A CASE STUDY OF STOCK INVESTMENT BASED ON NEURAL NETWORK TECHNIQUES

[1] MUH-CHERNG WU, [2]CHUN-TZU CHEN, [3]HUI-CHIH HUNG

[1,2,3]Department of Industrial Engineering and Management, National Chiao Tung University,Hsin-Chu City, 30010, Taiwan
E-mail: [1]mcwu@cc.nctu.edu.tw,[2]aeww452721@gmail.com,[3]hhc@cc.nctu.edu.tw

Abstract - Stock price predictions based on data mining techniques have been widely examined. Most studies focused on predicting the stock price or up/down in the next trading day. Considering a future time horizon (say, 100 days), this research attempts to predict whether a day is a "buy-day". A buy-day denotes a "good" day to purchase a particular stock if the stock closing price is expected to rise over 10% in the coming 100 days. The "buy-day" decision is a binary classification problem, which herein shall be solved by various artificial neural network (ANN) models. In an ANN model, the output involves two classification states: "buy-day" or "not-buy-day"; and the input can involve up to 15 variables (i.e., features). By selecting different portfolios of input features, three ANNs are established. The stock price ranging from Jan. 2007 to Dec. 2016 (10 years) of a Taiwanese company (Foxconn) is used as a test case. Numerical experiments reveal that the 3rd ANN outperforms the other ANNs; and its average annual return of investment is about 10.92%.

Keywords - Binary classification, Data mining, Neural network, Stock price prediction

I. INTRODUCTION

Predictions of stock prices have been widely studied by the application of various data mining techniques. Recent example studies include the use of artificial neural network (ANN) [1], support vector machine [2], Bayesian analysis [3], K-nearest neighbors [4], decision tree [5], and meta-heuristic algorithms [6]. Hybrid methods have also been developed; for example, the integration of ANN and meta-heuristic algorithms [7].

In prior studies, most works focused on predicting the stock price of next trading day based on different input variables. For example,Laboissiere *et al.* [8] attempted to forecast the maximum and minimum stock day prices of power distribution companies by considering 40 possible input variables. Yan *et al.* [9] used six input variables to predict the one-day future closing-index of Shanghai composite index. Zhang *et al.* [10] proposed a two-stage data mining method to predict four stock indices. Rather than predicting stock prices, some studies attempted to predict the up/down of a stock in next trading day [11]-[13]. Moreover, Oliveira *et al.* [14] proposed a method to assess the value of microblogs (e.g., Twitter) to forecast daily stock market variables like returns, volatility, and trading volume in next trading day.

A few studies attempted to predict stock prices or their up/down in a future time horizon. For example, Zhang *et al.* [15] propose a status box method, in which a status box is intended to model the stock price trend (up, down, or flat) in a certain period of time (e.g, 30 days). Machine learning techniques are used to classify the status of a box in order to identify buy/sell points. Shynkevich *et al.* [16] explore the performance of a predictive system in different combinations of forecast horizon (1-30 days) and input window length (3-30 days), in which technical indicators are input features andfuture directions of stock price movements are output vectors. Considering a future time horizon (say, 100 days), this research attempts to predict whether a day is a "buy-day". A buy-day denotes a "good" day to purchase a particular stock if the stock closing price is expected to rise over 10% in the coming time horizon (100 days). The "buy-day" decision is a binary classification problem, which herein is solved by various ANN models. In an ANN model, the output involves two classification states: "buy-day" or "not-buy-day"; and the input can involve up to 15 variables (i.e., features). By selecting different portfolios of input features, three ANNs are established. The stock price ranging from Jan. 2007 to Dec. 2016 (10 years) of a Taiwanese company (i.e., Foxconn Inc.) is used as a test case. Numerical experiments reveal that the 3rd ANN outperforms the other two ANNs; and its average annual return of investment is about 10.92%. The remainder of this paper is organized as follows. Section 2 presents the architecture of an ANN and its application methodology (training and testing). Section 3 introduces 15 possible input features. Three portfolios of input features are selected; each portfolio represents a particular ANN model. Section 4 reveals numerical experiments of the three ANNs on the prediction and investment of a Taiwanese stock (Foxconn). Conclusions are in Section 5.

