Stock price prediction based on SVM: The impact of the stock market indices on the model performance

Anass NAHIL  
Laboratory of Innovative Technologies (LTI)  
ENSA of Tangier  
Abdelmalek Essaadi University  
anassnahil@gmail.com

Abdelouahid Lyhyaoui  
Laboratory of Innovative Technologies (LTI)  
ENSA of Tangier  
Abdelmalek Essaadi University  
lyhyaoui@gmail.com

Abstract—The challenge of stock forecasting is appealing because a small forecasting improvement can increase profit significantly. However, the volatile nature of the stock market makes it difficult to apply linear models, simple time-series or regression techniques. Consequently, support vector machine (SVM) has become a good alternative. It is a popular tool in time series forecasting for the capital investment industry. This machine learning technique which is based on a discriminative classifier algorithm, forecasts more accurately the financial data. By examining the stock price of 5 Moroccan banks, the experiment shows that the SVM can perform better when we add the global evolution of the market to the independent variables. To express the global evolution of the market, three indices of the Casablanca Stock Exchange are used: MASI, MADEX and Banks Sector Index.

Keywords: Stock price prediction; Financial time series; Support vector machines; the Moroccan Stock Market; Casablanca Stock Exchange.

I. INTRODUCTION

Stock price forecasting is valuable for investors. It tells out the investment opportunities. Unlike other methods which are concerned with company fundamental analysis, in our approach, the independent variables are derived from the stock itself.

Research efforts have been made to find superior forecasting methods, and to enhance existing ones. Many studies have found that univariate time series models are an accurate forecasting models [1], [2], [3]. However, their accuracy require a linear and not very volatile data. The financial data doesn’t satisfy those conditions. Indeed, the stock prices are random and cannot be linearly predicted. Sharda et al.[4], Haykin [5] and Zhang et al. [6] studies’ have shown that, compared to traditional statistical models, neural networks describes more accurately the movement of financial time series. Those Machine learning techniques have been successfully used for modeling financial time series. Hall JW. [7] have used neural networks in an adaptive selection of U.S stocks. S.Kim et al. [8] have applied the probabilistic neural networks to a stock market index. Saad et al. [9] have predicted the trend of a stock market using time delay, recurrent and probabilistic neural networks.

The support vector machine [10], used in this paper, is a neural network technique that has been widely used in stock price predictions. It is a popular tool in time series forecast-

II. SVM REGRESSION THEORY

Given a set of data points \( \{(x_1, y_1), \ldots, (x_l, y_l)\} \), \( x_i \in \mathbb{X} \subset \mathbb{R}^n \) is the input vector, \( l \) is the total number of data patterns and \( y_i \in \mathbb{R} \) is the \( i^{th} \) value of the dependent variable. The estimating function \( f \) is approximated using the following:

\[
y = f(x) = (w.\phi(x)) + b
\]

where \( \phi(x) \) is the high dimensional feature space which is non-linearly mapped from the input space \( x \). The coefficients \( w \) and \( b \) are estimated by minimizing the risk function:

\[
R_{SV}(C) = \frac{1}{n} \sum_{i=1}^{n} L_{c}(d_i, y_i) + \frac{1}{2} \|w\|^2
\]

\[
L_{c}(d_i, y_i) = \begin{cases} |d - y| - \epsilon & |d - y| \geq \epsilon \\ 0 & \text{otherwise} \end{cases}
\]

\( L_{c} \) is the extension of \( \epsilon \)-insensitive loss function. Considering the slack variables \( (\zeta_i, \zeta_i^*) \), the problem can be reformulated as:

\[
(P1) : \begin{aligned}
\text{Minimize} & \quad C \sum_{i=1}^{l} \left( \zeta_i + \zeta_i^* \right) \frac{1}{2} \|w\|^2 \\
\text{Subject to} & \quad \begin{cases} y_i - (w.\phi(x_i)) - b \leq \epsilon + \zeta_i \\
y_i (w.\phi(x_i)) - b \leq \epsilon + \zeta_i^* \end{cases} \quad i = 1, 2, \ldots, l
\end{aligned}
\]

