

---

# EPSO-GHSOM Stock Selecting and Trading Strategy on Big Data

Wenqing Liu<sup>1</sup>, Tingyu Chen<sup>1,\*</sup>, Mike Y. J. Lee<sup>1,2</sup>

<sup>1</sup>Department of Management Information Systems, National Chengchi University, Taipei, Taiwan

<sup>2</sup>Department of Business Administration, China University of Technology, Taipei, Taiwan

## Email address:

w\_liou@nccu.edu.tw (Wenqing Liu), timvickie2003@gmail.com (Tingyu Chen), yjlee@cute.edu.tw (M. Y. J. Lee)

\*Corresponding author

## To cite this article:

Wenqing Liu, Tingyu Chen, Mike Y. J. Lee, EPSO-GHSOM Stock Selecting and Trading Strategy on Big Data. *International Journal of Intelligent Information Systems*. Vol. 6, No. 2, 2017, pp. 7-20. doi: 10.11648/j.ijis.20170602.11

**Received:** June 30, 2015; **Accepted:** December 2, 2015; **Published:** March 25, 2017

---

**Abstract:** Extracting trading information from the stock market to construct accurate forecasting models that filter signals and noise is a challenge. This research employs big data analytics to construct a computation platform for stock selection and trading strategies. It adopts elite particle swarm optimization (EPSO) to elucidate optimal trading opportunities and combines growing hierarchical self-organizing map (GHSOM) and EPSO in its stock selection strategy. EPSO-GHSOM distinguishes companies' operating profitability, identifies price signals, and sets decision rules for buying and selling.

**Keywords:** Particle Swarm Optimization (PSO), Growing Hierarchical Self-Organizing Map (GHSOM), Big Data Analytics, Stock Trading Strategies, Stock Market Forecasting, Stock Market Predicting

---

## 1. Introduction

Predicting stock market trends has always been a difficult problem. Often prediction failures are attributable to uncontrollable factors affecting the market [1], including macro- and micro-economic environments, floating funds in the market, and investors' decision behavior. These factors influence each other through their intricate relationships, thereby increasing forecasting complexity. Therefore, investigators must understand the operating mechanism of the stock market before constructing models.

An invisible mechanism generates price volatility in equities markets. However, scholars dispute whether equities markets are efficient [2] and whether stock prices correspond immediately to information. Fama [3] believes equity prices are randomly generated and all investors have equal access to all information (Efficient Market Theory). Therefore, prices fluctuate randomly, there is no way to forecast them, and buy-and-hold is the best investment strategy [2]. Other scholars [4-5] insist that markets are inefficient and information is not universally available. These hypotheses about investment information suggest that history will repeat even if future investors face similar situations. These scholars believe that price trends exist and can be forecasted. Investors who share that conclusion use fundamental and technical

analysis to assess stock values. Previous research combines artificial intelligence and technical analysis in models to forecast market trends, trading rules, and buy-sell signals. Their methods included genetic programming [6], k-nearest neighbor algorithms [2], neural networks [7-13], genetic algorithms [14], and particle swarm optimization (PSO) [15] to construct trading models based on technical indicators.

This research simulates investors' decision-making processes while incorporating fundamental and technical data into a model that assures comprehensive forecast information. Technical data form the basis of trading strategies, such as variable length moving averages (VLMAs) [16], golden crossover and death crossover rules, and PSO [17], in model construction to identify criteria to determine optimal purchase and sale timing. Its stock selection strategy employs automatic clustering and categorizing features of growing hierarchical self-organizing map (GHSOM) [18] to build the model. Input variables include fundamental data (e.g., indicators of company profitability). Output variables are return on investment (ROI) derived by the trading strategy from source data. Training and experiments define criteria and patterns that identify stocks for investment and maximize trading profits. A large quantity of increasingly volatile data is involved in conducting a successful experiment. Therefore, we employ big data analytics and methods. Doing so increases

computational efficiency and performance, provides insights into investment strategies, and facilitates investment decisions.

## 2. Literature Review

### 2.1. Strategy in Stock Prediction

Extensive scholarship has investigated stock prediction

strategies using methods including artificial intelligence, data mining, and statistical models. Technical indicators and daily transaction data constitute the majority of variables drawn from sourced data, followed by stock indices, economic indicators, and financial indicators. Recent research employs semi-structured and unstructured data from the Internet and social media. Table 1 identifies research concerning prediction and trading strategies [2, 6-11, 14-24].

*Table 1. Stock trading strategies researches.*

Author	Method/Variable/Solution
(I). find good trading rules	
Kwon, Yung-Keun and Byung-Ro Moon (2007) [7]	(1). Method: NN, ensemble model, GA (2). Variable: Stock price, MA, RSI, MACD, KD (3). Solution: NN weights are transformed into ensemble chromosome and use GA to find best trading rules.
Hsu Ling-Yuan et al. (2011) [15]	(1). Method: PSO (2). Variable: SMA (3). Solution: PSO is combined with the golden and the death crossover rule to find best trading rules.
Rodríguez-González et al. (2011) [8]	(1). Method: NN (2). Variable: RSI (3). Solution: NN is combined with RSI decision rule to find best trading rule.
Chien, Ya-Wen Chang and Yen-Liang Chen (2010) [14]	(1). Method: GA (2). Variable: Stock price, volume, MA, KD (3). Solution: GA is combined with the association rules as chromosomes consisted of technical indices.
(II). identify selling and buying signals	
Ohkawa et al. (2008) [6]	(1). Method: GNPcn (GNP with control node) , simulation (2). Variable: candle stick (3). Solution: GNPcn identifies candlestick chart patterns to distinguish selling from buying signals.
Teixeira et al. (2010) [2]	(1). Method: KNN (2). Variable: SMA, RSI, KD, Bollinger band (3). Solution: Experts identifies selling and buying signals from technical indicators and KNN matched patterns.
Chavarnakul, Thira and David Enke (2009) [9]	(1). Method: NN, Fuzzy, GA (2). Variable: VAMA (3). Solution: NN is combined with VAMA to generate the price and volume of the future, through Fuzzy and GA which identify selling and buying signals.
Wang Lipo and Shekhar Gupta (2013) [10]	(1). Method: NN, Wavelet (2). Variable: Stock Index (3). Solution: NN is combined with gradual data to predict prices, compares predicted prices with closing prices today and generates selling and buying signals.
Wang, Jar-Long and Shu-Hui Chan (2007) [19]	(1). Method: Pattern Recognition (2). Variable: Stock index (3). Solution: Charting patterns of technical indicators are transformed by image processing values (0, 1, 0.5, -1) to identify selling and buying signals.
(III). identify trends pattern	
Peng, Hsin-Tsung, Hahn-Ming Lee, and Jan-Ming Ho. (2005) [20]	(1). Method: Trading Decision Machenism (TDM) (2). Variable: Taiwan Stock Exchange Capitalization Weighted Index (3). Solution: TDM system find change points.
Chang Pei-Chann, Chin-Yuan Fan and Chen-Hao Liu. (2009) [21]	(1). Method: Piecewise Linear Representation Method, Backpropagation neural network, GA (2). Variable: Stock price, RSI, MA, MACD, KD, William, volume (3). Solution: PLR divides segments of history data. BPN and GA are combined with MACD and KD to find patterns of change points.
Lin Nana et al. (2014) [22]	(1). Method: Piecewise Linear Representation Method, Backpropagation neural network, GA (2). Variable: Stock price, RSI, MA, MACD, KD, volume, Web news media sentiment indicators (3). Solution: PLR divides segments of history data. GMKL system matches patterns of change points.
Lin Xiaowei, Zehong Yang and Yixu Song (2011) [11]	(1). Method: Echo State Network (2). Variable: Stock price (3). Solution: Price series transform values into a range from -1 to 1 using Echo State Network to identify change points.
Matthew Butler and Dimitar Kazakov (2012) [23]	(1). Method: Six supervised learning algorithm(SVM, KNN, MLP, AIS, J48, Naive Bayes), GARCH process, PSO (2). Variable: Stock price, Bollinger band (3). Solution: Simulating the processes of GARCH time series model to identify trends of stock market.
(IV). select stock	
Pham Hai V. et al., 2014 [24]	(1). Method: Fuzzy, SOM (2). Variable: survey and experts evaluation (3). Solution: Fuzzy is combined with surveys and evaluation from experts to select and classify stocks with SOM.

## 2.2. PSO

PSO [18] is an inspired algorithm that simulates swarm intelligence. It has memory and is easily calculated to derive optimal solutions garnered through accelerated convergence. The next step to particles' decision-making behaviors is influenced by past experiences of both the local and global best solution, and thereafter, the strategies are evolved. This method is commonly used to solve optimization questions.

$$V_i^{n+1} = \omega \cdot V_i^n + c_1 \cdot r_{i1}^n \cdot (P_i^n - X_i^n) + c_2 \cdot r_{i2}^n \cdot (G_i^n - X_i^n) \quad (1)$$

$$X_i^{n+1} = X_i^n + V_i^{n+1} \quad (2)$$

Equations (1) and (2) define parameters for the PSO algorithm.  $X_i$  is the particle position and  $V_i$  its velocity.  $P_i$  is the local best solution.  $P_g$  is the global best solution.  $\omega$  is the inertia weight.  $c_1$  is the self-confidence coefficient for the particle and the influencing best solution coefficient of its individual experience.  $c_2$  is the social confidence coefficient for the particle and the influencing best solution coefficient for all particles.  $r_{i1}$  and  $r_{i2}$  are random number generators spanning [0, 1].  $n$  is the current period and  $n+1$  is the next period.

