Journal of Systems Science and Information
Dec., 2014, Vol. 2, No. 6, pp. 481–504

An EPC Forecasting Method for Stock Index Based on Integrating Empirical Mode Decomposition, SVM and Cuckoo Search Algorithm

Xiangfei LI
School of Management, Tianjin Polytechnic University, Tianjin 300387*, China*

Zaisheng ZHANG
College of Management and Economics, Tianjin University, Tianjin 300072*, China*

Chao HUANG
School of Accountancy, Shanghai University of Finance and Economics, Shanghai 200433*, China*

Abstract In order to improve the forecasting accuracy, a hybrid error-correction approach by integrating support vector machine (SVM), empirical mode decomposition (EMD) and the improved cuckoo search algorithm (ICS) was introduced in this study. By using two indexes as examples, the empirical study shows our proposed approach by means of synchronously predict the prediction error which used to correct the preliminary predicted values has better prediction precision than other five competing approaches, furthermore, the improved strategies for cuckoo search algorithm has better performance than other three evolutionary algorithms in parameters selection.

Keywords error-correction; stock index forecasting; empirical mode decomposition; SVM; cuckoo search algorithm

1 Introduction

Stock investment which has long been one of the main activities for making profits has getting more and more attention. For stock market has characteristic of high risks, investors need to conjecture the volatility of the stock indexes in order to make sensible investment decisions. Therefore, effective forecasting models which can greatly reduce the personal decision mistakes through providing more accurate predictions are main concern for the investors and researches. For the inherently nonlinear and non-stationary properties of financial time series, instead of single method, hybrid models are becoming widely used to solve the limitations in financial time series forecasting. For example, Pai and Lin[1] integrate the ARIMA with support vector machine and get a better model for stock price forecasting. Wang[2] proposed a hybrid method provides higher predictability by integrate the GJR with GARCH. Chi et al.[3] use the grey theory and neural networks together and provide a prediction model which conquered the convergence problem caused by large amount input data. Wang et al.[4] gives a hybrid forecasting model which combine the smoothing model (ESM), ARIMA and BPNN. Lu et al.[5]

Received January 22, 2014, accepted April 14, 2014

provided an integrated model of NLICA, SVR and PSO which has effective predictive results comparing with 4 comparison models.

Among the forecasting methods, the support vector machine (SVM) is a promising methods which work more effectively than the traditional linear model in time series forecasting area, for it uses a risk function consisting of the empirical error and a regularized term which is derived from the structural risk minimization principle[6]. Because of its outstanding performance,the SVM has now successfully applied in forecasting the financial time-series[7, 8].

However, single SVM can hardly provide satisfying results either, therefore, like the above integration stragetics, there are many mixed methods integrate with the SVM model. For example, Chiu and Chen[7] proposed a dynamic fuzzy model based on the SVM method and choose the genetic algorithm (GA) to adjust the influential degree of each input variable, the experiment results show the model generated better accuracy rate than the traditional forecast methods. Owing to the easily ignoring of the non-stationary nature of stock price series, Hsu, Hsieh, Chih et al.[9] provide a solution by integrated the self-organizing map and SVM together. Lee[10] develope a prediction model based on SVM with a hybrid feature selection method to predict the trend of stock markets, in his study, it showes that SVM outperforms BPN to the problem of stock trend prediction. Kao, Chiu et al.[8] integrate the nonlinear independent component analysis and SVM for forecasting the stock price.

Generally speaking, no matter how perfect the forecasting method is, error is inevitably, the SVM is no exception. However, if we can effectively predict the error when we trying to predict certain time series by the SVM or other methods, it might be get better results by modify the error. The error predicting idea can be found widely used in engineering science, such as river and flood forecasting[11], electrical equipment load forecasting[12], weather forecasting[13, 14]. For example, Madsen and Skotner[11] proposed a new data assimilation procedure based on general filtering update combined with error forecasting, the error forecast model was used to propagate model error at measurement points in the forecast period. the results showed the new model significantly improved the flood forecasting ability. On the purpose of energy saving, Yao et al.[12] proposed a novel forecasting model called "RBFNN" with combined residual error correction to provide accurate air-conditioning load forecasting. The study case indicates the RBFNN with combined residual error correction has a much better forecasting accuracy. The application of error prediction can also be found in economic field. For example, Zhou et al.[15] proposed a novel ARIMA approach on forecasting electricity price with the improvement of predicted forecasting error for the first time, the results showed the presented approach improves the accuracy. Chen, Leung[16] proposed an adaptive forecasting approach which combines the strengths of neural networks and multivariate econometric models to predict the exchange rates. Anderson[17] discussed in detail the specification of a vector error correction forecasting model (called VECM) that is anchored by long-run equilibrium relationships suggested by economic theory. The model was proofed more accurate than the traditional forecasting model.

The key to the error correction forecasting method is how to predict the error value effectively, for the forecasting results after correction may have even bigger deviation if our error predictive value is not accurate. However, owing to the high-frequency, non-stationary and chaotic properties, it is hardly to get satisfying results for the error forecasting, therefore, to

solve this problem, a extraction technique which generally utilized to extract features contained in the signals is necessary, for the forecasting model based on these features could have better performance[18–20].

The empirical mode decomposition (EMD) which mainly used in extracting information contained in signals, is a signal processing technique primarily applied to image processing or signal processing, however, with its powerful feature extraction capability, it has now been successfully applied to time-series studies[21–25]. For example, Zhu, Sun and Li[21] used the EMD technique to decompose the load time series into a series of smooth intrinsic mode functions (IMFs) with different scales and then used the SVM forecast each IMF respectively and obtained the final results by summing up the forecasting results of each IMF together. Yu et al.[22] proposed an EMD based neural network ensemble learning paradigm for world crude oil spot price forecasting, owing to the decomposition work of EMD, each IMF was accurately predicted. The EMD is suitable for the time series in terms of finding fluctuation tendency, which simplifies the task into simple forecasting subtasks[26]. For its abilities of revealing the hidden patterns and trends of time series, we use the EMD technique to process the error time series caused by single SVM for easing the work of next step's forecasting.

Another problem in SVM forecasting is parameters selecting. The common practice is to optimize the penalty function and kernel function by intelligent optimization algorithms which most commonly includes: genetic algorithm (GA)[27–29] and particle swarm optimization (PSO)[30, 31]. With the continuous development of intelligent heuristic algorithm, a new optimization algorithm, cuckoo search (CS) was proposed in 2009 by Yang and Deb[32]. The algorithm is inspired by the reproduction strategy of cuckoos. The main component of CS is using Lévy flights which used as the searching pattern. For the Lévy flights is a random walk that is characterised by a series of instantaneous jumps chosen from a probability density function which has a power law tail, this kind of searching made CS can find all optima in a design space, therefore, it has been widely used in engineering science[33–35]. However, because the CS is a new algorithm, there are some defects like insufficiency searching energy, low accuracies that could not overcome completely. On basis of CS algorithm, in this paper we put forward several improved strategics and form a new algorithm called improved-cuckoo-search (ICS) algorithm which used to optimize the SVM parameters selecting. The following study shows that the ICS indeed has better performance than methods of grid search, GA, PSO and single CS.

The main highlights of this paper are:

1) An error forecasting and correcting method (called EPC for short) for stock forecasting which integrates the EMD, SVM and ICS was proposed. By forecasting the possible error simultaneously and using the predictive error to correct the preliminary results, we got better forecasting results with higher accuracy.

2) For the error sequence forecasting, to solve the problems of high frequency, non-stationary and chaotic properties, we introduced a feature extraction process, the empirical mode decomposition (EMD), into our study. By using the EMD technique we decomposed the error series into a series of smooth intrinsic mode functions (IMFs) with different scales and then used the SVM forecast each IMF respectively and obtained the final results by summing up them.

3) For the cuckoo search algorithm, for the defects of lacking search abilities and low ac-

curacy, we proposed several improved strategies. The results show that our improved cuckoo search algorithm (ICS) has better accuracy and less evolution steps than other 4 bench marking methods in selecting parameters for SVM model.

The basic chapter arrangement are as follows. Section 2 gives brief introduction of the SVM and EMD methods, then detail concepts and processes of the improved strategics for CS algorithm are explained. Section 3 presents the research scheme for this research. Section 4 containing the empirical results and robustness evaluation from two sample of SSEC indexes and NASDAQ indexes. Section 5 gives the conclusions.