\( C \) is an user specified constant. The solution of (P) using primal dual method leads to the following new problem:

\[
(P2) : \begin{aligned}
\text{Maximize} & \quad Q(\alpha_i, \alpha_i^*) = \sum_{i=1}^{l} y_i (\alpha_i - \alpha_i^*) \\
& - \epsilon \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} (\alpha_i - \alpha_i^*) (\alpha_j - \alpha_j^*) K(x_i, x_j) \\
\text{Subject to} & \quad \begin{cases} \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) = 0 \\
0 \leq \alpha_i \leq C \\
0 \leq \alpha_i^* \leq C \end{cases} \quad i = 1, 2, \ldots, l
\end{aligned}
\]

The typical examples of Kernel function are as follows:

- Linear: \( K(x_i, x_j) = x_i^T x_j \)
- Polynomial: \( K(x_i, x_j) = (\gamma x_i^T x_j + r)^d \), \( \gamma > 0 \)
- Radial basis function (RBF): \( K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \), \( \gamma > 0 \)
- Sigmoid: \( K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r) \).

The kernel parameters \( (\gamma, r \) and \( k \) should be carefully chosen. Those parameters control the complexity of the final solution by defining the structure of the high dimensional feature space \( \phi(x) \). The solution of the primal (P2) yields:

\[
w = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) \phi(x_i) + b
\]

and \( b \) is calculated using Karush-Kuhn-Tucker conditions (KKT):

\[
\begin{cases}
\alpha_i (\epsilon + \zeta_i + y_i + w.\phi(x_i) + b) = 0 \\
\alpha_i^* (\epsilon + \zeta_i^* - y_i + w.\phi(x_i) + b) = 0 \\
(C - \alpha_i) \zeta_i = 0 \\
(C - \alpha_i^*) \zeta_i^* = 0
\end{cases}
\]

Since \( \alpha_i, \alpha_i^*; \zeta_i^* = 0 \) for \( \alpha_i^* \in (0, C) \), \( b \) can be computed as follows:

\[
\begin{cases}
b = y_i - w.\phi(x_i) - \epsilon & 0 \leq \alpha_i \leq C \\
b = y_i - w.\phi(x_i) - \epsilon & 0 \leq \alpha_i \leq C
\end{cases}
\]

For those \( \alpha_i, \alpha_i^* \) for which \( x_i \) corresponds to \( 0 \leq \alpha_i \leq C \) and \( 0 \leq \alpha_i^* \leq C \) are called support vectors. The number of support vectors is a function of \( \epsilon \), the larger the \( \epsilon \), the fewer the number of support vectors and thus the sparser the representation of the solution. However, a larger \( \epsilon \) depreciate the model accuracy. In this regard, \( \epsilon \) should be a compromise between the sparseness of representation and the closeness to data.

Considering the previous results, \( f(x) \) can be computed as:

\[
f(x) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) K(x_i, x) + b
\]

III. PROPOSED METHODOLOGY

A. Experimental design

For each model in this paper, the SVMs regression forecasting follows the procedure in figure 1.

B. Data sets

Five stock prices of the five major Moroccan banks, collated from The Casablanca Stock Exchange, are examined in the experiment. They are: Banque Populaire (BP), Attijari Wafa Bank (AWB), Banque Marocaine pour le Commerce et l’Industrie (BMCI), BMCE Bank (BMCE) and CIH Bank (CHI). We have taken 1119 samples for each of the stocks mentioned above. The corresponding time period is from 2nd January, 2012 to 30th June, 2016. We use two-thirds of the research data points (the first 746 closing price) as the training data. The 373 remaining data points are used as the test data.
The collected data consists of daily closing price. They are used as the data sets. Table 1 shows high price, low price, mean, median and standard deviation (SD) of the five stocks collected for our experiment. The calculated standard deviation shows that the stocks of the banks sector are relatively not very volatile which explains the good performances of the models used in this paper.

C. Data preprocessing

The original closing price is transformed into a five-day relative difference in percentage of price (RDP), as suggested by Thomason [28], in order to enhance the forecasting ability of the model. The transformed data are more symmetrical and follow more closely a normal distribution. One more transformation called exponential moving average (EMA15) from the closing price. The subtraction is performed to eliminate the trend. EMA15 is obtained after a smoothing of the closing price with a three-day exponential moving average. The smoothing enhances the prediction performance of the model. The transformed data are more symmetrical by Thomason [28], in order to enhance the forecasting ability of the model. The transformed data are more symmetrical.