## 2.3. GHSOM

Kohonen's self-organizing map (SOM) [25] clustering algorithm maps high-dimensional data inputs onto low-dimensional output clusters using a neural network. However, its capability is limited. First, the number of nodes for topology must be pre-established. Second, the model has limited ability to represent similarities among data input under a distance concept rather than hierarchical (parents-child) relationships. To overcome these deficiencies, Dittenbach et al. [18] propose GHSOM to provide automatically a hierarchical structure for multiple layers wherein each comprises several independent SOMs (Fig. 1).

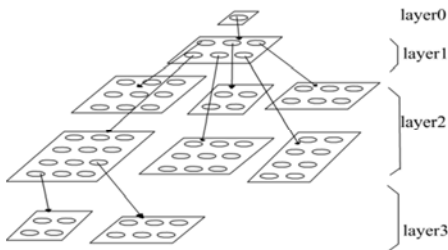


Figure 1. The GHSOM topology [18].

Training for the GHSOM algorithm involves four phases [18].

1. Initialize layer 0 and parameters.

Virtual layer 0 includes only one node whose weight vector is initialized as the average of all input data. Then the mean quantization error of layer 0 ( $MQE_0$ ) is calculated. The MQE of a node denotes the mean quantization error by Euclidean distance between the weight vector of the unit and all input vectors.

2. Train each SOM.

Input data imported individually are calculated as distances between the input data and the weighted vector of all nodes in the training process for individual SOMs. The clustering node with the shortest distance is chosen as the winner. The weight vectors of the winner and its neighborhood clustering nodes are adjusted under the competitive learning rule. The training process is repeated until the learning rate reduces to a specified value.

3. Grow each SOM horizontally.

Each SOM will grow until the mean value of the MQEs for all nodes in the current layer ( $MQE_m$ ) is less than the MQE of the previous layer ( $MQE_p$ ) multiplied by  $\tau_1$  (Equation (3)). If the stopping condition is not satisfied, find the error node that owns the largest MQE and insert one row or column of new nodes between the error node and its dissimilar neighbor.

$$MQE_m < \tau_1 \cdot MQE_p \quad (3)$$

4. Expand or terminate the hierarchical structure.

After the growing stage of individual SOMs, each  $MQE_i$  is compared with the value of  $MQE_0$  multiplied by  $\tau_2$ . Values greater than  $\tau_2 \times MQE_0$  generate a next layer of SOM. The hierarchy develops until all leaf nodes satisfy the stopping condition in Equation (4).

$$mqe_i < \tau_2 \cdot MQE_0 \quad (4)$$

## 2.4. Analysis of Stock Investment Decision

Assuming market trends and share prices are predictable, investors can use fundamental and technical analysis to value stocks.

### 2.4.1. Fundamental Analysis

Fama and French [26] argue that book-to-market ratio is a better predictor than other financial ratios of whether a stock could generate higher returns. Graham and Dodd [27] developed a value investing investment paradigm using indicators such as dividend yields and price-earnings ratios to assess stocks' intrinsic value. When applied with the price-to-book ratio (P/B share price/book value per share), investors determine whether a stock is undervalued.

Value investor Warren Buffett [28,29] followed Graham and Fisher [30] and improved upon Graham's theory by combining qualitative and quantitative analyses of four key considerations—a company's business and operating characteristics, management quality, financial soundness, and comparison between a stock's intrinsic and market values. Desirable businesses have high market shares, simple business models, stability, and profitability. Quality managers are capable and vigilant, are empowered to allocate capital, and are rational and honest. Criteria for financial soundness include return on equity (ROE), earnings per share (EPS), operating profits, and owner earnings.

Joseph D. Piotroski [31] and Mohanram [32] suggest F-scores and G-scores, respectively, as alternative methods and performance criteria for assessing stock values.

#### 2.4.2. Technical Analysis

Technical analysis indicators include momentum indicators (e.g., moving average crossover) and contrarian indicators (e.g., trend reversal indicators such as a relative strength index (RSI) and KD random indicators (stochastic)) [33].

The moving average crossover within momentum indicators is the most fundamental trend forecasting indicator. It generally uses number of days to calculate a stock's short-, mid-, or long-term price trend. It encompasses cost and support pressures (reference price point of support/pressure) and displays optimistic or pessimistic trends (bull or bear) and volatility. However, transmission of reversal signals is slow because the curve is smooth. The slower the pace at which the moving average crosses over, the less information is represented.

The moving average crossover depends on whether the number of days is stable. It is a fixed length moving average (FLMA); if the number of days varies, it presents a VLMA [16]. Equation (5) captures the simple moving average.

$$SMA_k^T = \frac{1}{T} \sum_{t=k-T+1}^k Price_t \quad (5)$$

$T$  is the time interval for the simple moving average.  $k$  is the  $k^{\text{th}}$  day now measured within all of days.  $Price_t$  is the closing price on previous day  $t$ . Day  $t$  ( $t=k-T+1$ ) covers actual trading days. Daily prices along the moving average line have equal weight; accordingly, research usually employs the moving average crossover line to model forecasting trends and make trading decisions.

#### 2.5. Big Data Analytics

There are several definitions of "big data." Gartner [34] defines it as "high volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making."

Zikopoulos *et al.* [35] say that big data are characterized by volume, velocity, and variety. Volume indicates quantities of data greater than that appearing in traditional settings. Velocity suggests that information is generated faster than that in traditional systems. Variety refers to multiple forms of emerging data that interest business enterprises. These V-words constitute the classic description for big data applications.

Zikopoulos *et al.* [36] introduce the additional concepts of veracity and value. Veracity refers to the accuracy, truthfulness, and reliability of data. Value refers to the potential for big data to provide cost-beneficial additions to an enterprise's technology portfolio. Ultimately, value is the key cost-beneficial criterion for determining whether big data analytics should be used. However, big data analytics also require infrastructure to support gathering, storing, processing, and using the accumulated information.

Big data analysis adopts a three-part framework: data accessing and computing, data privacy and domain knowledge, and a mining algorithm [37]. Data accessing and computing provide platforms capable of handling massive data for

computing and programming. Processing and computing platforms are divided into batch processing and stream processing architectures [38].

Data privacy and domain knowledge pertain to data falling within the domain of privacy issues, such as equity-related knowledge. Big Data Mining mainly employs artificial intelligence, simulation, data mining, and statistics and combines with domain knowledge using parallel computing to process voluminous and complex data.

Chen CL Philip and Chun-Yang Zhang [38] propose principles for designing a big data analysis system. The most important consideration is an architecture that supports multiple analytic methods and is adequate for selecting the appropriate platform that combines parallel computing, distributed storage, and blending capacity.

EMC proposes solutions [39] oriented toward application, analysis, management, and infrastructure layers. Planning is top-down, whereas construction is bottom-up. The application layer requires understanding the ecosystem of positions and how to solve the relevant connotations, objectives, and properties of a problem. The data analysis layer involves selecting and constructing platforms for data analysis and visualization. The management layer involves effective application of unified and automatic interfaces to manage all four layers. The infrastructure layer involves supporting big data applications and should consider automated scalable storage and computing.

### 3. The Proposed Methodology

Our big data analytics framework [37] focused on three areas: domain knowledge, data mining and analysis, and computational platforms. To answer questions about stock trading strategy, we incorporated domain knowledge, PSO, GHSOM, and other artificial intelligence tools into the analytics platform. We developed a mash-up strategy for selecting and trading to identify quality stocks and to determine optimal timing and opportunities for buying and selling. This research advances the implementation of big data analytics methods, adds insights into information processing, and improves decision-making.

We assume that the stock market is inefficient in that some investors can use informational asymmetry to predict market trends and to make trading decisions. We also assume that investors seek maximum trading profits and that the market allows unlimited trading volume.

This research puts forth the EPSO stock trading prediction model and the EPSO-GHSOM stock selection prediction model. Each requires phases of building, training, analyzing, testing, and adapting architecturally the big data analytics platform before being deployed phase-by-phase in modules within the platform's architecture. Fig. 2 displays the system architecture.

#### 3.1. Stock Trading Prediction Model – EPSO

To identify buy–sell signals, we propose improvements to the PSO algorithm that combine with the golden crossover and

death crossover rules of SMA in Equation (6) – (7). It calculates ROI as a fitness function to obtain the SMA crossover trading principle that maximizes profits on high and low trading spreads.