2 Research methodology

2.1 Support vector machine

For the support vector machine plays a domain role in the forecasting model, in this section, we will have a brief introduce of its principal as well as the process. SVM was proposed by Vapnik[36] at 1986. It has built up statistical learning theory and has got more and more attention because of its outstanding ability in solving nonlinear regression estimation problems. The SVM, based on the structural risk minimization principle, its basic thought is mapped the data X_i into the space F which has high altitudes characteristics and then set linear regression equation. The equation can expressed as:

$$f(X) = (w,\varphi(X))+b \tag{1}$$

Where w is the weight vector, b is bias, $\varphi(X)$ is the nonlinear mapping of R^m space to F space. The traditional prediction or classification method is to find $f \in F$ to make sure minimize the structural risk value. The structural risk equation is expressed as:

$$R_{\text{reg}}=\lambda\|w\|^2+R_{\text{emp}}[f]=\sum_{i=1}^{S} C(e_i)+\lambda\|w\|^2 \tag{2}$$

Where $\|w\|^2$ is the incredible risk which reflects the complexity of the model, $R_{emp}[f]$ is the empirical risk, λ is constant used to balance the complexity and the loss error of the model, $C(e_i)$ is the experience losses of the model, S is the sample capacity. For established loss function, this problem can be transformed into the optimal solution of the quadratic programming problem. According to Vapnik[36], the ε-insensitivity loss function can be defined as:

$$|y-f(x)|_{\varepsilon}=\begin{cases} |y-f(x)|-\varepsilon, & \text{if } |y-f(x)|\geq\varepsilon \\ 0, & \text{if } |y-f(x)|\leq\varepsilon \end{cases} \tag{3}$$

Where ε controls regression error range, the smaller the value, the higher of the accuracy, but with generalization ability decreases. Based on this kind of loss function, the empirical risk can be defined as:

$$R_{\text{emp}}^{\varepsilon}[f]=\frac{1}{S}\sum_{i=1}^{S}|y-f(x)|_{\varepsilon} \tag{4}$$

Combine with the Eq.(1)∼Eq.(4), the original problem can be transformed into a functional problem of minimize the linear risk as follows:

$$\min\eta=\frac{1}{2}w^{\mathrm{T}}w+C\sum_{i=1}^{S}(\zeta_i+{\zeta_i}^*)$$

$$\text{s.t.}\begin{cases} y_i-(w,\varphi(X_i))-b \le \varepsilon+\zeta_i \\ (w,\varphi(X_i))+b-y_i \le \varepsilon+\zeta_i \\ \zeta_i,\zeta_i{}^* \ge 0 \end{cases} \tag{5}$$

Where $C=1/\lambda$, ε is the estimate of the accuracy, ζ_i, $\zeta_i{}^*$ are slack variables. To facilitate the solution of the problem, we transformed it into the dual problem as follows:

$$\max \mu = -\frac{1}{2}\sum_{i,j=1}^{S}(\alpha_i-\alpha_i^*)(\alpha_j^*-\alpha_j)(\phi(X_i),\phi(X_j))+\sum_{i=1}^{S}\alpha_i^*(Y_i+\varepsilon)-\sum_{i=1}^{S}\alpha_i(Y_i+\varepsilon)$$
$$\text{s.t.}\begin{cases} \sum_{i=1}^{S}\alpha_i^* = \sum_{i=1}^{S}\alpha_i \\ 0 \le \alpha_i \le C \\ 0 \le \alpha_i^* \le C \end{cases} \tag{6}$$

Solve the Eq.(6) and obtain the w and b which taken to Eq.(1) can get the nonlinear function $f(X)$:

$$f(X)=\sum_{i=1}^{S}(\alpha_i-\alpha_i^*)(\phi(X_i),\phi(X))+b \tag{7}$$

The kernel function which define the computation of high-dimensional space can be expressed as $K(x_i,x_j)=\varphi(x_i)\varphi(x_j)$. In this study we use the type of radial basis function $K_{rbf}(x_i,x_j)=\exp(-\gamma\|x_i-x_j\|^2)$. Based on this, the prediction problem can be transformed into a solution of quadratic programming decision function problem. This study will use EMD and ICS algorithm to optimize the parameter of the prediction SVM model.

2.2 Empirical mode decomposition

Empirical mode decomposition (EMD) is a promising nonlinear, non-stationary data processing method proposed by Huang et al.[37, 38]. It considers the real time series as fast oscillations super imposed on slow oscillations. Those oscillations in data are extracted based on the principle of local scale separation, and are approximated by "intrinsic mode functions". An IMF must satisfy the following two conditions: 1) the number of extrema and zero-crossings are the same, or differ at the most by one; 2) they are symmetric with respect to local zero mean.

A sifting process is designed to extract IMFs level by level. First, the IMF with the highest frequency riding on the lower frequency part of the data is extracted, and then the IMF with the next highest frequency is extracted from the differences between the data and the extracted IMF. The iterations continue until no IFM is contained in the residual. The overall sifting procedure for a time series S(t) is described as follows:

If the extrema number of the original series is more than zero-crossings over two, the decomposition shall begin.

1) Identify all the maxima and minima of the original series, then using the cubic spline interpolation method estimate the envelope function.

2) Calculate the average value of the maximum and minimum envelope, expressed as $m_1(t)$.

Define the difference of S(t) and $m_1(t)$ as

$$S(\mathrm{t})-m_1(t)=h_1(t) \tag{8}$$

Where $h_1(t)$ is the lower frequency series that get ride of the high frequency part of S(t).

3) If $h_1(t)$ is still not smooth, the EMD process will continue, repeat the above process, theoretically, until the average envelope value is zero. However, if the average envelope is zero, some physical meanings of the amplitude or frequency modulation might be eliminated, therefore, it is necessary to adjust the standard, let

$$\mathrm{SD}=\sum_{\mathrm{t}=0}^{T}\left[\frac{\left[m_{(k-1)}(t)-m_k(t)\right]^2}{m^2{}_{(k-1)}(t)}\right] \tag{9}$$

Where $m_k(t)$ is the average envelope function of the k loops. In this study, we let the SD ranged between 0.1 and 0.2, this standard will relax the requirements of the average envelope properly and help retained the physical meaning of IMF to a certain extent.

4) According to the above process, we get the first component $C_1(t)$ which defined as

$$h_{1(k-1)}(t)-m_{1k}(t)=C_1(t) \tag{10}$$

Where $C_1(t)$ as the first component has the highest frequency, with the original series $S(t)$ minus it can get a slightly smooth series $r_1(t)$, for which repeat the above operation can get the second component of $C_2(t)$ and $r_2(t)$. Repeat it until get $r_n(t)$ can not be decomposed again, this is the end of EMD decomposition.

$$r_{n-1}(t)-C_n(t)=r_n(t) \tag{11}$$

Where $r_n(t)$ represents the overall trend of the original sequence $S(t)$, then the original sequence is decomposed into several components and a overall trend.

$$S(t)=\sum_{j=1}^{n}C_j(t)+r_n(t) \tag{12}$$

Every IMF components have different vibration frequencies and amplitudes which represent different scales information of the original sequence.

Actually, there will be a mode mixing problem if the data has intermittency. Mode mixing is defined as a single IMF consisting of signal of widely disparate scales, or a signal of a similar scale residing in different IMF components. To overcome this problem, Wu and Huang[39] proposed the ensemble EMD method. The main procedure of EEMD is to add a white noise series to the targeted data series and then decompose the data with white noise added into IMFs. Then repeat the steps iteratively and obtain the means of corresponding IMFs of the decompositions as the final results.

Therefore, this paper uses the EEMD (we collectively referred to as EMD) method to decompose the training error and predicting error sequence into several components with different time scale characteristics that convenient for predict the error sequence. When processing the data, in order to get the effective information, the original sequence extremum problem of end point is overcame by using the polynomial fitting method[40]. The main process of EMD-SVM is illustrated in Fig.1.

2.3 Cuckoo search algorithm

2.3.1 Principle of cuckoo search algorithm

Cuckoo search actually belongs to a kind of random search way caused by their unique brood parasitism behavior on zoology principle. According to the research conclusions of zoologist, some kinds of cuckoos have lazy temperament, they never nesting, hatching, or brooding in breeding season, instead, they adopt a way by brood parasitism to reproduction. That is mean they will look for some host birds which have similar physiological and diets with themselves. Usually these cuckoos will quickly lay their eggs in the nest of host species when the hosts go out for food or something else, then leaving those parents to hatch and nurture for its young[41–43].

The cuckoo search algorithm actually is a simulation of the random walk search process that cuckoos looking for suitable host nest for laying eggs. The same way of traveling is very common in other animals' foraging process, such as the albatross[44], bees[45], fruit fly[46], spider monkey[47], baboon[48], etc. They all followed with the Lévy flight distribution which is the best search strategy when there are several independent searchers and the target is randomly distributed. The general process of cuckoo search simulation is to initialize several bird nests, calculate the fitness value of each nest, and then let the bird update its habitat location followed the Lévy flight way until the global best solution point founded. The cuckoo Lévy flight search pattern is illustrated in Fig.2.

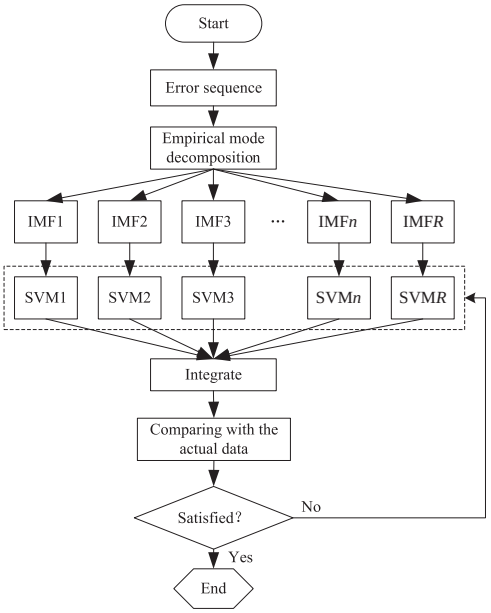


Figure 1 The process of EMD-SVM

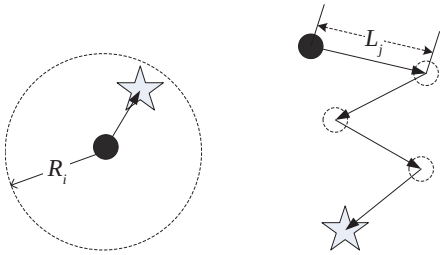


Figure 2 Lévy flight way of cuckoo

The R_i means cuckoo's search radius which is the largest step that cuckoo can update at on time. When the goal host's nest is within the radius, the cuckoo will fly directly to it in a straight line, in contrast, it will search in Lévy flight way[49]. Its random walk step L_j is drawn from a Lévy distribution.

$$Lévy \sim P(L_j) = {L_j}^{-\mu}, 1 < \mu \leq 3 \tag{13}$$

That is to say, when have new host's nest x_i^{t+1} for, say, a cuckoo i, its travel flight way is performed as

$$x_i^{t+1} = x_i^t + \alpha L \oplus L\acute{e}vy(\mu) \tag{14}$$

Usually the α= 1. The above equation is essentially the stochastic equation for random walk. The product $\oplus$ means entry wise multiplications which is similar with those used in PSO, but more efficient in exploring the search space as its step length is much longer in the long run.