D. Kernel function and parameters selection

A kernel is a function that satisfied the Mercers condition. The two typical kernels used in the Vapniks SVM for regression are the polynomial Kernel and the Gaussian Kernel. The Gaussian kernel function performs well under general smoothness assumptions. Polynomial Kernel takes a longer time in training SVMs and gives inferior result compared to Gaussian Kernel. When training a model, the parameters that gives the best performance should be selected. Generally, in the SVR, those values of $\epsilon$ that produces the best result on the validation set of our data are selected. In the case of a Gaussian Kernel, we introduce two additional parameters: $C$ and $\gamma$. A model selection must be done to identify good $(C,\gamma)$ so that the classifier can accurately perform in the test data. We use a grid-search on $C$ and $\gamma$ based on a cross-validation process. The values of the parameters used in the selection process are illustrated in table 3.

E. Performance Criteria

The prediction accuracy is evaluated using the statistical metrics in table 4, namely, the normalized mean squared error (NMSE), mean absolute error (MAE), directional symmetry (DS) and the coefficient of determination $R^2$.

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Calculation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>$P(i) = EMA_{15}(i)$</td>
</tr>
<tr>
<td>RDP</td>
<td>$RDP_i = \frac{P(i) - P(i-2)}{P(i-2)} \times 100$</td>
</tr>
<tr>
<td>Output</td>
<td>$RDP^{EMA}<em>{i+5} = \frac{EMA</em>{i+5} - EMA_{i}}{EMA_{i}} \times 100$</td>
</tr>
</tbody>
</table>

Table 2: Performance indicators

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Calculation</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMSE</td>
<td>$NMSE = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{x}_i)^2$</td>
</tr>
<tr>
<td>MAE</td>
<td>$\frac{1}{N} \sum_{i=1}^{N}</td>
</tr>
<tr>
<td>DS</td>
<td>$DS = \frac{100}{N} \sum_{i=1}^{N} \delta_i$ where $\delta_i = \begin{cases} 1 &amp; (x_i - x_{i-1})(\hat{x}<em>i - \hat{x}</em>{i-1}) \geq 0 \ 0 &amp; \text{otherwise} \end{cases}$</td>
</tr>
<tr>
<td>$R^2$</td>
<td>$1 - \frac{\sum_{i=1}^{N} (x_i - \hat{x}<em>i)^2}{\sum</em>{i=1}^{N} (x_i - \mu)^2}$</td>
</tr>
</tbody>
</table>

Table 4: Calculation of the performance metrics

The prediction accuracy is evaluated using the statistical metrics in table 4, namely, the normalized mean squared error (NMSE), mean absolute error (MAE), directional symmetry (DS) and the coefficient of determination $R^2$. Recommended by Thomason [28]. The second model (SVM+ model) applies the same indicators to each bank then adds the three indices of the Casablanca stock exchange (MASI, MADEX and Banks Sector Index) to the input (independent variables). Since outliers may make it difficult to arrive to an effective solution, values beyond a set limit of standard deviation are selected as outliers. They are replaced with the closest marginal values. A limit of ±2 SD is set for the RDP values and a limit ±50 SD is set for indexes values. Another pre-processing technique used in this study is data scaling. In support vector machines, feature scaling improves the convergence speed of the algorithm [29]. For that reason, All the data points are scaled into the range of [0.9; 0.9].
IV. RESULTS AND DISCUSSION

The SVM model applied to the data sets let as draw the following figures. The selected model from the training set is applied to the test set. The figures illustrate the real values and the predicted values for the test set of each bank using SVM and SVM+. Globally, SVM+ performs better, but, in some data points, predicted values by SVM are closer to the real values. However, The prediction made by the SVM+ model remains more risk-averse. Indeed, the predicted values by SVM+ are, globally, bellow those predicted by SVM. This observation is due to the addition of the market global evolution to the model. This variable reflects the average evolution of the market which is more risk-averse.