II. ARTIFICIAL NEURAL NETWORK

The architecture of a typical artificial neural network (ANN) model is shown in **Fig.1.**, which involves three layers: input layer, hidden layer, and output layer. Each layer is composed of a set of nodes; and a link exists between each node of a particular layer (e.g., input layer) to each node of its next layer (e.g., hidden layer). A weighting parameter w_{ij} shall be defined on a link that connecting node *i* and node *j*.

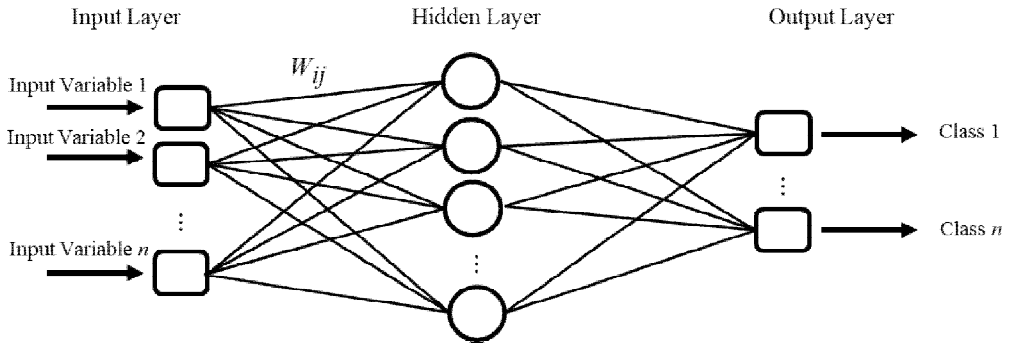


Fig.1. A three-layer neural network used for classification

The ANN technique can be used to solve prediction and classification problems based on a given data set $S = \{\bar{x}_t, \bar{y}_t | t = 1, \dots, n\}$, in which $\bar{x}_t$ represents an input vector and $\bar{y}_t$ represents the output vector at time t. The data set S is divided into two sets: training set $S_1 = \{\bar{x}_t, \bar{y}_t | t = 1, \dots, m\}$ and testing set $S_2 = \{\bar{x}_t, \bar{y}_t | t = m + 1, \dots, n\}$. Training data set S_1 is used to train or develop an ANN model by determining the weights of links (w_{ij}) so that a mapping function $\bar{\hat{y}}_t = f(\bar{x}_t)$ is formed. Testing data set S_2 is used to test the prediction capacity of $\bar{\hat{y}}_t = f(\bar{x}_t)$; for example, by a classification error rate defined as $\sum_{m+1}^{n} \frac{|\bar{\hat{y}}_t - \bar{y}_t|}{n-m}$.

An ANN model is developed and tested by a particular data set $S = \{\bar{x}_t, \bar{y}_t | t = 1, \dots, n\}$. Input vector $\bar{x}_t$ represented by a distinct set of *input features* forms a different ANN model. Now suppose we have *N* possible input features (say, *N* =15). Selecting three portfolios out of the *N* input features, we can develop three ANN models. Different ANN models have different prediction capacity.

This research addresses a binary classification problem. As shown in **Table 1**, the classification results can be of four categories: true positive (TP), false positive (FP), false negative (FN), and true negative (TN). TP and TN are of *accurate* predictions. In TP category, a "buy-day" is also predicted as a "buy-day" and TN denotes that a "not-buy-day" is predicted as a "not-buy-day". In contrast, FP and FN are of inaccurate prediction. In FP category, a "not-buy-day" is yet predicted as a "buy-day"; and FN denotes that a "buy-day" is predicted as a "not-buy-day".