Here the steps essentially form a random walk process obey power law step-length distribution with a heavy tail[50]. Some of the new solutions should be generated by Lévy walk around the best solution obtained so far, this will speed up the local search. However, a substantial fraction of the new solutions should be generated by far field randomization and whose locations should be far enough from the current best solution, will make sure the system will not be trapped in a local optimum.

2.3.2 Improvement of cuckoo search algorithm

The cuckoo search algorithm proposed by Yang and Deb[32] based on three ideal situations: firstly, each cuckoo only can lay one egg at a time and dump its egg in randomly chosen nest; secondly, the best nests with high quality of eggs will carry over to the next generations; thirdly, the number of available host nests is fixed. In this assumption, the cuckoo has characteristics of succinct and easy to realize, while, it also leads some defects that like other evolutionary algorithm, for lack search ability and low accuracy. Based on this, this paper proposed several improved solutions according to the cuckoo's natural habitat, we called it the improved cuckoo search (ICS).

Similar to the chromosome in genetic algorithm and particle location in particle swarm algorithm, we use nest location as the data point in ICS. The basic information is illustrated in Fig.3.

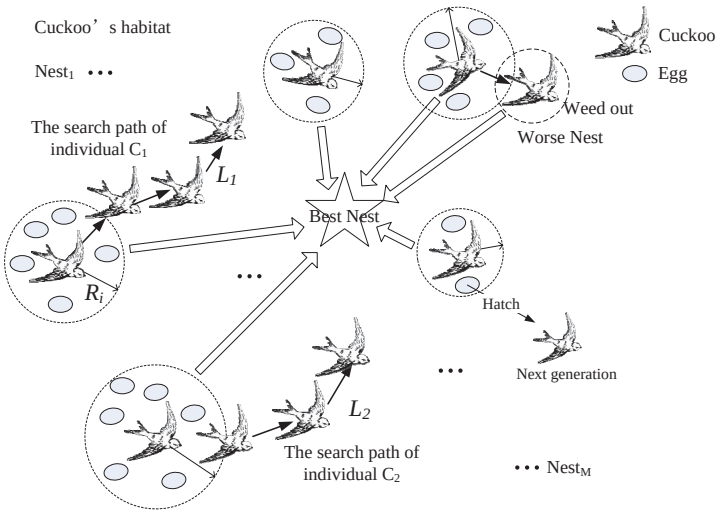


Figure 3 The migration and brood parasitism of cuckoos

1) The figure contains the whole habitat where there are totally M nests and n birds. Each nest location is a multi-dimensional vector and has a fitness value. For a m dimensional problem, say, for cuckoo i, $Nest_i = [N_1, N_2, \cdots, N_m]$, its fitness value is performed as $f(Nest_i)$, $i \in [1, M]$.

2) The cuckoo only lay eggs in a certain range of space which measured by the radius, as shown in the figure, the radius is R_i. Actually the cuckoo would not lay only one egg every time in breeding period, in fact it will lay eggs randomly in its spawn space and the number are proportional to the radius and satisfy the following relation[51]:

$$C_i(R_i) = \phi \times \left(C_i\text{eggs} \Big/ \sum_{\text{i=1}}^{\text{n}} C_i\text{eggs} \right) \times \text{Var}_{\max} - \text{Var}_{\min} \tag{15}$$

For cuckoo i, $C_i(R_i)$ means its searching radius, $C_i$eggs means the eggs number it lay at one time, n is the total number of all cuckoos, $\text{Var}_{\max} - \text{Var}_{\min}$ means the interpolation between the max step and min step that determines the search range and accuracy.

3) Use the individual C_1 and C_2 as examples, each generation, the cuckoo will search for goal nest followed with the Lévy distribution, and its searching step L_j is random variable. On the one hand, the random searching step will not be easily trapped in local optimum, on the other, it will lose accuracy because of sudden big step may leads the fitness value being far away from the optimal solution, therefore, self-adaptive searching step is necessary. According to the relationship between the searching radius and egg number that reflecting in Eq.(15), we assumed that the searching step is changing with the laying eggs number. To explain in more detail, suppose that cuckoos searching step take random value between 0 to its searching radius, then make the step length $L = \alpha C_i(R_i)$, $\alpha \in \text{rand}[0, 1]$. In the beginning of the search, cuckoo will lay more eggs in order to adapt to the hostile environment, while along with the increasingly close to the optimal solution, it will lay less eggs to reduce burden. Therefore, the searching step is self-adaptive which will be shorter with closing to optimal solution.

$$C_i\text{eggs} = round\left[C_i\text{eggs}(O) \cdot K \cdot \frac{|f_{\text{all}}(N_{\text{best}}) - f_i(N_j)|}{f_{\text{all}}(N_{\text{best}})} \right] \tag{16}$$

In Eq.(16), the product *round* will assure the egg number is integer. For cuckoo i, $C_i$eggs(O) means the original number of the eggs, $f_i(N_j)$ means the fitness value of the j^{th} nest, $f_{\text{all}}(N_{\text{best}})$ means the best solution of all nest, K is the adjustment coefficient which controls the egg numbers. By controlling the searching step, accordingly the accuracy will improved. It is noticed that no matter how much cuckoo lays eggs, there is only one can survive in the nest, for once hatched, the chicks will push out the other eggs and enjoy the only tending.

4) The eggs laid by cuckoos will be discovered by the host birds with a probability P (usually 10%). In this case, the host birds can either throw the egg away or abandon the nest. Undetected eggs have chance to be hatched and become the next generation birds that will searching for better hosts' nests in their spawning space, in this paper, the goal nest means better fitness values.

5) When a cuckoo migrates to a new position with the fitness value lower than its last nest, it will be regarded as that the bird is deviating from its goal nest which has better fitness value, therefore, we weed out these kinds of bird in order to avoid the unnecessary calculation. Moreover, for the purpose of increasing the random properties, let the random eggs being

hatched into mature birds in the next generation and continue their searching follows the Lévy flight.

Therefore, the improved CS has properties as follows: Firstly, as mentioned in [50], the Lévy flight search will not be the best strategy unless there are several independent searchers, thus, the ICS introduced several cuckoos to search independently in parallel with the purpose of increasing the searching ability and efficiency as well as meeting the requirements of best searching strategics. Secondly, the cuckoo eggs hatched and become the mature birds in next generation, this biological evolution style increase the diversity of cuckoo populations and search randomness. Thirdly, by controlling the egg spawning way, we get self-adaptive searching steps, which assured that the wider searching range in the beginning while higher accuracy at later stage.

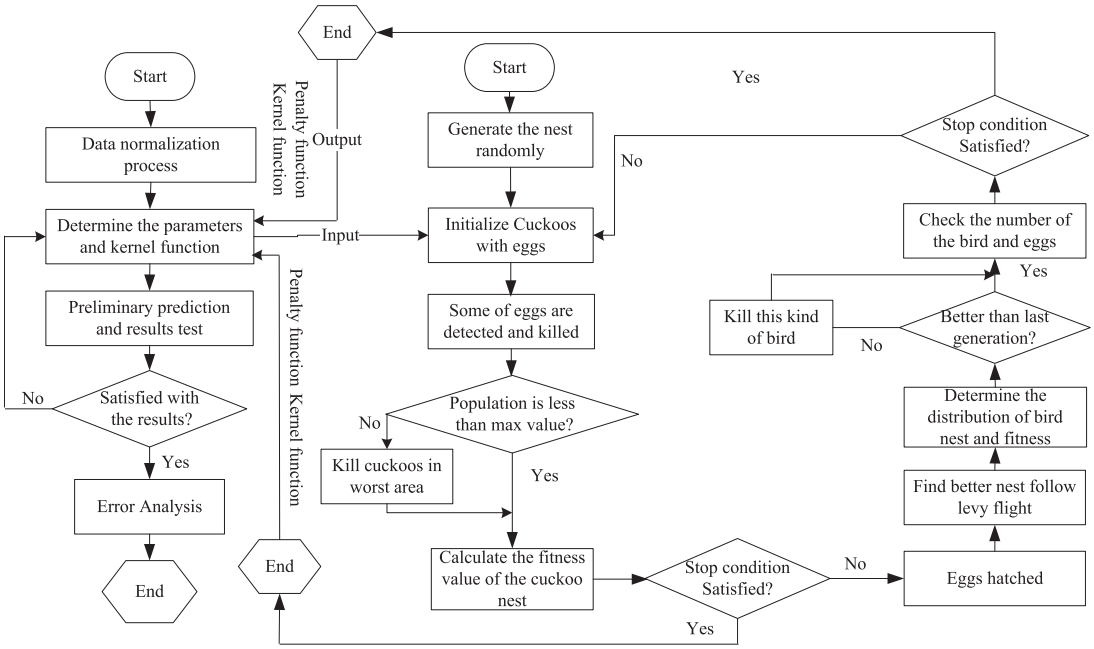


Figure 4 The process of SVM optimized by ICS

The central factors for a SVM model are the penalty function C and kernel function g. Therefore, we put forward to optimize them by using the ICS algorithm method. The process of the step for the optimization is illustrated in Fig.4 and the detailed illustration is provided as follows:

Step 1 Data initialize. Generate n bird's nests and the same number of bird, the location of which also recorded, indicated as $Nest^{(0)}= [N_1^{(0)}, N_2^{(0)}, \cdots, N_n^{(0)}]$. Set the largest population $Pop_{\max}$, the maximum number of iterations $iter_{\max}$, the $Var_{\max}$ and $Var_{\min}$, let the $egg_i \in [2, 5]$. For the j^{th} nest, it has M dimensions feature when it refers to M dimensional optimization problems.