For example, we present short- and long-moving averages (SMA<sup>60</sup> and SMA<sup>200</sup>) and identify the golden crossover as the point at which SMA<sup>60</sup> crosses SMA<sup>200</sup> from below. The golden crossover presents a buy signal. Similarly, the death crossover—SMA<sup>60</sup> crosses SMA<sup>200</sup> from above—as a sell signal. Therefore, our trading decision originates in the golden crossover and death crossover. When circumstances conform to these conditions, buy–sell decisions are triggered immediately. Transaction costs are considered and profitability calculated at the end, so a comparison of the forecasting models is possible:

$$ROI = \frac{GainFromInvestment - CostOfInvestment}{CostOfInvestment} \quad (6)$$

$$signal_k = \begin{cases} \text{buy, if } (SMA_k^{T_1} > SMA_k^{T_2}) \text{ and } (SMA_k^{T_3} > SMA_k^{T_4}) \\ \text{sell, if } (SMA_k^{T_5} < SMA_k^{T_6}) \text{ and } (SMA_k^{T_7} < SMA_k^{T_8}) \\ \text{hold, otherwise} \end{cases} \quad (7)$$

In Equation (7), T<sub>1</sub>, T<sub>2</sub>, T<sub>3</sub>, T<sub>4</sub>, T<sub>5</sub>, T<sub>6</sub>, T<sub>7</sub>, and T<sub>8</sub> represent parameters of days of SMA. If (SMA<sup>T<sub>1</sub></sup> > SMA<sup>T<sub>2</sub></sup>) and (SMA<sup>T<sub>3</sub></sup> > SMA<sup>T<sub>4</sub></sup>) conditions are satisfied and then EPSO will generate a buying signal. T<sub>1</sub>, T<sub>2</sub>, T<sub>3</sub> and T<sub>4</sub> parameters are buying solutions for the EPSO algorithm. If (SMA<sup>T<sub>5</sub></sup> < SMA<sup>T<sub>6</sub></sup>) and (SMA<sup>T<sub>7</sub></sup> < SMA<sup>T<sub>8</sub></sup>) conditions are satisfied and then EPSO will generate a selling signal. T<sub>5</sub>, T<sub>6</sub>, T<sub>7</sub> and T<sub>8</sub> parameters are selling solutions for the EPSO algorithm.

We divide the buying solutions (T<sub>1</sub>, T<sub>2</sub>, T<sub>3</sub>, T<sub>4</sub>) to generate two subsets of the solutions, ((T<sub>1</sub>, T<sub>2</sub>) and (T<sub>3</sub>, T<sub>4</sub>)). We divide the selling solutions (T<sub>5</sub>, T<sub>6</sub>, T<sub>7</sub>, T<sub>8</sub>) to generate two subsets of the solutions, ((T<sub>5</sub>, T<sub>6</sub>) and (T<sub>7</sub>, T<sub>8</sub>)). There are two subsets of the buying solutions, ((T<sub>1</sub>, T<sub>2</sub>) and (T<sub>3</sub>, T<sub>4</sub>)) and two subsets of the selling solutions, ((T<sub>5</sub>, T<sub>6</sub>) and (T<sub>7</sub>, T<sub>8</sub>)). Four subsets feature two SMA parameters and are comparable. Purchases are indicated when (SMA<sup>T<sub>1</sub></sup> > SMA<sup>T<sub>2</sub></sup>) and (SMA<sup>T<sub>3</sub></sup> > SMA<sup>T<sub>4</sub></sup>). Selling is indicated when (SMA<sup>T<sub>5</sub></sup> < SMA<sup>T<sub>6</sub></sup>) and (SMA<sup>T<sub>7</sub></sup> < SMA<sup>T<sub>8</sub></sup>).

We propose the EPSO algorithm to improve the PSO algorithm [41]. After each stock runs  $\gamma$  experiments, they will get  $\gamma$  global best solutions including the qualities of good trading strategies. Each solution can be divided into  $2\gamma$  trading strategy subsets, denoted as (T<sub>i</sub>, T<sub>j</sub>). Subsets featuring sound trading strategies are permuted and combined, and we derive the best  $(2\gamma)^4$  solutions for each subset as candidates for elite solutions. These elite solutions are calculated by a fitness function and the highest value selected as the best trading strategy.

In addition, the V<sub>IT</sub> and X<sub>IT</sub> of MMPSO [15,41] are trapped within the boundary value so that the optimal solution is passivated. This EPSO algorithm will modify the boundary value rule of V<sub>IT</sub> and X<sub>IT</sub>, as shown in Equation (8) – (11). EPSO-related algorithms are as shown in Table 2.

$$V_{IT}^{n+1} = \omega \cdot V_{IT}^n + c_1 \cdot r_{i1}^n \cdot (P_{IT}^n - X_{IT}^n) + c_2 \cdot r_{i2}^n \cdot (G_{IT}^n - X_{IT}^n) \quad (8)$$

$$X_{IT}^{n+1} = X_{IT}^n + V_{IT}^{n+1} \quad (9)$$

$$V_{IT} = \begin{cases} rand[V_{max}, -V_{max}], & \text{if } V_{IT} > V_{max} \\ rand[V_{max}, -V_{max}], & \text{if } V_{IT} < -V_{max} \\ -V_{max} + 2 \times V_{max} \times rand(), & \text{if } |V_{IT}| > V_s \\ V_{IT}, & \text{otherwise} \end{cases} \quad (10)$$

$$X_{IT} = \begin{cases} rand[X_{max}, -X_{max}], & \text{if } X_{IT} > X_{max} \\ rand[X_{max}, -X_{max}], & \text{if } X_{IT} < -X_{max} \\ X_{IT}, & \text{otherwise} \end{cases} \quad (11)$$

### 3.2. Stock Selecting Prediction Model - EPSO-GHSOM

To identify low-risk quality stocks, we focus on a company’s long-term operating ability and financial soundness. We begin with fundamental analysis. It is measured and differentiated according to whether the company was financially healthy, and determines the company’s profitability and evaluates the stock’s intrinsic value and potential.

In selecting a method for cluster analysis, we adapted GHSOM to solve the grouping problem and to build the predictive model. GHSOM aids in determining cluster patterns through averages and variances in data, identifying self-organized groups and constituent data and highlighting data characteristics within clusters. We extract information, patterns, and features related to financial fundamentals from the model. We then use them to differentiate qualities of stocks and to aid decisions intended to reduce risks, increase safety, and maximize profit. The steps are as follows.

1. Define input to GHSOM (X). For the grouping variables, we chose as inputs the fundamental financial indicators based on value investing [28-29]. These variables were derived from literature reviews, making their metrics representative. They helped in differentiating stocks’ quality characteristics, reducing numbers of input variables, and grouping data appropriately and meaningfully. The financial indicators of value investing are as follows.

- 1) EPS = Net profits / number of shares.
- 2) ROE = EPS / Equity.
- 3) Net profits = operating profits + net profit (loss).
- 4) Dividend rate = EPS / Dividends.
- 5) Earnings reinvestment rate [41].
- 6) = the 4<sup>th</sup> year Fixed Assets + the 4<sup>th</sup> year Long-term Investment – the 0<sup>th</sup> year Fixed Assets + the 0<sup>th</sup> year Long-term Investment) / (4<sup>th</sup> year Net Profit + the 3<sup>rd</sup> year Net Profit + 2<sup>nd</sup> year Net Profit + 1<sup>st</sup> year Net Profit).
- 7) Shareholding ratio of directors and officers.

2. Define output from GHSOM (Y). The output variables are suitable indicators of the groups formed after the data have been categorized and segmented. Since the objective is to earn stock returns, a company’s financial soundness and profitability would influence investing profits. Therefore, we chose ROI as the output. However, ROI could be affected by price differences attributable to timing, bid-ask spreads, and investors’ behaviors when deciding to buy or sell. This uncertainty raised questions about assessing objectivity when using ROI. Therefore, we used ROI processed through the EPSO stock trading prediction model as the significant output variable. These ROI values are

more objective as they are calculated based on the same criteria and method as the trading strategy.

3. Define breadth ( $\tau_1$ ) and depth ( $\tau_2$ ) parameters for controlling the boundary of clustering experiments in GHSOM.

4. Perform cluster grouping and establish clustering criteria. Because of its competitive training feature, GHSOM could detect regularities and statistically significant characteristics within clusters, and its topology expressions could capture each cluster’s most salient features. Further, by using a small set of leaf nodes, GHSOM’s training process could categorize data in hierarchical relationships rather than merely dichotomous outcomes. It has a refined ability to dissect and classify uncovered hidden financial characteristics within each group, what they represent, and how they relate to ROI. This allowed observing whether the selected input variables are sufficiently discriminable and adequate for evaluating a firm’s profitability.

### 3.3. Big Data Analytics Platform Construction Phase

The huge volume of trading data needing analysis required

an extremely efficient system architecture to support long-term mining and decision-making. We constructed our big data analytics platform using the framework in Wu *et al.* [37] and solutions from EMC [39], including the application, analysis, management, and foundation layers. Since characteristics of the data being analyzed were structured, we selected NoSQL as the method for data storage. As there was no requirement for real-time return of results, we used batch processing to process and return data. Because we used PSO and GHSOM algorithms, the analytics processes required transformation procedures where data were interlocked between iterations, making them unsuitable for distributed parallel processing. Thus, we performed serial data processing for the algorithms.

We adopted the Apache Hadoop open source framework as our software and HBase for data management and storage. We completed the development of the analytic routines using Java and serial processing. A high-performance big data computation platform with independent parallel processing capability was necessary for completing solutions.