To justify the capacity of an ANN model, the following four performance metrics shall be reported in the training stage as well as in the testing stage. The four metrics are defined as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

$$Specificity = \frac{TN}{TN+FP}$$

$$Sensitivity = \frac{TP}{TP+FN}$$

$$Precision = \frac{TP}{TP+FP}$$

Accuracy denotes the average classification capacity of the ANN model. *Specificity* measures the proportion of "not-buy-days" that are correctly identified as such. *Sensitivity* (*probability of buy-point detection*) is the proportion of "buy-days" that are correctly identified as such.*Precision* denotes the probability of being a "*real buy-day*" in each "*predicted buy-day*".

Table1: Performance metric

		Original data	
		buy-day	not-buy-day
Prediction	buy-day	True Positive (TP)	False Positive (FP)
	not-buy-day	False Negative (FN)	True Negative (TN)

III. INPUT FEATURES AND PORTFOLIOS

In this research, 15 variables (input features) are hypothesized to have an impact on the future trend of stock price. These variables are categorized into three groups: macroeconomics, industry, and stock trading factors.

The macroeconomics group involves four variables (X_1, X_2, X_3, and X_4). Variable X_4denotes the TAIEX Index closing price; the first three variables model the impacts of *money supply*. Money supply is an aggregate measure of money, which includes currency (notes or coins) and liquid instruments (deposits in banks). Money supply can be defined in different scope. For example, according to Central Bank of Taiwan, M1B is of narrow scope definition which includes currency and some bank deposits that can be converted easily to cash. M2 is of broad scope definition which includes M1B and certain deposits

in banks (e.g., Certificate Deposit) and funds that takes a longer time to be converted to cash.

X_1= M1B annualgrowth rate
$= \frac{(M1B \text{ in month } i \text{ of year } t) - (M1B \text{ in month } i \text{ of year } t-1)}{(M1B \text{ in month } i \text{ of year } t-1)}$

X_2=
M1B annualgrowth rate– M2 annual growth rate

X_3= M1B short-termgrowth rate
$= \frac{(M1B \text{ of month } i) - (\text{average } M1B \text{ of month } i-1,\ i-2,\ i-3)}{(\text{average } M1B \text{ of month } i-1,\ i-2,\ i-3)}$

X_4= TAIEX Index closing price

The industry group involves only one variable (X_5), which is the *stock price index of electronic industry* in Taiwan stock exchange. The electronic industry involves companies that are in the supply chain of electronic products. These companies may be the upstream/downstream providers of the addressed company (i.e., Foxconn). As a result, their stock price may have an impact on the stock price of Foxconn.

The stock trading factors involve 10 variables (X_6, …, X_{15}). Of the 10 variables, four ones are to model the stock closing prices of four consecutive days (X_6, …, X_9), in which X_6 = the closing price of day t; X_7 = the closing price of day *t*-1; X_8 = the closing price of day *t*-2; X_9 = the closing price of day *t*-3. Herein, day *t* denotes "today" and we attempt to justify if "tomorrow" (day *t*+1) is a "buy-day". Of the 10 variables, two ones (X_{10} and X_{11}) are to model the psychological line of the last *k* days. The psychological line of the last *k* days is (*k*-1)/*k*, in which *k*-1 denotes the number of days that are "up" against "yesterday" in terms of closing price. Herein, variable X_{10} denotes *k* = 12 and variable X_{11} denotes *k* = 24. In addition, variable X_{12} models the trading volume of day *t*; variable X_{13}models the monthly trading turnover rate (monthly trading shares/outstanding shares) at day *t*. Furthermore, variable X_{14}= (f_t-f_{t-1}) wheref_t models the balance of financing at day *t*; variable X_{15} = (s_t-s_{t-1}) where s_t models balance of securities loans at day *t*.

Out of the 15 input variables, we select three portfolios of input variables; and each portfolio is used to develop an ANNof the three portfolios (PF_1, PF_2, PF_3), PF_1 involves 13 variables, in which X_3 and X_{11} are excluded. Portfolio PF_2 involves 13 variables, in which X_3 and X_{10} are excluded. Portfolio PF_3 involves 14 variables, in which X_{11} is excluded.