The optimal values of the parameters of SVM chosen based on the validation set are given in Table 5. The results obtained are given in Table 6. Obviously, SVM+ converges to a better performance indicators on the test set. SVM+ provides a smaller NMSE and MAE and larger DS and $R^2$ than SVM+ in all the cases (tables 6 and 7).

<table>
<thead>
<tr>
<th>Bank</th>
<th>SVM Parameters</th>
<th>SVM+ Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCP</td>
<td>$C = 0.00001$, $\gamma = 100$</td>
<td>$C = 0.00001$, $\gamma = 1000$</td>
</tr>
<tr>
<td>BMCI</td>
<td>$C = 0.00001$, $\gamma = 10$</td>
<td>$C = 0.00001$, $\gamma = 10$</td>
</tr>
<tr>
<td>AWB</td>
<td>$C = 0.001$, $\gamma = 1000$</td>
<td>$C = 0.001$, $\gamma = 10$</td>
</tr>
<tr>
<td>BMCE</td>
<td>$C = 0.0001$, $\gamma = 100$</td>
<td>$C = 0.001$, $\gamma = 100$</td>
</tr>
<tr>
<td>CIH</td>
<td>$C = 0.01$, $\gamma = 10$</td>
<td>$C = 0.001$, $\gamma = 10$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bank</th>
<th>SVM NMSE</th>
<th>SVM MAE</th>
<th>SVM+ NMSE</th>
<th>SVM+ MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCP</td>
<td>0.002811</td>
<td>1.208029</td>
<td>0.002585</td>
<td>1.201884</td>
</tr>
<tr>
<td>BMCI</td>
<td>0.001260</td>
<td>6.014052</td>
<td>0.001250</td>
<td>5.976358</td>
</tr>
<tr>
<td>AWB</td>
<td>0.002137</td>
<td>2.034262</td>
<td>0.001341</td>
<td>0.269665</td>
</tr>
<tr>
<td>BMCE</td>
<td>0.000273</td>
<td>1.185295</td>
<td>0.001759</td>
<td>1.129769</td>
</tr>
<tr>
<td>CIH</td>
<td>0.001268</td>
<td>3.297368</td>
<td>0.001016</td>
<td>3.217085</td>
</tr>
</tbody>
</table>

TABLE 5: The selected simulation parameters

TABLE 6: The converged indicators : NMSE and MAE
TABLE 7: The converged indicators : DS and R²

<table>
<thead>
<tr>
<th></th>
<th>SVM</th>
<th>SVM+</th>
<th>SVM</th>
<th>SVM+</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCP</td>
<td>65.19</td>
<td>65.78</td>
<td>95.24%</td>
<td>95.25%</td>
</tr>
<tr>
<td>BMCI</td>
<td>84.37</td>
<td>85.55</td>
<td>97.85%</td>
<td>97.86%</td>
</tr>
<tr>
<td>AWB</td>
<td>62.54</td>
<td>62.54</td>
<td>96.29%</td>
<td>96.79%</td>
</tr>
<tr>
<td>BMCE</td>
<td>69.62</td>
<td>70.50</td>
<td>95.92%</td>
<td>96.35%</td>
</tr>
<tr>
<td>CH</td>
<td>66.37</td>
<td>67.55</td>
<td>98.13%</td>
<td>98.20%</td>
</tr>
</tbody>
</table>

Reviewing results in regards to the NMSE, the model accuracy improvement is the greatest and increases by 26.97% for AWB. For BMCI, the NMSE remains almost constant (table 8). This result is supported by the market capitalization of each bank that reflect the importance and economic weight of the bank in the market. A stock price that has an important market capitalization is more correlated to the global trend of the market, thus the information provided by the global evolution of the market contributes more significantly to the model improvement.

V. CONCLUSIONS

The information contained in the trend of the stock market helps to improve the performance of the SVM regression. In order to evaluate the accuracy improvement in the model, this paper proposes to add the trend of the stock market to the input variables of the SVM model. The global evolution of the market is represented by the indices of the stock market and the sectoral index. The enhanced model provides a better fitting for the stock prices, particularly for those with an important market capitalization.

REFERENCES