Step 2 Let the cuckoos spawning randomly within its radius, because some of the eggs (usually 10%), which are not similar to the host's eggs, are detected and killed by the hosts. Each generation, for each cuckoo, generate the probability P obeying uniform distribution, that is $P \in rand[0, 1]$, if $P < 10\%$ means the current nest has less profit values and the egg in which

are killed by host birds.

Step 3 Count the current existing cuckoo bird number n_{birds} and egg number n_{eggs}, calculate the fitness value of their corresponding nest location. If $n_{\text{birds}} + n_{\text{eggs}} > Pop_{\max}$, then kill the extra birds or eggs which have less fitness value. Record the best nest location $Nest_{\text{best}}^{(0)}$ with best fitness value $Fitness_{\text{best}}{}^{(0)}$=$f(Nest_{\text{best}}^{(0)})$.

Step 4 If the $Fitness_{\text{best}}{}^{(0)}$ of step 3 has not yet reach the precision requirements, then make the current bird eggs hatched and grown to the adult birds, then let all of the adult birds update their habitats through Lévy flight migration. Then all the birds have their new location of habitats set, as well as the corresponding fitness value set. Record the current best nest location $Nest_{\text{best}}^{(1)}$ with best fitness value $Fitness_{\text{best}}{}^{(1)}$.

Step 5 Compare $Fitness_{\text{best}}{}^{(0)}$ with $Fitness_{\text{best}}{}^{(1)}$, choose the larger one and use its nest location set to replace the smaller one. Do it like this is let all birds migrate to the better nest set which has better fitness value. Meanwhile, weeding out the birds whose fitness values are lower than the previous generation, for those birds may migrate to worse environments.

Step 6 Determine whether the best solution in step 5 satisfy the precision requirements, if yes then output the global optimal solution, otherwise back to step 2, repetitive compute until get the satisfying results.

It is noticed that there is a significant difference the weed out way among step 2, step 3 and step 5. Step 2 only weed out eggs, not only reflects the true state of cuckoo's breeding procedure in nature, but also to a certain extent increases the random disturbance and searching ability of the ICS; Step 3 weed out both adults and eggs, for the purpose of control the total population; Step 5 only weed out the adults, for only adults can migrate following Lévy flight, moreover, bird eggs will not taken into consideration until the next generation, not eliminate the eggs can retain the diversity of the searching way of next generation and increase the chances of finding the global optimal solution.

3 Research scheme

Owing to the high frequency, non-stationary and chaotic properties of the stock index data, a stock price forecasting model utilizing the original stock index data fails to provide satisfying forecast results. To solve this problem, before constructing a forecasting model, many studies would first utilize an information extraction technique to extract features contained in data, then use these extracted characteristics to construct the forecasting model. The following is the illustration of our proposed EPC method.

According to the previous study, it is inevitably produce certain error which lead to the unsatisfactory results. Thus, we assume that if the error can effectively predicted, then the model can produce higher precision of prediction through feedback the error prediction results, using which to modify the preliminary results. Therefore, this section proposes our synchronization error prediction idea based on the hybrid of SVM, EMD and ICS.

As shown in Fig.5, the procedure of the error-correction forecasting can be expressed as follows:

Step 1 Data preprocessing. For a known stock price time series $\{x_t, t=1, 2, \cdots, n\}$, it needs to reconstruct the data set space in order to satisfy the prerequisite of SVM analysis,

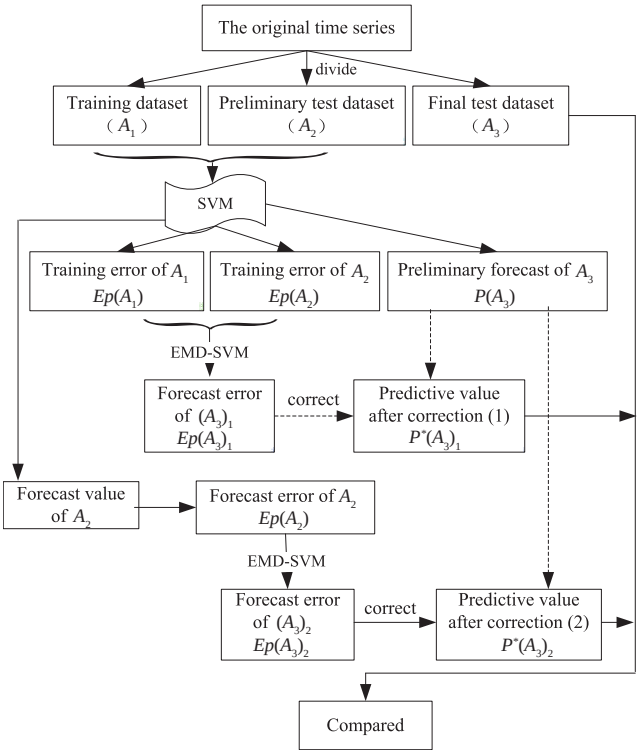


Figure 5 The procedure of the EPC method

that is transform the time series into matrix form and construct sample (X_t, Y_t) where $X_t = \{x_{t-m}, x_{t-m+1}, \cdots, x_{t-1}\}$ and $Y_t = x_t$. Set m for the sliding time window size which represents using the first m days trading price to predict the $(m+1)^{\text{th}}$ day's price. In order to use the existing data effectively, it is necessary to do segmentation first. According to different purposes of each stage, we divide the raw sequence into training dataset A_1, preliminary test dataset A_2 and final test dataset A_3, then respectively reconstruct them based on the above space reconstruction principle.

Step 2 Preliminary forecast of the original sequence. On condition that the dataset A_1 and A_2 as the training samples, predict the value $P(A_3)$ of dataset A_3 based on the SVM method. $P(A_3)$ is the preliminary forecast value which shall be corrected by the predictive error in the following steps. In terms of the preliminary forecasting process with SVM model, for the stock indices are subject to many factors of influence, we can not expect to achieve good results by applying single factor indictor, therefore, by reviewing of domain experts and prior research, we choose 4 technical indicators as the input variables. Table 1 shows these technical indictors and formulas.

Step 3 Predict the error sequence by EMD-SVM. For the error sequence has high frequency, non-stationary and chaotic properties, we decompose the error sequence into several IMFs with different time scales. On the basis of this, choosing different kernel functions and parameters for each IMF and make predictions respectively, finally obtained the predictive results by summing up each IMF's predictive result. Actually forecasting the error in A_3 dataset can be realized in two ways:

Table 1 Initial input features and their formulars

Feature name	formula	Refs
CCI (Commodity Channel Index)	$\frac{M_t - SM_t}{0.015 \times D_t}$	[52–54]
RSI (Relative Strength Index)	$100 - \frac{100}{1+(\sum_{i=0}^{n-1} Up_{t-i}/n)/(\sum_{i=0}^{n-1} Dw_{t-i}/n)}$	[52, 53]
CPP (Current Price Position)	$\frac{1}{1+e^{-(C_t - MA_{t-1,t-i}/MA_{t-1,t-i}) \times 100}}$	[55]
ROC (Price Rate-of-Change)	$\frac{C_t}{C_{t-n}} \times 100$	[56]

Note: C_t is the closing price at time t, L_t is the low price, H_t is the high price, $M_t = (H_t + L_t + C_t)/3$, $SM_t = (\sum_{i=1}^{n} M_{t-i+1})/n$, $D_t = (\sum_{i=1}^{n} |M_{t-i+1} - SM_t|)/n$, Up_t means upward-price-change and Dw_t means downward-price-change, MA_t is the moving average of t days.

Method a. Using the training errors of A_1 and A_2 set as the training samples, predict the error values in A_3 and get the sequence results $Ep(A_3)_1$;

Method b. Firstly use the data of A_1 set as the training samples and get the predictive values of A_2 set, then utilize the $Ep(A_2)$ calculated by using the real value minus the predict value as the training samples and get the predictive error values $Ep(A_3)_2$ in A_3 set.

We called Method a as forecast error and correct the initial predictive result as "Training error prediction & correction (TEPC)"; Method b as "Forecast error prediction & correction (FEPC)".

Step 4 Final data prediction. Respectively utilize the predictive error value $Ep(A_3)_1$ and $Ep(A_3)_2$ of A_3 set to correct $P(A_3)$ and get the corrected results $P^*(A_3)_1$ and $P^*(A_3)_2$.