IV. NUMERICAL EXPERIMENTS

A stock (Foxconn) listed in Taiwan Stock Exchange is used to test the three portfolios (PF_1, PF_2, and PF_3) and their corresponding ANNs (ANN_1, ANN_2, and ANN_3). The training and testing of these ANN are implemented by *R* programming language based on *neuralnet* package and carried out in the environment of Inter(R) Core(TM) CPU 3.20GHz and 12.0GB. Parameters predefined for the training of an ANN are listed in **Table 2**.

Table 2: Parameter Values

Parameters	Given Values
Seed (Pseudo Random Number)	123
Number of hidden layers	1
Number of nodes	numbers of PF_S' variables
Threshold	0.01
Learning rate	0.01
Stepmax	10^7

Data set for training and testing the ANNS ranges from Jan. 2007 to Dec. 2016 (10 years); 70% of the data set is used for training and the remaining 30% is for testing. **Table 3** shows the distributions of "buy-day" and "not-buy-day" in the data sets. Notice that the percentage of "buy-day" and that of "not-buy-day" are both close to 50% (i.e., data sets are quite "balance"). The four performance metrics (accuracy, specificity, sensitivity and precision) for the training stage and testing stage for three ANNs are shown in **Table 4**, which reveals that the performance of training stage is much better than that of testing stage. To integrate the four performance metrics into one, we propose a trading strategy in order to compute the rate of return in stock investment. The trading strategy is called *one batch trading policy*. Suppose we have a current account for stock investment, initially with a budget (say, *B*). Once a "predicted buy-day" appears, money in the current account is *wholly* devoted to buy the stock and the balance becomes zero. In the coming time horizon (100 days), whenever the stock rises over 10%, we sell all the stocks. If the stock price never rises over 10% in the coming time horizon, we sell all the stock at the end of the time horizon. Under the trading strategy, the annual rate of return of three ANNs can be computed as shown in **Table 5**. The table reveals that the annual rate of return of ANN_3 is 10.92%, substantially outperforming the other two ANN models.

Table 3: Percentage of each data set

Data set	Percentage of raw data	buy-day		not-buy-day	
		Numbers	Percentage	Numbers	Percentage
Raw Data	100%	1,160	48.82 %	1,213	51.18 %
Training	70%	824	49.61 %	837	50.39 %
Testing	30%	336	47.19 %	376	52.81 %

Table 4: Performance of three ANNs

Models	Number of Variables	Training				Testing			
		Accuracy	Specificity	Sensitivity	Precision	Accuracy	Specificity	Sensitivity	Precision
ANN_1	13 (PF_1)	97.11%	98.81%	95.39%	98.74%	37.22%	29.99%	46.43%	36.88%
ANN_2	13 (PF_2)	99.10%	99.76%	98.42%	99.75%	49.58%	81.38%	13.99%	40.17%
ANN_3	14 (PF_3)	97.23%	99.88%	94.54%	99.87%	57.44%	78.46%	33.93%	58.46%

Table 5: Annual Rate of return of three ANNs

Models	Annual Rate of Return				
	1st year	2nd year	3rd year	Sum	Average
ANN_1	30.11%	-12.58%	6.52%	24.05%	8.02%
ANN_2	10.03%	-9.36%	10.04%	10.71%	3.57%
ANN_3	20.03%	4.93%	7.80%	32.76%	10.92%

CONCLUSIONS

This paper presents a neural network approach to predict the "buy-day" for stock investment. Neural network and other data mining techniques have been widely used in predicting stock prices; yet most studies focused on the prediction of next trading day. In contrast, this research is distinguished in predicting whether a trading day is a "buy-day" by considering a future time horizon (say, 100 days). Namely, a trading day is a "buy-day" if the stock is expected to rise over 10% in the coming 100 days. We address 15 variables as possible input features as an ANN model. Out of the 15 variables, we select three portfolios of input features and develop three ANNs. Numerical experiments reveal that the best ANN can have an annual rate of return about 11%. These experiments imply that scope and selection of input features have a substantial impact on the rate of return and shall be further explored.