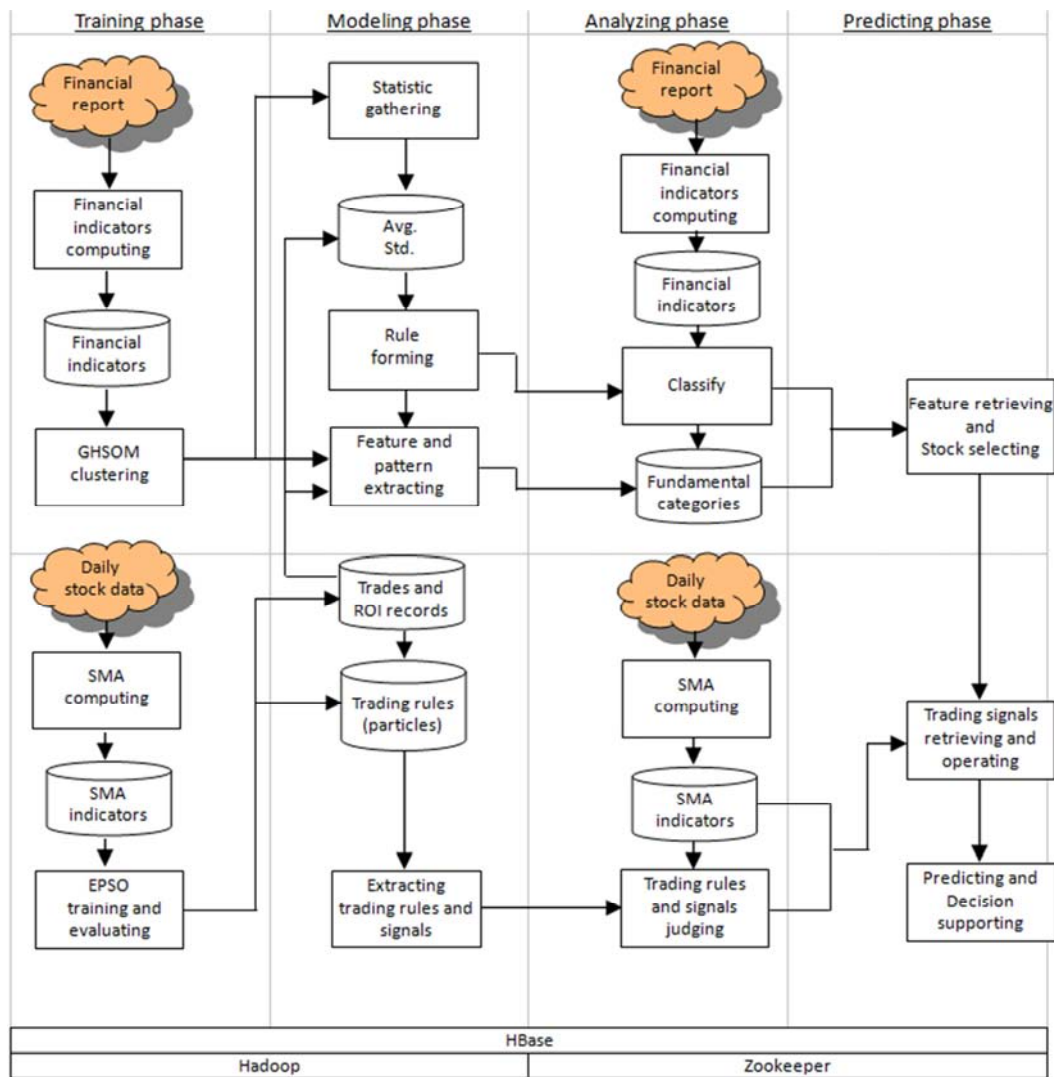


Figure 2. System architecture of the proposed approach.

**Table 2.** EPSO algorithm.

(1) EPSO main algorithm.

0	Perform $\gamma$ test runs.
0-1	Initial elite candidate queue, the queue can store $\gamma$ global best solutions.
1	Initial step:
	Give $m$ particles, $m$ is the swarm's size. $X_i$ express one of the particles.
	Each particle has 8 attributes. Each attributes has two vectors (position, velocity).
1-1	$X_{iT}$ express one of the attributes' position and the time interval of SMA.
	$V_{iT}$ express one of the attributes' velocity.
1-2	For each particle $X_{iT}$ , $1 \leq i \leq m$ , $1 \leq T \leq 8$ do.
	Initialize the particle's position $X_{iT}$ with a uniformly distributed random vector.
1-3	$X_{iT} \sim U(b_{lo}, b_{up})$ , where $b_{lo}$ and $b_{up}$ are the lower and upper boundaries of the searching space.
1-4	Initialize the particle's best known position to its initial position: $P_{iT} \leftarrow X_{iT}$
1-5	If ( $f(P_{iT}) < f(G_T)$ ) update the swarm's best known position: $G_T \leftarrow P_{iT}$ , $f(G_T)$ express calculating fitness function (ROI) of $P_i$ , $G$ according fitness function algorithm.
	Initialize the particle's velocity $V_{iT}$ with a uniformly distributed random vector.
1-6	$V_{iT} \sim U(b_{lo}, b_{up})$ , where $b_{lo}$ and $b_{up}$ are the lower and upper boundaries of the moving space.
2	Main Step:
	While the stop condition is not satisfied do.
2-1	The stop condition express the optimal solution is found or the maximal moving steps is reached.
2-2	For each particle $X_{iT}$ , $1 \leq i \leq m$ , $1 \leq T \leq 8$ do.
2-3	Pick random numbers: $r_{i1}, r_{i2} \sim U(0, 1)$ .
2-4	Update the particle's velocity according Eq.(8) and (10): $V_{iT} \leftarrow \omega V_{iT} + c_1 r_{i1} (P_{iT} - X_{iT}) + c_2 r_{i2} (G_T - X_{iT})$
2-5	Update the particle's position according Eq.(9) and (11): $X_{iT} \leftarrow X_{iT} + V_{iT}$
2-6	If ( $f(X_{iT}) < f(P_{iT})$ ) do.
2-7	Update the particle's best known position: $P_{iT} \leftarrow X_{iT}$
2-8	If ( $f(P_{iT}) < f(G_T)$ ) do.
2-9	Update the swarm's global best position: $G_T \leftarrow P_{iT}$
3	Elite Step:
3-1	Save the global best position of this elite candidates queue : $\Phi_{RT} \leftarrow G_T$ , $1 \leq R \leq \gamma$ , $1 \leq T \leq 8$ .
3-2	While all runs ( $\gamma$ times) have been executed then do.
	For each run $\Phi_{RT}$ , $1 \leq R \leq \gamma$ do.
	Each $\Phi_{RT}$ is one strategy solution and has 8 the time intervals of SMA ( $1 \leq T \leq 8$ ).
	Divided $\Phi_{RT}$ into 4 pairs sub-strategies - ( $\Phi_{R1}, \Phi_{R2}$ ), ( $\Phi_{R3}, \Phi_{R4}$ ), ( $\Phi_{R5}, \Phi_{R6}$ ), ( $\Phi_{R7}, \Phi_{R8}$ ).
	( $\Phi_{R1}, \Phi_{R2}$ ), ( $\Phi_{R3}, \Phi_{R4}$ ) are sub-strategies for buy signal and ( $\Phi_{R5}, \Phi_{R6}$ ), ( $\Phi_{R7}, \Phi_{R8}$ ) are sub-strategies for sell signal.
3-3	( $\Phi_{R1}, \Phi_{R2}$ ) express $E_{R1}$ sub-strategy. ( $\Phi_{R3}, \Phi_{R4}$ ) express $E_{R2}$ sub-strategy. ( $\Phi_{R5}, \Phi_{R6}$ ) express $E_{R3}$ sub-strategy. ( $\Phi_{R7}, \Phi_{R8}$ ) express $E_{R4}$ sub-strategy.
	Then we get one set of sub-strategies ( $E_{R1}, E_{R2}, E_{R3}, E_{R4}$ ).
3-4	All buy and sell signal sub-strategies of pairs in all runs for elite candidates have permuted and combined: $EC \leftarrow (E_{R1}, E_{R2}, E_{R3}, E_{R4})$
3-5	Find the best solution $E$ from the maximum fitness value $f(E_C)$ in all $E_C$ according fitness function algorithm: $E \leftarrow E_C$

(2) EPSO fitness function algorithm (calculate ROI).

1	Initial step:
1-1	Accept the particle $X_{iT}$ for main algorithm of EPSO as 8 time intervals of SMA, $1 \leq i \leq m$ , $1 \leq T \leq 8$ do.
1-2	Initialize holding capital=NT\$1,000,000 and sheet of hold stock=0.
2	Main Step:
2-1	For (d = start date of training/testing ; d ≤ end date of training/testing ; d++)
2-2	Calculate SMA of the particle $X_{iT}$ on d date.
2-3	Evaluate trading strategy signal on d date according Eq.(6).
2-4	If (signal on d date == buy) then Buy stock , and sheets of hold stock += holding capital / (close price of this d date of this stock + transaction cost).
2-5	If (this d date == end date of training/testing) then Sell stock, and holding capital += (sheet of hold stock * close price of this d date of this stock) - transaction cost.
2-6	return holding capital.

## 4. Result Analysis and Discussion

### 4.1. Experimental Backgrounds and Criteria

We selected 28 categories of stocks from Taiwan's stock market. From each category, we selected three stocks with similar prices as the stock of lowest, the medium, the highest in that category. Closing prices were on November 14, 2014,

and each stock selected from each category was listed before 2008. To assure the completion of the experiment, we divided the calculation of the EPSO algorithm for each stock into a training period and a testing period. The training period ran from the date the stock was listed until December 31, 2011. For each run, 1,000 times were executed and 10 runs were conducted. After all runs were completed during the training period, we chose the best combination of evolutionary decisions for SMA parameter day testing to calculate  $ROI_{EPSO}$ .