One important issue should be considered is determining the parameters in setting up SMV model. It should be noticed that the parameters determination in our research is with the purposed of optimized by the ICS algorithm. The parameter especially including penalty function C and kernel function g which have decisive influence to the prediction accuracy and generalization ability, thus, in order to assure the effectiveness of selected parameters, we use several methods including single cuckoo search (CS), the grid search (GS), genetic algorithm (GA) and particle swarm optimization (PSO) as the comparing groups for our proposed ICS algorithm. All the algorithms are realized on MATLAB software (R2012 version). To control the ICS, we set the raw bird number $n = 5$, the $Pop_{\max} = 10$, $egg_i \in [2, 5]$, $Var_{\max} = 0.5$ and $Var_{\min} = 0$.

4 Empirical study

4.1 Datasets and performance criteria

To evaluate the performance of the proposed forecasting model, two stock market indexes (SSE Composite Index of China (SSEC) and National Association of Securities Dealers Automate Quotation (NASDAQ)) are used herein. All of the data collected in this study are cash closing indexes. The time period for each closing index is summarized and shown in Fig.6 and Fig.10. There are total of 437 data points for SSEC and 466 data points for NSDAQ. The first 280 data points are utilized as the training samples (A_1), the following 100 data points are used as the preliminary testing samples (A_2), while the remaining data points are used as the final testing samples (A_3). For the sake of identify the 3 dataset mentioned above easily, we use different color to mark them, the green curve reflects training dataset which has 280 data

points, the cyan curve reflects preliminary test dataset which has 100 data points and the blue curve reflects the final test dataset which has the remaining data points. The meaning of these three colors stays the same in the flowing process, while we use the red color curve reflects the predict data series.

To certify the performance of our proposed error-correction forecasting method, the forecasting results of the proposed model are compared to the BP neural network, the integrated wavelet-network, the single SVM, single ARIMA, single ANFIS, meanwhile, in order to compare the effectiveness of ICS, it also introduced grid search method (GS), some other evolutionary algorithms like genetic algorithm (GA), particle swarm optimization algorithm (PSO) as the comparative methods in the following study.

The metrics of forecasting performance are utilizing the root mean square error (RMSE), the mean absolute percentage error (MAPE) and the mean absolute error (MAE). Table 3 reflects the numerical results of these three metrics which being used to measure the deviation between real data and predict data of SSEC index and NSDAQ index in our empirical study.

4.2 SSEC index

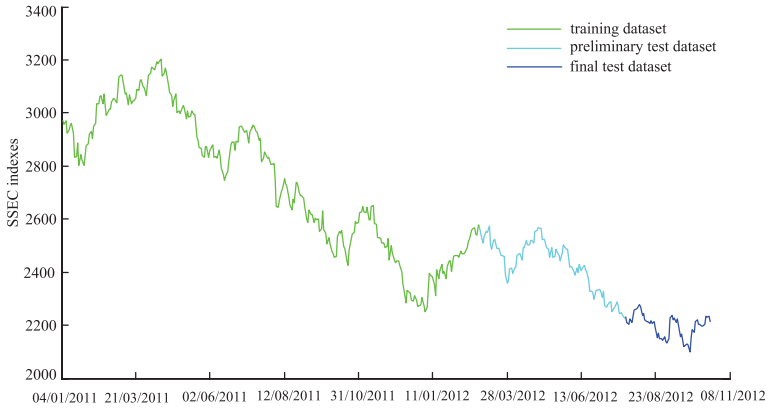


Figure 6 The daily SSEC closing indexes from 04/01/2011 to 22/10/2012

According to the basic process in Fig.5, the preliminary forecast result is shown in Fig.7. Compared with the real data, the predicted data can basically reflects the volatility pattern of SSEC, while there are two significant problems: first, the prediction curve has obvious time lag

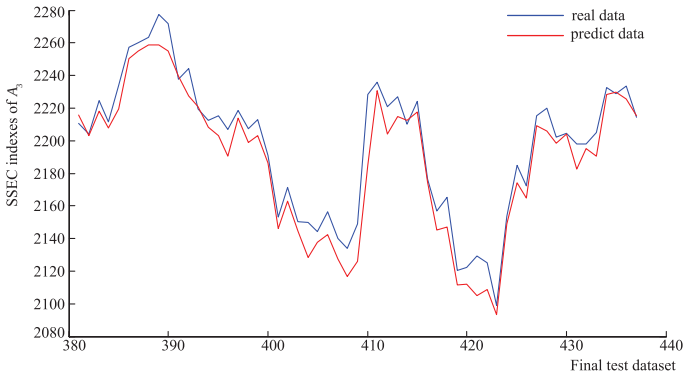
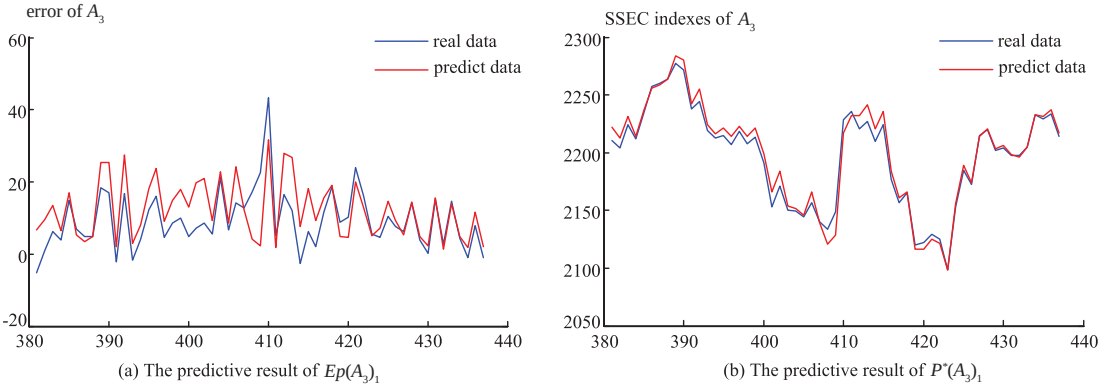


Figure 7 The preliminary predictive result of SSEC through single SVM

with the real data curve; second, there are big errors at the inflection points. That means through the single SVM model can hardly get satisfying results.

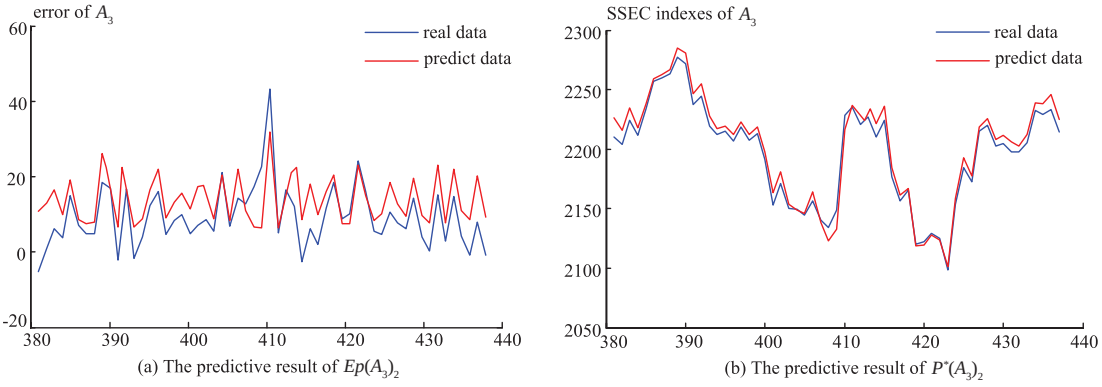
As mentioned in Section 3, there are 2 ways to set up simultaneous error-prediction forecasting model, TEPC and FEPC. For TEPC, it needs to decompose the training error sequence in A_1 and A_2 data set, as shown in Fig.(a) of Appendix, the green curve means sequences have same time span with A_1 dataset, the cyan curve means sequences have same time span with A_2 dataset. The training error sequence $TE(A_1, A_2)$ can finally be decomposed into 7 IMFs and 1 overall trend. Taking the best C and g selected by ISC algorithm (shown in Table 2), set up SVM model and get the predictive value of each IMF, integrate all of them then get the final predictive error $Ep(A_3)_1$, as shown left in Fig.8. Use the $Ep(A_3)_1$ to correct the preliminary predictive result $P(A_3)$ and get the result of $P^*(A_3)_1$ which illustrated right in Fig.8, the red curve means the predict data which is the final predictive result of SSEC through TEPC method.



(a) The predictive result of $Ep(A_3)_1$ (b) The predictive result of $P^*(A_3)_1$

Figure 8 The $Ep(A_3)_1$ and $P^*(A_3)_1$ of SSEC through TEPC

From the Fig.8 we can clearly see, through the correction by the predictive error, the two obvious problems mentioned above have been largely improved compared with the preliminary results. The metrics of forecasting performance results are summarized in Table 3.