REFERENCES

[1] L. Xi, H. Muzhou, M. H. Lee, J. Li, D. Wei, H. Hai, and Y. Wu, "A new constructive neural network method for noise processing and its application on stock market prediction," Applied Soft Computing, vol. 15, pp. 57–66, February 2014.

[2] X. Gong, Y.W. Si, S. Fong, and R. P. Biuk-Aghai, "Financial time series pattern matching with extended ucr suite and support vector machine," Expert Systems with Applications, vol. 55, pp. 284–296, August 2016.

[3] J. Miao, P. Wang, and Z. Xu, "A bayesian dynamic stochastic general equilibrium model of stock market bubbles and business cycles," Quantitative Economics, vol. 6, pp. 599–635, November 2015.

[4] L. A. Teixeira, and A. L. I. De Oliveira, "A method for automatic stock trading combining technical analysis and nearest neighbor classification," Expert Systems with Applications, vol. 37, no. 10, pp. 6885–6890, October 2010.

[5] Z. Hu, J. Zhu, and K. Tse, "Stocks market prediction using support vector machine," IEEE. 6th International Conference on Information Management, Innovation Management and Industrial Engineering, pp. 115–118, 2, November 2013.

[6] H. F. Rahman, R. Sarker, and D. Essam, "A genetic algorithm for permutation flow shop scheduling under make to stock production system," Computers & Industrial Engineering, vol. 90, pp. 12–24, December 2015.

[7] W.C. Chiang, D. Enke, T. Wu, and R. Wang, "An adaptive stock index trading decision support system," Expert Systems with Applications, vol. 59, pp. 195–207, October 2016.

[8] L. A. Laboissiere, R. A. Fernandes, and G. G. Lage, "Maximum and minimum stock price forecasting of Brazilian power distribution companies based on artificial neural networks," Applied Soft Computing, vol. 35, pp. 66-74, 2015

[9] D. Yan, Q. Zhou, J. Wang, and N. Zhang, "Bayesian regularisation neural network based on artificial intelligence optimization," International Journal of Production Research, vol. 55, no. 8, pp. 2266-2287, 2017.

[10] N. Zhang, A. Lin, and P. Shang, "Multidimensional k-nearest neighbor model based on EEMD for financial time series forecasting," Physica A: Statistical Mechanics and its Applications, vol. 477, pp. 161-173, July 2017.

[11] M. Inthachot, V. Boonjing, and S. Intakosum, "Artificial neural network and genetic algorithm hybrid intelligence for predicting Thai stock price index trend," Hindawi Publishing Corporation Computational Intelligence and Neuroscience, vol. 2016, Article ID 3045254, 8 pages, 2016.

[12] C. Liu, J. Wang, D. Xiao, and Q. Liang, "Forecasting S&P 500 Stock Index Using Statistical Learning Models," Open Journal of Statistics, vol. 6, no. 6, pp. 1067, December 2016.

[13] E. Chong, C. Han, and F. C. Park, "Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies," Expert Systems with Applications, vol. 83, pp. 187-205, October 2017.

[14] N. Oliveira, P. Cortez, and N. Areal, "The impact of microblogging data for stock market prediction: using Twitter to predict returns, volatility, trading volume and survey sentiment indices," Expert Systems with Applications, vol. 73, pp. 125-144, May 2017.

[15] X. D. Zhang, A. Li, and R. Pan, "Stock trend prediction based on a new status box method and AdaBoost probabilistic support vector machine," Applied Soft Computing, vol. 49, pp. 385-398, December 2016.

[16] Y. Shynkevich, T. M. McGinnity, S. A. Coleman, A. Belatreche, and Y. Li, "Forecasting price movements using technical indicators: Investigating the impact of varying input window length," Neurocomputing, vol. 264, pp. 71-88, November 2017.

★★★