That done, we compared  $ROI_{EPSO}$  and the buy-and-hold strategy's  $ROI_{BH}$ . Other parameters are set to produce 50 particles ( $n = 50$ ). Particle  $X_{i,T}$  is SMA's parameter day, and the range is 6 to 400 days.  $V_{i,T}$  is the direction of movement and the velocity of a particle, and the range is  $-80$  to  $80$ .  $1 \leq i \leq n$ ,  $1 \leq T \leq 8$ .  $c_1$  is the confidence coefficient ( $c_1 = 2$ ).  $c_2$  is the social confidence coefficient ( $c_2 = 2$ ).  $w$  is the inertia weight, initially set to 1.4. As the number of runs increased,  $w$  gradually declined to 0.4. The initial capital is NT\$1 million, and at least 50 stocks are purchased and transacted at each step. We deducted transaction costs for each transaction. Total transaction cost for buying is 1.425‰ of the transacted amount, and for selling is 4.425‰ of the transacted amount. Considering transaction costs, 80% of available cash was used to purchase stocks.

The EPSO-GHSOM algorithm's parameters are as follows:  $\tau_1$  is between 0.1 to 0.005, and  $\tau_2$  is between 0.01 to 0.001. The sources of EPSO-GHSOM training are the same as EPSO testing stocks and are reported by the Taiwan Stock Exchange. We sourced secondary data from investment consulting companies.

#### 4.2. Stock Trading Prediction Strategy

##### 1. EPSO ROI performance.

Among the 84 stocks in the experiment, only one reported negative ROI. ROI exceeded 50% for 54 stocks (64.29%), 100% for 31 stocks (36.90%), and 200% for eight stocks (9.52%). During the period examined, TSEC weighted index (TPE:TAIEX) rose 29.2% from 6952 on January 1, 2012, to 8982 on November 14, 2014. EPSO accurately estimated ROI for 69 stocks (82.14%) identified as outperform. Therefore, the result of the performance of the methodologies used in this research was good and then the stock market was upswing to increase ROI (Table 3).

Table 3. ROI profits during the EPSO test period.

$ROI_{EPSO}$	Num. of shares	Cumulative Num.	Cumulative percentage
$ROI_{EPSO} \geq 300\%$	3	3	3.57%
$200\% \leq ROI_{EPSO} < 300\%$	5	8	9.52%
$100\% \leq ROI_{EPSO} < 200\%$	23	31	36.90%
$50\% \leq ROI_{EPSO} < 100\%$	23	54	64.29%
$0\% \leq ROI_{EPSO} < 50\%$	29	83	98.81%
$ROI_{EPSO} < 0\%$	1	84	100.00%

We picked the best EPSO's ROI performance-Inotera and Merry companies to discuss. When Stock Code 3474(Inotera) and 2439 (Merry) executed EPSO's ROI performance, it found that their performance were better than all and highest at 1,548.07% and 495.64%. Several possibilities can explain the difference in performance. First, these were high-priced stocks, and percentage swings represented extreme price differences. The highest price for each stock during the period examined was two to four times greater than the lowest price. Second, generally stock prices were rising.

We picked the worst EPSO's ROI performance- Formosa Petrochemical and Tsann Kuen companies to discuss. When Stock Code 6505 (Formosa Petrochemical) and 2430 (Tsann

Kuen) executed EPSO's ROI performance, we found that their performance was worse than all and lowest at 5.27% and  $-40.59\%$ . Several reasons underlie the biggest performance difference. First, these also were high-priced stocks for which percentage swings represented extreme price differences. The lowest price for each stock during the examined period was 86% of the highest price. Second, stock prices were trending downward.

##### 2. Comparison of buy-and-hold strategies.

After comparing the buy-and-hold strategy and ROI performance of EPSO (Table 4), among the 84 stocks, only one stock with the buy-and-hold strategy performed slightly better than EPSO. Among the EPSO stocks that out-performed the buy-and-hold strategy, 33 exhibited ROI exceeding 50%. Performance of individual stocks might differ because of prospects for their respective industries, operational excellence of their companies, and investors' speculative behavior. Thus, selection of stocks to generate higher profit was of paramount importance. Our experimental results suggested that Taiwan's stock market was inefficient during the examined period, allowing us to predict price trends.

Table 4. ROI of comparison between EPSO and Buy&Hold.

$ROI_{EPSO}-ROI_{BH}$	Num. of shares	Cumulative Num.	Cumulative percentage
$ROI_{EPSO}-ROI_{BH} \geq 100\%$	6	6	7.14%
$50\% \leq ROI_{EPSO}-ROI_{BH} < 100\%$	27	33	39.29%
$0\% \leq ROI_{EPSO}-ROI_{BH} < 50\%$	49	82	97.62%
$ROI_{EPSO}-ROI_{BH} < 0\%$	2	84	100.00%

##### 3. Golden crossover and death crossover rule.

The EPSO algorithm is intended to solve a set of 8 days of SMA parameters, the study referred to a group of buying and selling decision-making solution ( $T_1, T_2, T_3, T_4, T_5, T_6, T_7, T_8$ ). This solution can be divided into two subsets indicating buying ( $T_1, T_2$ ) and ( $T_3, T_4$ ) and two indicating selling ( $T_5, T_6$ ) and ( $T_7, T_8$ ). There are 84 stocks and 336 combinations of subsets for buying and selling. Two pairs of subset strategies emerged:  $T_1, T_3, T_5$ , and  $T_7$  as the first parameter and  $T_2, T_4, T_6$ , and  $T_8$  as the second.

The experiment revealed five decision patterns (A, B, C, D, and E). A, B, and C follow the golden and death crossover decision rules. In Pattern A, the first parameter is the number of days of short-term and the second the number of days of mid-term for SMA. In Pattern B, the first parameter is the number of days of short-term and the second the number of days of long-term for SMA. In Pattern C, the first parameter is the number of days of mid-term and the second the number of days of long-term for SMA. Pattern D features two parameters as the number of long-term, mid-term, short-term, compared with no decision-making significance. Pattern E features two parameters that do not meet the golden crossover and death crossover rules.

The experimental results indicate that of the 336 combinations, 10% conform to the principle, and of these 60% combinations, 38% displayed meaningless decision criteria, but still are able to attain reasonably attractive ROI. The



golden crossover and death crossover decision rules and EPSO's ROI performance have no significant relationship.

According to the observation of this research from overall subset solution strategies of SMA, the long term (exceeding 120 days) is approximately 76%. The mid-term (60 to 120

days) is approximately 20%. The short term (fewer than 60 days) is approximately 4%. The reason may be that SMAs are lagging suitable indicators for detecting long-term trends (Table 5).

*Table 5. EPSO subsets for buying and selling rules.*

Rule Type	subsets for buying 1	subsets for buying 2	subsets for selling 1	subsets for selling 2	Num. of subsets	Proportion
A	1	0	0	0	1	0.3%
B	0	2	3	2	7	2.08%
C	4	5	7	11	27	8.04%
D	21	26	27	24	98	29.17%
E	58	51	47	47	203	60.42%

An examination of poorly performing stocks indicates that the SMA crossover rule did not apply to declining markets. The reason might be that stock prices generally trended downward during the examined period, and our sampled stocks might have been overvalued with respect to their issuing companies' fundamentals. Further methods are needed to identify this kind of stock.

#### 4. EPSO algorithm accuracy.

The correct prediction of EPSO stock trading is that before the second sale price was greater than or equal to the previous purchase price, otherwise is the wrong prediction. Among the aggregated 84 stocks and 282 times for selling price were greater than buying price, a predictive accuracy of 83.69%. Predictive accuracy was greater when the SMA had a positive slope and less when its slope was negative. Perhaps our method is more applicable to the former case, or perhaps Taiwan's rising stock market aided its predictive accuracy.

### 4.3. Stock Selecting Prediction Strategy

#### 1. Discussion on the patterns after clustering.

We employed indicators from the value investing paradigm as input variables in the GHSOM stock-selection strategy and ROI from EPSO as the output. The experiment found that clustering using  $\tau_1 = 0.065$  and  $\tau_2 = 0.003$  produced the best results. As Table 6 shows, 16 groups were formed.

In Table 6, the overall observation is that the relationships between average ROI and the average slope of the net profit were most telling. When the average slope of the net profit was positive, average ROI of the cluster was good. When it was negative, average ROI of the cluster was likely to be worse. The majority of EPSS and ROES average slopes influence cluster average ROI, but some do not.

We also observed that the bigger the standard deviation in the population's property values, the greater is the degree of dispersion and the easier it can be used to differentiate. For example, the standard deviation in the net profit of 220.08 ranks highest among all input values, and its ability to differentiate was highest. Therefore, knowing the degree of dispersion in input values affects clustering criteria in GHSOM.

Since financial data were ordered by year, after GHSOM calculated the groupings, it was able to reveal trends and changes in financial indicators, differences in levels, and other features. The 16 groups derived from the experiment were classified into four categories:

(1) Net profit and EPS of groups in this category showed rising trends over the years, and the slopes were positive. These groups represented "star growth stocks" and were the highest in the recommendation list. Groups in this category included C-7, C-8, C-9, and D-11. The financial indicators, slopes, and graphs for Largan Precision (3008) and Highwealth Construction (2542) are as shown in Table 7-(1), 7-(2) and Figure 3-(1), 3-(2).