(a) The predictive result of $Ep(A_3)_2$ (b) The predictive result of $P^*(A_3)_2$

Figure 9 The $Ep(A_3)_2$ and $P^*(A_3)_2$ of SSEC through FEPC

Similarly, for FEPC, it is also need to decompose the error sequence, while the error sequence $Ep(A_2)$ is calculated by the real data minus the predictive data produced by the SVM model

using the A_1 dataset as training samples, therefore, the $Ep(A_2)$ has the same time span of A_2 and being marked with cyan. The EMD results is shown in Fig.(b) of Appendix, the $Ep(A_2)$ finally decomposed into 5 IMFs and 1 overall trend. The number of components is different from the $TE(A_1, A_2)$ that mainly because the $TE(A_1, A_2)$ has 280 data points which is more than $Ep(A_2)$'s 100. Same procedure with the TEPC, select the best C and g for each IMF, get the predictive results of the error sequence $Ep(A_3)_2$ and final predictive results $P^*(A_3)_2$ of SSEC. The parameter selection results are summarized in Table 2 and the final predictive results are shown in Fig.9, from which we can easily distinguish that the predictive results has obviously improved by FEPC. However, compared with the results in Fig.8, it is hardly to distinguish which has higher accuracy, thus, by summarize the metrics of forecasting performance results in Table 3, we can infer that the TEPC is slightly better than FEPC.

Table 2 The parameter selecting results for each IMF through 5 different methods (SSEC)

Error sequence	Compo-nents	MSE (fitness value) / Evolution steps					Best C	Best g
		GS	GA	PSO	CS	ICS		
$TE(A_1, A_2)$	Imf1	0.00224/-	0.0056/521	0.04187/354	0.00876/652	**0.00163/267**	2^{-13}	2^{-5}
	Imf2	0.00614/-	0.0121/502	0.05102/407	0.00456/552	**0.00187/255**	2^{-12}	2^{-4}
	Imf3	0.00216/-	0.0133/531	0.06351/442	0.00823/660	**0.00192/243**	2^{-12}	2^{-7}
	Imf4	0.00457/-	0.0154/487	0.03123/356	0.01157/621	**0.00201/299**	2^{-11}	2^{-7}
	Imf5	0.00556/-	0.0182/466	0.02230/457	0.01354/557	**0.00262/247**	2^{-11}	2^{-5}
	Imf6	0.00354/-	0.0155/454	0.01190/401	0.00567/563	**0.00207/156**	2^{-9}	2^{-4}
	Imf7	0.00196/-	0.0144/375	0.03231/388	0.09657/489	**0.00135/132**	2^{-10}	2^{-4}
	r	0.00247/-	0.0137/331	0.01232/424	0.02015/587	**0.00122/94**	2^{-9}	2^{-6}
$Ep(A_2)$	Imf1	0.00190/-	0.0166/498	0.03319/399	0.00878/552	**0.00148/239**	2^{-7}	2^{-8}
	Imf2	0.00233/-	0.0202/513	0.02037/446	0.00954/521	**0.00189/242**	2^{-7}	2^{-4}
	Imf3	0.00201/-	0.0130/487	0.04155/504	0.02247/483	**0.00163/201**	2^{-8}	2^{-6}
	Imf4	0.00183/-	0.0187/466	0.01162/397	0.01822/590	**0.00122/114**	2^{-10}	2^{-4}
	r	0.00126/-	0.0140/371	0.03101/362	0.00650/532	**0.00105/88**	2^{-9}	2^{-4}

From Table 2 we can see that ICS has better fitness value while lesser evolution steps, indicating that the ICS method has great adaptability in SVM model's parameter selection.

4.3 NASDAQ index

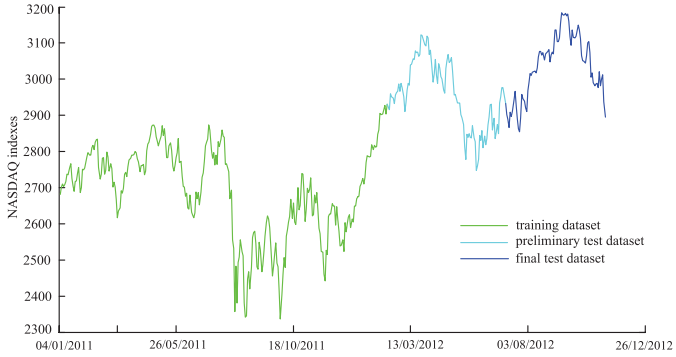


Figure 10 The daily NASDAQ closing indexes form 04/01/2011 to 08/11/2012

The forecasting procedure of NASDAQ index is similar with that of SSEC index illustrated in Section 4.2. Respectively, Fig.11∼Fig.13 show the results of the forecasting process of NASDAQ and the metrics of forecasting performance are summarized in Table 3.

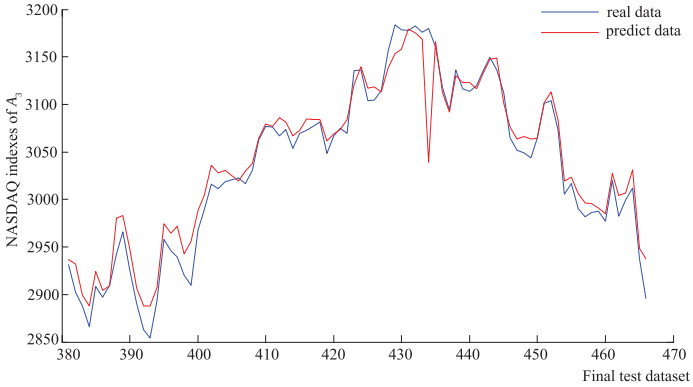


Figure 11 The preliminary predictive result of NASDAQ through single SVM

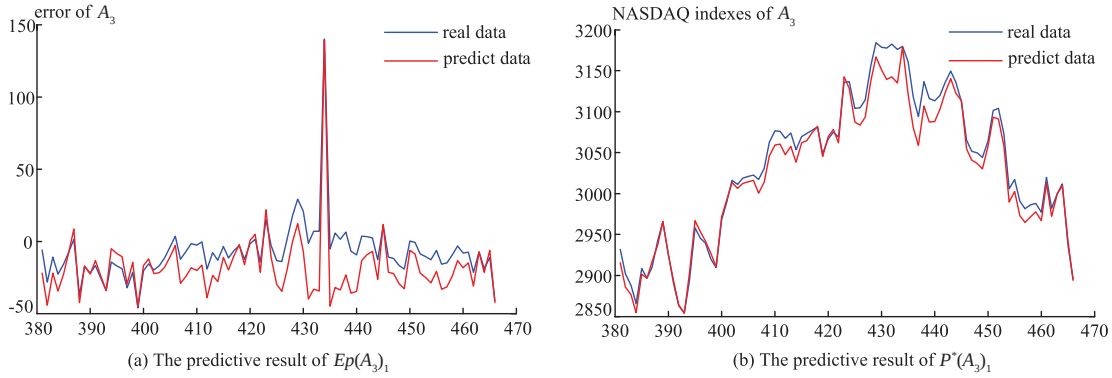


Figure 12 The $Ep(A_3)_1$ and $P^*(A_3)_1$ of NASDAQ through TEPC

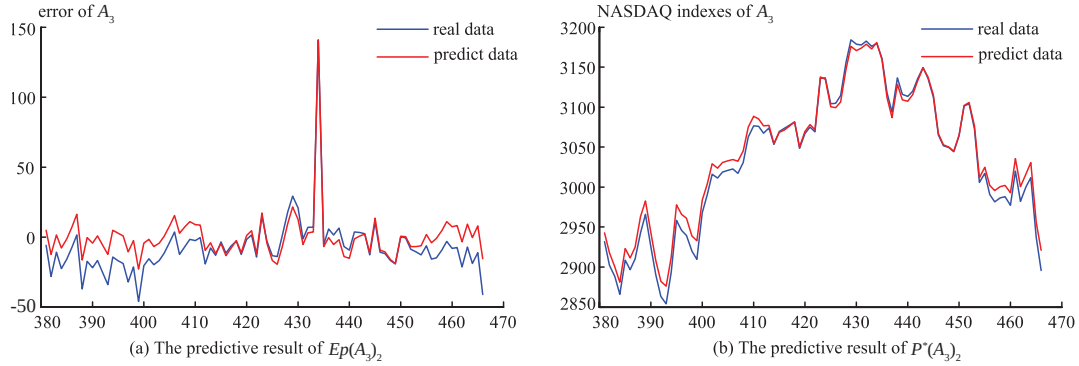


Figure 13 The $Ep(A_3)_2$ and $P^*(A_3)_2$ of NASDAQ through FEPC

Judging from Fig.11 we can see that the single SVM model failed to produce satisfying predictive results of NASDAQ index as it did in SSEC index. Because the NASDAQ index is in much more mature market which means it may has properties of much higher frequency,

more chaotic, etc. Therefore, it seems much more necessarily handing the large error in order to better the predictive results. The following process shows our proposed method can greatly improve the accuracy of the predictive results of NASDAQ index as it did for SSEC.

Firstly, using the TEPC, we get the training error sequence through the single SVM model. The $TE(A_1, A_2)$ of the NASDAQ index has similar characteristics with SSEC, it can be decomposed into 7 IMFs and 1 overall trend, the pattern of which is illustrated in Fig.(c) of Appendix. Taking each IMF using different penalty function C and kernel function g selected by ICS algorithm (summarized in Table 4) respectively and set SVM model for each IMF and get the final predictive error $Ep(A_3)_1$ which being used to correct the preliminary predictive result $P(A_3)$ and finally get the result of $P^*(A_3)_1$. As being illustrated in Fig.12, the red curve means the predict data which is the final predictive result of NASDAQ through TEPC method. Through the correction by the predictive error, the two obvious problems mentioned above have also largely improved compared with the preliminary results. That means the TEPC method is also effective for the predicting of NASDAQ.

