comput. Sci. **47**, 214–221 (2015)
12. Fernandez-Lozano, C., Canto, C., Gestal, M., et al.: Hybrid model based on genetic algorithms and SVM applied to variable selection within fruit juice classification. Sci. World J. (Article ID 982438), 1797–1805 (2013)
13. García, S., Fernández, A., Luengo, J., Herrera, F.: Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: experimental analysis of power. Inf. Sci. **180**(6), 1506–1518 (2010)
14. Garcia, L.P.F., de Carvalho, A.C.P.L.F., Lorena, A.C.: Effect of label noise in the complexity of classification problems. Neurocomputing **160**, 108–119 (2015)
15. Sáez, J.A., Galar, M., Luengo, J., Herrera, F.: Tackling the problem of classification with noisy data using multiple classifier systems: analysis of the performance and robustness. Inf. Sci. **247**, 1–20 (2013)
16. Leigh, W., Modani, N., Hightower, R.: A computational implementation of stock charting: Abrupt volume increase as signal for movement in New York stock exchange composite index. Decis. Support Syst. **37**(4), 515–530 (2004)
17. Xie, B., Passonneau, R.J., Wu, L., Creamer, G.G.: Semantic frames to predict stock price movement. In: The Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Bulgaria, pp. 873–883 (2013)
18. Han, H., Men, K.: How does normalization impact RNA-seq disease diagnosis?. J. Biomed. Informat. **85**, 80–92 (2018)
19. Jain, S., Shukla, S., Wadhvani, R.: Dynamic selection of normalization techniques using data complexity measures. Exp. Syst. Appl. **106**, 252–262 (2018)
20. Chen, Y., Hao, Y.: A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction. Expert Syst. Appl. **80**, 340–355 (2017)
21. Żbikowski, K.: Using volume weighted support vector machines with walk forward testing and feature selection for the purpose of creating stock trading strategy. Expert Syst. Appl. **42**(4), 1797–1805 (2015)
22. Hsu, M.-W., Lessmann, S., Sung, M.-C., Ma, T., Johnson, J.E.V.: Bridging the divide in financial market forecasting: machine learners vs. financial economists. Expert Syst. Appl. **61**, 215–234 (2016)
23. Neely, C.J., Rapach, D.E., Jun, T., Zhou, G.: Forecasting the equity risk premium: the role of technical indicators. Manag. Sci. **60**(7), 1617–1859 (2014)
24. Cao, L.J., Tay, F.E.H.: Support vector machine with adaptive parameters in financial time series forecasting. IEEE Trans. Neural Netw. **14**(6), 1506–1518 (2003)
25. Yeh, T.-L.: Capital structure and cost efficiency in the Taiwanese banking industry. Serv. Ind. J. **31**(2), 237–249 (2011)
26. Nikfarjam, A., Emadzadeh, E., Muthaiyah, S.: Text mining approaches for stock market prediction. In: The 2nd International Conference on Computer and Automation Engineering (ICCAE), Singapore, pp. 1–2 (2010)