(2) Net profit and EPS of groups in this category decline over the years, and the slopes are negative. These groups represented stocks of companies operating poorly. These "landmine stocks" rank lowest on the list of recommendations. Groups in this category included B-4, B-5, B-6, and C-10. Financial indicators, slopes, and graphs for Formosa Plastics (1301) and Formosa Chemicals & Fibers (1326) are as shown in Table 7-(3), 7-(4) and Figure 3-(3), 3-(4).

(3) Net profit and EPS of these groups display U-shaped or W-shaped patterns with a mix of positive and negative slopes. After applying filters to remove groups with negative slopes, groups with U-shaped or W-shaped net profit and EPS show an overall positive aspect, representing companies with stable operating profitability. These stocks would stand second-highest on the list of recommendations. Groups in this category included A-1, A-2, B-3, E-12, F-13, F-14, F-15, and F-16. Financial indicators, slopes, and graphs for Merry (2439) and Grape King Bio (1707) are as shown in Table 7-(5), 7-(6) and Figure 3-(5), 3-(6). The remaining groups with U-shaped or W-shaped net profit and EPS display a negative trend; these were companies with deteriorating profitability. These stocks would be the next-highest (i.e., second-lowest) on the recommendation list. Groups in this category included A-1, A-2, B-3, E-12, F-13, F-14, F-15, and F-16. Financial indicators, slopes, and graphs for Ambassador Hotel (2704) and Dah San (1615) are as shown in Table 7-(7), 7-(8) and Figure 3-(7), 3-(8).

(4) Special cases: These groups of companies display negative net profit and EPS trends, but ROI values from EPSO indicate that their stocks had performed well. That finding was considered abnormal. Cases with the biggest anomalies are as shown in Table 7-(9), 7-(10) and Figure 3-(9), 3-(10).

Taroko Textile (1432) had negative net profit and EPS slopes, indicating poor financial fundamentals. However, its ROI of 277.23% is remarkably high. A likely explanation is that the company's stock price had been lingering in a trough. Recent

news about it being a concept stock linked to a political candidate and development of the entertainment market across the Taiwan Strait had boosted the stock price.

Hey-Song Corporation (1234) exhibits negatively sloping net profit and EPS, whereas its ROI is 110.23%. The negative slopes are attributable to a net operating loss of NT\$519 million in the second quarter of 2013. The positive ROI performance could be due to the resolution of a dispute among the company's board.

2. Discussion of differentiation capability using financial indicators.

In terms of the discriminant abilities among input variables in the GHSOM calculation, the highest was net profit, followed by EPS, ROE, and other financial indicators.

The order within groups A1–F16 is based primarily on net profit from highest to lowest, and then on EPS (Table 6). ROE exhibits discriminant influences on most stocks. Remaining indicators such as reinvestment rate, interest rate, and ratio of inside shareholders showed no significant discriminant effect.

3. Discussion of characteristics and models.

The first run of the experiment clarified that the orientations of the net profit, ROE%, and EPS slopes for most examined stocks had significant correlations with the EPSO ROI. As financial data were arranged sequentially by year, with its clustering criteria, GHSOM would first evaluate stocks based on their five-year financial trend. There were three patterns: straight up, straight down, and U-shaped or W-shaped. The latter, referred as “mixed groups,” display positive and negative values for net profit, ROE%, and EPS slopes of the stocks. Therefore, alongside the initial GHSOM criteria from the first run of the experiment, we added a new rule that when a mixed group was encountered, a filter was applied to further differentiate its pattern. This was aimed to identify the stocks' financial quality more accurately, to reduce trading risks, and improve investment profitability. The filter rules are as follows: if  $(slope(ROE) \geq 0) \ \&\& \ (slope(EPSS) \geq 0) \ \&\& \ (slope(Net\ Profits) \geq 0)$  then {safe stock};

Table 6. Financial indicators means and slopes of GHSOM leaf nodes.

group	Node no.	Num	Avg of ROE %	Avg of Return investment %	Avg of Net profits \$b	Avg of Divi - dends %	Avg of EPSS	Avg of Shares for director / supervisor	Avg of ROI <sub>PSO</sub>	Slope of Avg. ROE%	Slope of Avg Return investment %	Slope of Avg Net profits \$b	Slope of Avg Divi - dends %	Slope of Avg EPSS
A	1	1	26.97	58.61	1748.80	56.35	6.75	0.07	0.80	-0.01	-0.03	103.37	0.01	0.03
A	2	1	16.39	47.52	945.11	19.34	7.86	0.13	0.57	-0.01	0.11	181.58	-0.09	0.70
B	3	2	14.44	-8.18	252.49	82.77	11.21	0.44	0.59	-0.01	0.08	10.05	-0.00	0.29
B	4	2	10.90	17.12	264.56	72.16	4.43	0.25	0.27	-0.03	0.07	-66.88	-0.10	-1.17
B	5	1	7.30	187.77	192.31	85.95	1.32	0.21	0.23	-0.02	0.14	-41.31	0.04	-0.35
B	6	1	30.53	14.92	225.36	46.46	26.82	0.12	0.16	-0.21	0.01	-148.37	-0.25	-17.82
C	7	2	36.77	260.58	198.79	73.48	10.87	0.08	0.67	0.16	-5.78	110.56	-0.22	6.89
C	8	1	11.84	92.11	99.63	0.00	1.82	0.37	15.48	0.12	0.60	79.09	0.00	2.62
C	9	3	22.43	53.18	136.19	52.47	8.72	0.10	1.26	0.00	-0.03	21.17	0.00	1.27
C	10	1	17.04	7.27	197.50	72.65	5.13	0.29	0.34	-0.01	-0.09	-0.71	0.02	-0.03
D	11	8	23.79	147.58	85.04	62.54	16.13	0.27	1.41	0.01	-1.21	12.00	-0.01	3.48
E	12	13	15.36	23.04	23.39	64.28	4.11	0.24	0.73	-0.02	0.20	-2.62	0.04	-0.21
F	13	4	26.14	21.98	4.81	64.13	4.61	0.26	0.94	-0.05	0.02	0.43	0.04	-0.18
F	14	5	12.36	34.14	4.11	69.33	2.86	0.26	0.64	-0.01	0.10	0.06	0.05	-0.16
F	15	23	8.94	60.32	2.98	74.79	1.57	0.23	1.14	0.09	0.48	1.06	0.11	0.62
F	16	16	15.76	87.79	9.21	70.51	4.06	0.28	0.68	-0.01	-0.21	-1.28	0.08	0.0
Avg. of all			15.66	67.11	77.25	67.65	5.30	0.24						
Std. of all			12.59	183.77	220.08	32.50	7.94	0.17						

Table 7. Data of financial indicators.

(1) Largan Precision (3008)

Largan Precision (3008)			DS	25%	
3008	ROE%	4ER%	NetP\$b	Divends%	EPSS
2010	30	15	40.47	53	30.2
2011	32	26	51.73	45	38.6
2012	28	28	55.41	44	41.3
2013	42	18	96.32	41	71.8
2014t	50	12	153.30	40	114.3
slope	0.05	-0.014	27.025	-0.03	20.14

(2) Highwealth Construct. (2542)

Highwealth Construct. (2542)			DS	8%	
2542	ROE%	4ER%	NetP\$b	Divends%	EPSS
2010	58	2	69.56	77	9.8
2011	37	(4)	63.97	62	8.8
2012	27	(5)	54.99	34	9.2
2013	29	3	64.07	33	10.7
2014t	45	1	98.19	19	10.9
slope	-0.033	0.005	5.736	-0.146	0.427

(3) Formosa Plastics (1301)

Formosa Plastics (1301)			DS	24%	
Year	ROE%	4ER%	NetP\$b	Divends%	EPSS
2010	20	4	45,262	89	7.4
2011	14	(15)	35,573	92	5.8
2012	6	6	13,635	69	2.2
2013	9	35	20,582	54	3.2
2014t	8	65	21,473	59	3.4
slope	-0.029	0.172	-62,569	-0.098	-1.06

(4) Formosa Chem. & Fib. (1326)

Formosa Chem. & Fib. (1326)			DS	27%	
Year	ROE%	4ER%	NetP\$b	Divends%	EPSS
2010	21	31	474.56	88	8.3
2011	12	15	329.57	90	5.8
2012	3	(1)	70.43	69	1.2
2013	11	11	244.76	53	4.2
2014t	6	20	161.01	60	2.7
slope	-0.031	-0.026	-71.191	-0.093	-1.28

(5) Merry (2439)

Merry (2439)			DS	10%	
Year	ROE%	4ER%	NetP\$b	Divends%	EPSS
2010	13	8	6.01	98	3.8
2011	7	19	3.15	99	2.0
2012	11	11	4.81	76	2.9
2013	23	23	10.38	69	5.9
2014t	27	27	14.29	63	7.7
slope	0.044	0.042	2.379	-0.1	1.17

(6) Grape King Bio (1707)