Then using the FEPC, decompose the error sequence $Ep(A_2)$ and get another series of components, as shown in Fig.(d) of Appendix, the $Ep(A_2)$ being decomposed into 5 IMFs and 1 overall trend, because the $Ep(A_2)$ has same time span with A_2 dataset, it marked with cyan as well. Repeat the above forecasting process for each IMF and finally get the results as shown in Fig.13. The parameter selection results are summarized in Table 4 and the metrics of forecasting performance are summarized in Table 3.

Table 3 Summary of forecast results of SSEC and NASDAQ

Indexes	Model		RMSE	MAPE	MAE
SSEC	BP-neural network		81.2453	0.569%	12.3454
	Wavelet-neural network		53.2455	0.343%	7.4363
	Single SVM		73.7306	0.450%	9.7659
	Single ARIMA		75.5647	0.475%	10.3131
	Single ANFIS		77.5780	0.516%	11.2013
	EPC	**TEPC**	**42.7637**	**0.261%**	**5.6642**
		FEPC	43.2455	0.264%	5.7546
NASDAQ	BP-neural network		152.3786	0.529%	15.5248
	Wavelet-neural network		101.4542	0.384%	11.2415
	Single SVM		129.5699	0.463%	13.9719
	Single ARIMA		132.5453	0.490%	14.3580
	Single ANFIS		135.2486	0.501%	14.6854
	EPC	TEPC	89.4576	0.336%	9.8453
		FEPC	**89.3404**	**0.321%**	**9.6338**

Comparing the results in Fig.12 with Fig.13, it can be roughly judged that the FEPC is somewhat better than the TEPC. For further proofs, through the metrics in Table 3, we can see that the FEPC has smaller deviation, which means higher accuracy. By now, for the SSEC index forecasting, the TEPC is slightly better than the FEPC, while for the NASDAQ index

forecasting, the FEPC has better performance. Therefore, it is hard for us to judge the two ways of EPC method, TEPC and FEPC, which one is better.

4.4 Robustness evaluation

To evaluate the robustness of the proposed method, we test all the methods mentioned in Table 3 under different ratios of training sample sizes as well as the testing sample. Because the data points in our sample are in small amount, we make the training and testing sample ratios in 75%, 80%, 85%, 90% and 95%. The forecasting results for the SSEC and NASDAQ under the 6 ratios in different methods are summarized in Table 5. It can be observed that the proposed EPC method outperforms the other bench marking tool under the different ratios in terms of the 3 performance measures. The results show that the EPC method indeed provides better forecast accuracy.

Table 4 The parameter selecting results for each IMF through 5 different methods (NSDAQ)

Error sequence	Compo-nents	MSE (fitness value) / Evolution steps					Best C	Best g
		GS	GA	PSO	CS	ICS		
$TE(A_1, A_2)$	Imf1	0.00463/-	0.0083/552	0.03233/430	0.01032/641	**0.00155/331**	2^{-11}	2^{-5}
	Imf2	0.00314/-	0.0096/521	0.06217/421	0.01223/652	**0.00253/320**	2^{-11}	2^{-4}
	Imf3	0.00626/-	0.0101/493	0.07324/439	0.00975/633	**0.00312/275**	2^{-12}	2^{-6}
	Imf4	0.00433/-	0.0166/487	0.03235/384	0.02087/589	**0.00330/364**	2^{-11}	2^{-5}
	Imf5	0.00514/-	0.0172/483	0.03088/473	0.01723/566	**0.00289/298**	2^{-10}	2^{-4}
	Imf6	0.00402/-	0.0138/477	0.01022/412	0.02389/602	**0.00257/201**	2^{-9}	2^{-4}
	Imf7	0.00188/-	0.0169/431	0.03731/392	0.08323/511	**0.00188/198**	2^{-10}	2^{-3}
	r	0.00263/-	0.0137/396	0.01845/413	0.06312/602	**0.00161/113**	2^{-9}	2^{-6}
$Ep(A_2)$	Imf1	0.00450/-	0.0183/511	0.03417/378	0.01025/611	**0.00170/302**	2^{-7}	2^{-8}
	Imf2	0.00373/-	0.0197/523	0.02368/436	0.01756/586	**0.00303/312**	2^{-7}	2^{-2}
	Imf3	0.00262/-	0.0203/485	0.03255/494	0.03440/503	**0.00189/287**	2^{-11}	2^{-6}
	Imf4	0.00192/-	0.0136/463	0.01870/402	0.02314/632	**0.00152/165**	2^{-10}	2^{-4}
	r	0.00138/-	0.0095/363	0.03361/377	0.02001/571	**0.00143/102**	2^{-9}	2^{-3}

In the robustness evaluation research, the FEPC has better performance than TEPC on the whole, the reason can be explained that, for the SVM model, it is set up by train the training data sample before the forecasting, the training error in A_1 and A_2 time span is caused by our known samples, while the predictive error in A_2 time span, forecasted by the training error in A_1, is caused by unknown samples which is much more similar with the predictive error in A_3 time span, that means they have analogical regularity and feature.

5 Conclusion

This paper proposed a simultaneous error prediction method (EPC) for stock index forecasting by integrating the empirical mode decomposition (EMD), support vector machine (SVM) and improved cuckoo search algorithm (ICS). In terms of the cuckoo search algorithm, for the

Table 5 Robustness evaluation

Ratio	Model		SSEC			NASDAQ		
			RMSE	MAPE	MAE	RMSE	MAPE	MAE
75%	BP-neural network		90.3210	0.6133%	14.0154	167.4531	0.550%	18.9915
	Wavelet-neural network		60.2154	0.370%	8.4663	122.4560	0.464%	13.7505
	Single SVM		80.5740	0.524%	11.9784	143.1057	1.0255%	16.0312
	Single ARIMA		81.3540	0.524%	12.6754	144.3581	0.486%	16.7740
	Single ANFIS		83.1461	0.555%	13.8451	148.7543	0.518%	17.8752
	EPC	TEPC	50.3458	0.281%	6.4245	98.6417	0.351%	12.1241
		FEPC	**49.6482**	**0.280**%	**6.4010**	**98.5378**	**0.348**%	**12.0012**
80%	BP-neural network		87.3241	0.602%	13.8512	162.3054	0.545%	18.6420
	Wavelet-neural network		58.3215	0.359%	8.2460	118.1057	0.386%	13.2004
	Single SVM		77.6844	0.515%	11.8454	139.0354	0.461%	15.7566
	Single ARIMA		79.2354	0.524%	12.0488	140.2778	0.475%	16.2521
	Single ANFIS		82.3452	0.574%	13.1864	145.5420	0.506%	17.3101
	EPC	**TEPC**	**47.6442**	**0.266**%	**6.1258**	**95.6121**	**0.344**%	**11.7754**
		FEPC	48.1245	0.277%	6.3782	95.7080	0.351%	12.0138
85%	BP-neural network		85.5624	0.588%	12.7821	159.3742	0.537%	16.3421
	Wavelet-neural network		56.9850	0.362%	7.8641	115.6547	0.396%	12.0425
	Single SVM		75.3204	0.488%	10.6027	136.8579	0.492%	14.9821
	Single ARIMA		77.9056	0.504%	10.9472	137.9540	0.504%	15.3473
	Single ANFIS		79.7856	0.577%	12.5525	140.1275	0.526%	16.0036
	EPC	TEPC	45.6414	0.278%	6.0457	94.6477	0.362%	11.0210
		FEPC	**45.5970**	**0.274**%	**5.9643**	**93.7204**	**0.357**%	**10.8542**
90%	BP-neural network		85.3785	0.581%	12.6541	157.8512	0.534%	15.9123
	Wavelet-neural network		56.5432	0.351%	7.6420	112.2306	1.0255%	0.04187
	Single SVM		74.6241	0.460%	10.0113	134.3412	0.494%	14.2304
	Single ARIMA		77.5614	0.492%	10.7054	134.4021	0.507%	14.7125
	Single ANFIS		78.4520	0.533%	11.6121	137.3342	1.0255%	15.1004
	EPC	TEPC	44.0274	0.274%	5.9642	91.6801	0.346%	10.3246
		FEPC	**43.8720**	**0.270**%	**5.8767**	**91.5902**	**0.340**%	**10.1206**
95%	BP-neural network		83.3241	0.573%	12.3579	154.2472	0.534%	15.8437
	Wavelet-neural network		54.5231	0.349%	7.5631	107.0247	0.394%	11.6855
	Single SVM		73.8614	0.457%	9.9216	130.2104	0.475%	14.1021
	Single ARIMA		76.4264	0.488%	10.5852	133.2423	0.494%	14.6682
	Single ANFIS		77.8460	0.529%	11.4672	136.5784	0.502%	14.9031
	EPC	TEPC	43.8214	0.272%	5.8930	90.7764	0.341%	10.1129
		FEPC	**43.3145**	**0.267**%	**5.7873**	**90.5430**	**0.338**%	**10.0357**

defects of lacking search abilities and low accuracy, we proposed several improved strategies, the results show that our improved cuckoo search algorithm has better accuracy and less evolution steps than other 4 bench marking methods. The basic steps are using the SVM model to get a preliminary results first, then use the EMD method to decompose the error sequence which caused in the first step into several IMFs, then set up SVM model for each IMFs and get forecasting results by integrate them, finally use the error predictive to correct the preliminary results. The ICS algorithm play parameters selecting role in the whole research.