Grape King Bio (1707)			DS	19%	
Year	ROE%	4ER%	NetP\$b	Divends%	EPSS
2010	22	49	4.31	69	3.3
2011	22	38	4.66	76	3.6
2012	28	29	6.18	84	4.7
2013	34	30	8.36	78	6.4
2014t	33	15	9.11	78	7.0
slope	0.034	-0.076	1.33	0.02	1.02

(7) The Ambassador Hotel (2704)

The Ambassador Hotel (2704)			DS	26%	
Year	ROE%	4ER%	NetP\$b	Divends%	EPSS
2010	3	(91)	2.51	47	0.7
2011	5	(88)	4.30	58	1.2
2012	3	30	2.83	43	0.8
2013	4	(20)	3.35	52	0.9
2014t	5	(26)	4.20	55	1.1
slope	0.003	0.198	0.243	0.01	0.05

(8) Dah San (1615)

Dah San (1615)			DS	22%	
Year	ROE%	4ER%	NetP\$b	Divends%	EPSS
2010	15	245	1.95	68	1.8
2011	4	276	0.58	66	0.5
2012	6	306	0.86	96	0.8
2013	2	139	0.31	52	0.3
2014t	3	149	0.46	108	0.4
slope	-0.050	0.013	-1.299	-0.027	-0.529

(9) Taroko Textile (1432)

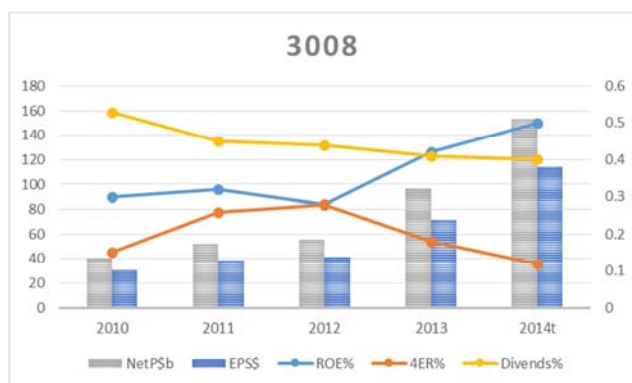
Taroko Textile (1432)			DS	12%	
Year	ROE%	4ER%	NetP\$b	Divends%	EPSS
2010	208	(51)	8.10	0	9.1
2011	2	(84)	0.23	0	0.3
2012	(10)	(88)	(1.07)	0	(1.2)
2013	(10)	36	(1.48)	0	(0.9)
2014t	(10)	loss	(2.32)	0	(1.4)
slope	-0.448	0.222	-2.255	0	-2.22

(10) Hey-Song (1234)

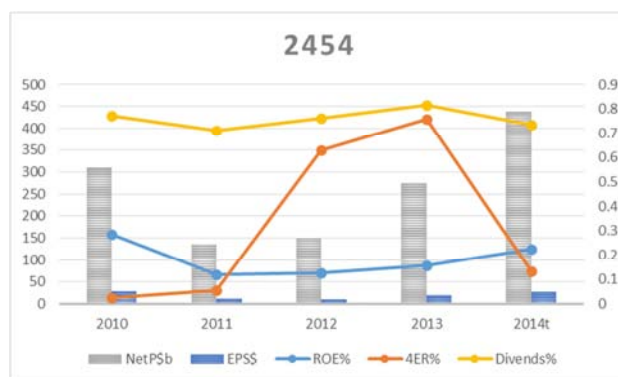
Hey-Song (1234)			DS	18%	
Year	ROE%	4ER%	NetP\$b	Divends%	EPSS
2010	3	-31	330	76	0.6
2011	4	264	394	81	0.7
2012	3	169	378	82	0.7
2013	-0	251	-39	354	(0.1)
2014t	3	388	452	0	1.1
slope	-0.006	0.824	-0.189	0.121	0.019

4ER%: Earnings reinvestment rate; NetP\$b: Net profits; DS: Shareholding ratio of directors and supervisors.

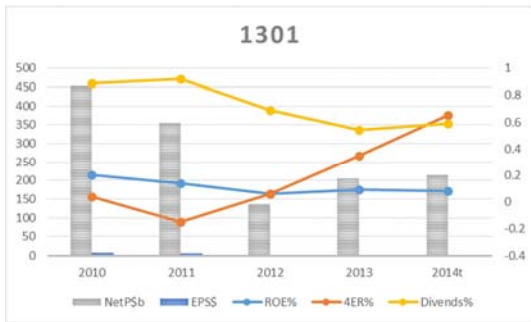
(1) Largan Precision (3008)



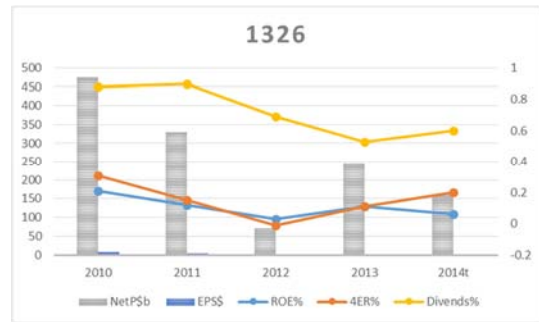
(2) Highwealth Construct. (2542)



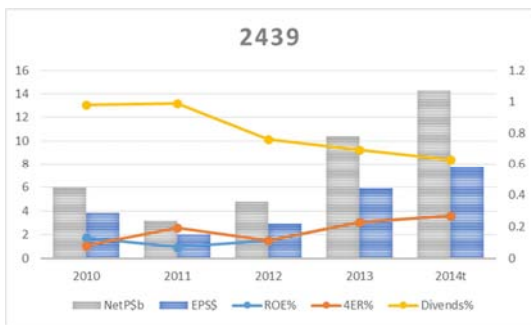
(3) Formosa Plastics (1301)



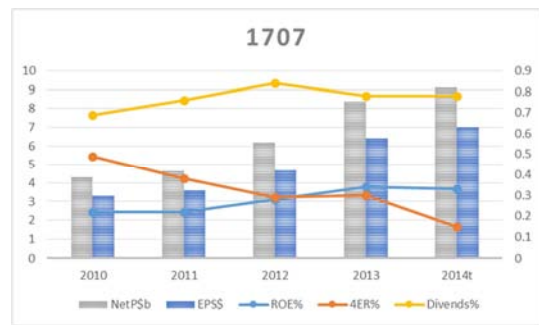
(4) Formosa Chem. & Fib. (1326)



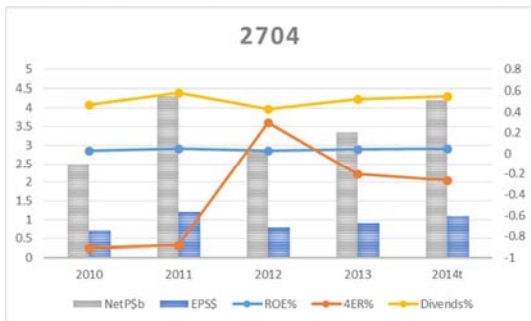
(5) Merry (2439)



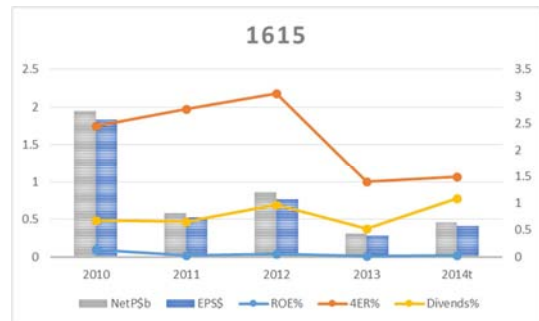
(6) Grape King Bio (1707)



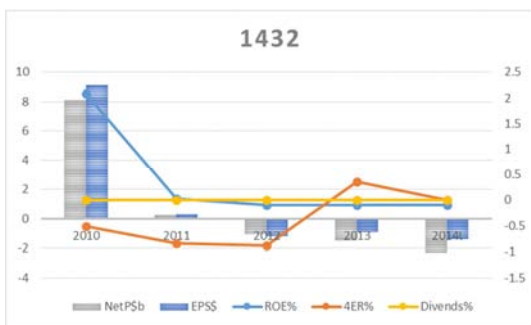
(7) The Ambassador Hotel (2704)



(8) Dah San (1615)



(9) Taroko Textile (1432)



(10) Hey-Song (1234)

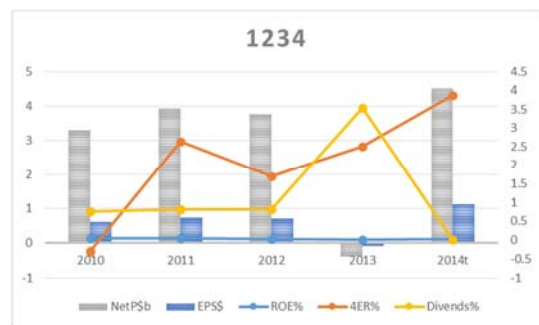


Figure 3. Graphs of financial indicators.

## 5. Conclusion

This research employed big data analytics to construct a computation platform for stock selection and trading strategies. It's combined the EPSO-GHSOM method and distinguished four stock patterns: star growth stocks, landmine stocks of poorly operating companies, stocks underpinned by stable or rising operating profitability, and stocks underpinned by falling operating profitability. These categories aid investors in defining standards for selecting safe, quality stocks.

Next, we applied the EPSO method and adjusted the time frame in simple moving average crossovers to determine timing for trading stocks. Performance was 82.14% accurate for stocks identified as outperform and 97.62% accurate on stocks selected as long-term holds. Trends in share price and bid-ask spreads emerged as the primary drivers of performance by individual stocks. Therefore, methods used in this research produced compelling results, and stock selection and trading strategies proved excellent. However, Taiwan's generally bullish equity markets aided ROI during the examined period. In addition, although the SMA golden cross and death cross criteria failed in this study, SMA crossovers remain valuable references in identifying trading opportunities, especially on medium-to-long-term simple moving averages.

SMA is a lagging indicator that reacts slowly to undulations in the market. Therefore, questions remain whether it is necessary to select leading indicators to move forward, or select high-diversity technical indicators to differentiate mixed signals and acquire information to enhance the accuracy of this forecasting model.

## References

- [1] Silver, Nate, *The signal and the noise: Why so many predictions fail-but some don't*. Penguin, 2012.
- [2] Teixeira, Lamartine Almeida, and Adriano Lorena Inacio De Oliveira. "A method for automatic stock trading combining technical analysis and nearest neighbor classification." *Expert systems with applications*, 2010, vol.37, no.10, pp. 6885-6890.
- [3] Fama, Eugene F., "Efficient capital markets: A review of theory and empirical work\*." *The journal of Finance*, 1970, vol.25, no.2, pp. 383-417.
- [4] Haugen, Robert A., *The inefficient stock market: What pays off and why*. Upper Saddle River, NJ: Prentice Hall, 1999.
- [5] Los, Cornelis A., "Nonparametric efficiency testing of Asian stock markets using weekly data." *Centre for Research in Financial Services Working Paper*, 1998, pp. 99-01.
- [6] Ohkawa, E., Chen, Y., Mabu, S., Shimada, K., & Hirasawa, K. "Evaluation of varying portfolio construction of stocks using Genetic Network Programming with control nodes." *SICE Annual Conference*, 2008.
- [7] Kwon, Yung-Keun, and Byung-Ro Moon., "A hybrid neurogenetic approach for stock forecasting." *Neural Networks, IEEE Transactions on*, 2007, vol. 18, no. 3, pp. 851-864.
- [8] Chavarnakul, Thira, and David Enke., "A hybrid stock trading system for intelligent technical analysis-based equivolume charting." *Neurocomputing*, 2009, vol. 72, no. 16, pp. 3517-3528.
- [9] Rodríguez-González, A., García-Crespo, Á., Colomo-Palacios, R., Iglesias, F. G., & Gómez-Berbis, J. M. "CAST: Using neural networks to improve trading systems based on technical analysis by means of the RSI financial indicator." *Expert systems with Applications*, 2011, vol. 38, no. 9, pp. 11489-11500.
- [10] Wang, Lipo, and Shekhar Gupta., "Neural networks and wavelet de-noising for stock trading and prediction." *Time Series Analysis, Modeling and Applications*. Springer Berlin Heidelberg, 2013, pp. 229-247.
- [11] Lin, Xiaowei, Zehong Yang, and Yixu Song., "Intelligent stock trading system based on improved technical analysis and Echo State Network." *Expert systems with Applications*, 2011, vol. 38, no. 9, pp. 11347-11354.
- [12] Olatunji, S. O., Al-Ahmadi, M. S., Elshafei, M., & Fallatah, Y. A. "Forecasting the Saudi Arabia Stock Prices Based on Artificial Neural Networks Model." *International Journal of Intelligent Information Systems*, 2013, vol. 2, no. 5, pp. 77.
- [13] Rath, Smita, Alok Kumar Jagadev, and Manoj Ranjan Nayak., "Performance Analysis of Stock Market Using Artificial Neural Network." *International Journal of Applied Engineering Research*, 2015, vol. 10, no. 4.
- [14] Chien, Ya-Wen Chang, and Yen-Liang Chen., "Mining associative classification rules with stock trading data—A GA-based method." *Knowledge-Based Systems*, 2010, vol. 23, no. 6, pp. 605-614.
- [15] Hsu, L. Y., Horng, S. J., He, M., Fan, P., Kao, T. W., Khan, M. K.,... & Chen, R. J. "Mutual funds trading strategy based on particle swarm optimization." *Expert Systems with Applications*, 2011, vol. 38, no. 6, pp. 7582-7602.
- [16] Gunasekarage, Abeyratna, and David M. Power., "The profitability of moving average trading rules in South Asian stock markets." *Emerging Markets Review*, 2001, vol. 2, no. 1, pp. 17-33.
- [17] Kennedy, James., "Particle swarm optimization." *Encyclopedia of Machine Learning*. Springer US, 2010, pp. 760-766.
- [18] Dittenbach, Michael, Andreas Rauber, and Dieter Merkl. "Uncovering hierarchical structure in data using the growing hierarchical self-organizing map." *Neurocomputing*, 2002, vol. 48, no. 1, pp. 199-216.
- [19] Wang, Jar-Long, and Shu-Hui Chan., "Stock market trading rule discovery using pattern recognition and technical analysis." *Expert Systems with Applications*, 2007, vol. 33, no. 2, pp. 304-315.
- [20] Peng, Hsin-Tsung, Hahn-Ming Lee, and Jan-Ming Ho., "Trading decision maker: Stock trading decision by price series smoothing and tendency transition inference." *-Technology, e-Commerce and e-Service, 2005. IEEE'05. Proceedings. The 2005 IEEE International Conference on*. IEEE, 2005.

- [21] Chang, Pei-Chann, Chin-Yuan Fan, and Chen-Hao Liu., "Integrating a piecewise linear representation method and a neural network model for stock trading points prediction. Systems," *Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 2009, vol. 39, no. 1, pp. 80-92.
- [22] Lin, Nana, Xinwei Zhang, and Siqi Lv. "CAN WEB NEWS MEDIA SENTIMENTS IMPROVE STOCK TRADING SIGNAL PREDICTION?". 2014.
- [23] Butler, Matthew, and Dimitar Kazakov., "Testing implications of the adaptive market hypothesis via computational intelligence." *Computational Intelligence for Financial Engineering & Economics (CIFER), 2012 IEEE Conference on. IEEE*, 2012.
- [24] Pham, H. V., Cooper, E. W., Cao, T., & Kamei, K. "Hybrid Kansei-SOM model using risk management and company assessment for stock trading." *Information Sciences*, 2014, vol. 256, pp. 8-24.
- [25] Kohonen, Teuvo., "Self-organizing maps." *Springer Science & Business Media*, 2001, pp. 30.
- [26] Fama, Eugene F., and Kenneth R. French., "Size and book - to-market factors in earnings and returns." *The Journal of Finance*, 1995, vol. 50, no. 1, pp. 131-155.
- [27] Graham, Benjamin, and David L. Dodd., *Security analysis: principles and technique*. McGraw-Hill, 1934.
- [28] Buffett, Warren, and Carol Loomis., "Warren Buffett on the stock market." *Fortune*, 2001.
- [29] Hagstrom, Robert G., *The Warren Buffett way: Investment strategies of the world's greatest investor*. John Wiley & Sons, 1997.
- [30] Zhong Hua., "Buffett stock valuation investment strategy applied research." *Securities bimonthly*, 2011, vol. 586, pp. 80-109.
- [31] Piotroski, Joseph D., "Value investing: The use of historical financial statement information to separate winners from losers." *Journal of Accounting Research*, 2000. pp. 1-41.
- [32] Mohanram, Partha S., "Separating Winners from Losers among Low Book-to-Market Stocks using Financial Statement Analysis." *Review of Accounting Studies*, 2005, vol. 10, no. 2-3, pp. 133-170.
- [33] Liaw, Siqin., *Technical Analysis: An Asian Perspective*. 2012.
- [34] Gartner. <http://www.gartner.com/it-glossary/big-data/>.
- [35] Zikopoulos, P. C., Eaton, C., DeRoos, D., Deutsch, T., & Lapis, G. *Understanding big data*. New York et al: McGraw-Hill, 2012.
- [36] Zikopoulos, P., Parasuraman, K., Deutsch, T., Giles, J., & Corrigan, D. *Harness the Power of Big Data The IBM Big Data Platform*. McGraw Hill Professional, 2012.
- [37] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding., "Data Mining with Big Data." *Transactions on Knowledge and Data Engineering*, 2014, vol. 26, no. 1, pp. 1041-4347.
- [38] Chen, CL Philip, and Chun-Yang Zhang., "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data." *Information Sciences*, 2014, vol. 275, pp. 314-347.
- [39] EMC Institute, VMware research team, Hu Jia Xi(translation)., *Next step of Big Data: Big Data new strategy, technology and large-scale Web applications*. Saatchi Times Press, 2014.
- [40] Palvia, P., Leary, D., Mao, E., Midha, V., Pinjani, P., & Salam, A. F. "Research methodologies in MIS: an update." *Communications of the Association for Information Systems*, 2004, vol. 14, no. 1, pp. 24.
- [41] Hong Ruitai., *Buffett stock magic book*. Smart Inv Press, 2004.
- [42] Liu, Wenqing, Tingyu Chen, and Mike YJ Lee., "A Method for Stock Trading Strategy Combining Technical Analysis and Particle Swarm Optimization." *Journal of Convergence Information Technology*, 2014, vol. 9, no. 5.