For the empirical research, we using one emerging daily stock market index (SSEC) and one mature daily stock market indexes (NASDAQ) as samples. In order to compare the performance of our proposed method, the BP neural network, the wavelet-network, the single SVM, single ARIMA and single ANFIS are used as the reference methods. The empirical research and robustness evaluation results show that our proposed method has a better performance in forecasting the stock index than the 5 reference methods. Moreover, the ICS algorithm has good application prospects in the best parameters selecting, which reflects that it might be used in solving other optimization problems. Future research can aim at combining the EPC idea with other forecasting tools, like neural networks, ARIMA to improve their own forecasting abilities.

References

[1] Pai P F, Lin C S. A hybrid ARIMA and support vector machines model in stock price forecasting. Omega, 2005, 33(6): 497–505.

[2] Wang Y H. Nonlinear neural network forecasting model for stock index option price: Hybrid GJR-GARCH approach. Expert Systems with Applications 2009, 36(1): 564–570.

[3] Chi S C, Chen H P, Cheng C H. A forecasting approach for stock index future using grey theory and neural networks. International Joint Conference on Neural Networks, IJCNN'99, IEEE, 1999, 6: 3850–3855.

[4] Wang J J, Wang J Z, Zhang Z G, et al. Stock index forecasting based on a hybrid model. Omega, 2012, 40(6): 758–766.

[5] Lu C J, Wu J Y, Chiu C C, et al. Predicting stock index using an integrated model of NLICA, SVR and PSO. Advances in Neural Networks, ISNN 2011, Springer, 2011: 228–237.

[6] Kim K J. Financial time series forecasting using support vector machines. Neurocomputing, 2003, 55(1): 307–319.

[7] Chiu D Y, Chen P J. Dynamically exploring internal mechanism of stock market by fuzzy-based support vector machines with high dimension input space and genetic algorithm. Expert Systems with Applications, 2009, 36(2): 1240–1248.

[8] Kao L J, Chiu C C, Lu C J, et al. Integration of nonlinear independent component analysis and support vector regression for stock price forecasting. Neurocomputing, 2013, 99(1): 534–542.

[9] Hsu S H, Hsieh J, Chih T C, et al. A two-stage architecture for stock price forecasting by integrating self-organizing map and support vector regression. Expert Systems with Applications, 2009, 36(4): 7947–7951.

[10] Lee M C. Using support vector machine with a hybrid feature selection method to the stock trend prediction. Expert Systems with Applications, 2009, 36(8): 10896–10904.

[11] Madsen H, Skotner C. Adaptive state updating in real-time river flow forecasting? A combined filtering and error forecasting procedure. Journal of Hydrology, 2005, 308(1): 302–312.

[12] Yao Y, Lian Z, Hou Z, et al. An innovative air-conditioning load forecasting model based on RBF neural network and combined residual error correction. International Journal of Refrigeration, 2006, 29(4): 528–538.

[13] Orrell D, Smith L, Barkmeijer J, et al. Model error in weather forecasting. Nonlinear Processes in Geophysics, 2001, 8(6): 357–371.

[14] Allen M R, Kettleborough J, Stainforth D. Model error in weather and climate forecasting. ECMWF

Predictability of Weather and Climate Seminar, European Centre for Medium Range Weather Forecasts, Reading, UK, http://www. ecmwf. int/publications/library/do/references/list/209, 2002.

[15] Zhou M, Yan Z, Ni Y X, et al. A novel arima approach on electricity price forecasting with the improvement of predicted error. Proceedings of the CSEE, 2004, 12: 013.

[16] Chen A S, Leung M T. Regression neural network for error correction in foreign exchange forecasting and trading. Computers & Operations Research, 2004, 31(7): 1049–1068.

[17] Anderson R G, Hoffman D L, Rasche R H. A vector error-correction forecasting model of the US economy. Journal of Macroeconomics, 2002, 24(4): 569–598.

[18] Kao L J, Chiu C C, Lu C J, et al. A hybrid approach by integrating wavelet-based feature extraction with MARS and SVR for stock index forecasting. Decision Support Systems, 2013, 54(3): 1228–1244.

[19] Chang P C, Fan C Y. A hybrid system integrating a Wavelet and TSK fuzzy rules for stock price forecasting. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 2008, 38(6): 802–815.

[20] Lu C J, Lee T S, Chiu C C. Financial time series forecasting using independent component analysis and support vector regression. Decision Support Systems, 2009, 47(2): 115–125.

[21] Zhu Z, Sun Y, Li H. Hybrid of EMD and SVMs for short-term load forecasting. IEEE International Conference on Control and Automation, 2007: 1044–1047.

[22] Yu L A, Lai K K, Wang S Y, et al. Oil price forecasting with an EMD-based multiscale neural network learning paradigm. Computational Science-ICCS 2007, Springer, 2007: 925–932.

[23] Lin A, Shang P, Feng G, et al. Application of empirical mode decomposition combined with k-nearest neighbors approach in financial time series forecasting. Fluctuation and Noise Letters, 2012, 11(2): 1–14.

[24] Yu L A, Wang S Y, Lai K K. Financial crisis modeling and prediction with a Hilbert-EMD-based SVM approach. Intelligent Data Analysis: Developing New Methodologies Through Pattern Discovery and Recovery, 2009: 286–299.

[25] Yu L A, Wang S Y, Lai K K. Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm. Energy Economics, 2008, 30(5): 2623–2635.

[26] Lin C S, Chiu S H, Lin T Y. Empirical mode decomposition-based least squares support vector regression for foreign exchange rate forecasting. Economic Modelling, 2012, 29(6): 2583–2590.

[27] Nguyen T, Gordon-Brown L, Wheeler P, et al. GA-SVM based framework for time series forecasting. Fifth International Conference on Natural Computation, ICNC'09, IEEE, 2009, 1: 493–498.

[28] Yuan F C. Parameters optimization using genetic algorithms in support vector regression for sales volume forecasting. Applied Mathematics, 2012, 30(3): 1480–1486.

[29] Wu C H, Tzeng G H, Goo Y J, et al. A real-valued genetic algorithm to optimize the parameters of support vector machine for predicting bankruptcy. Expert Systems with Applications, 2007, 32(2): 397–408.

[30] Abolhassani A M, Yaghoobi M. Stock price forecasting using PSOSVM. Advanced Computer Theory and Engineering (ICACTE), 2010 3rd International Conference on, IEEE, 2010, V3: 352.

[31] Lu C J, Wu J Y, Chiu C C, et al. Predicting stock index using an integrated model of NLICA, SVR and PSO. Advances in Neural Networks-ISNN 2011, Springer, 2011: 228–237.

[32] Yang X S, Deb S. Cuckoo search via Lévy flights. World Congress on Nature & Biologically Inspired Computing, NaBIC 2009, IEEE, 2009: 210–214.

[33] Gandomi A H, Yang X S, Alavi A H. Cuckoo search algorithm: A metaheuristic approach to solve structural optimization problems. Engineering with Computers, 2013, 29(1): 17–35.

[34] Walton S, Hassan O, Morgan K, et al. Modified cuckoo search: A new gradient free optimisation algorithm. Chaos, Solitons & Fractals, 2011, 44(9): 710–718.

[35] Yang X S, Deb S. Engineering optimisation by cuckoo search. International Journal of Mathematical Modelling and Numerical Optimisation1, 2010(4): 330–343.

[36] Cortes C, Vapnik V. Support-vector networks. Machine Learning, 1995, 20(3): 273–297.

[37] Huang N E, Shen Z, Long S R, et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of the Royal Society of London, Series A: Mathematical, Physical and Engineering Sciences, 1998, 454(1971): 903–995.

[38] Huang N E, Shen Z, Long S R. A new view of nonlinear water waves: The Hilbert spectrum 1. Annual Review of Fluid Mechanics, 1999, 31(1): 417–457.

[39] Wu Z, Huang N E. Ensemble empirical mode decomposition: A noise-assisted data analysis method. Ad-

vances in Adaptive Data Analysis, 2009, 1(1): 1–41.

[40] Boudraa A, Cexus J, Saidi Z. EMD-based signal noise reduction. International Journal of Signal Processing, 2004, 1(1): 33–37.

[41] De Ramirez S S, Enquobahrie D, Nyadzi G, et al. Prevalence and correlates of hypertension: A cross-sectional study among rural populations in sub-Saharan Africa. Journal of Human Hypertension, 2010, 24(12): 786–795.

[42] Payne R B, Sorensen M D. The cuckoos. OUP Oxford, 2005, 15.

[43] Wheatcroft D J. Co-evolution: A behavioral 'spam filter' to prevent nest parasitism. Current Biology, 2009, 19(4): R170–R171.

[44] Viswanathan G, Afanasyev V, Buldyrev S V, et al. Lévy flights search patterns of biological organisms. Physica A: Statistical Mechanics and its Applications, 2001, 295(1): 85–88.

[45] Reynolds A. Cooperative random Lévy flight searches and the flight patterns of honeybees. Physics Letters A, 2006, 354(5): 384–388.

[46] Reynolds A M, Frye M A. Free-flight odor tracking in drosophila is consistent with an optimal intermittent scale-free search. PloS ONE, 2007, 2(4): e354.

[47] Ramos-Fernández G, Mateos J L, Miramontes O, et al. Lévy walk patterns in the foraging movements of spider monkeys (ateles geoffroyi). Behavioral Ecology and Sociobiology, 2004, 55(3): 223–230.

[48] Schreier A L, Grove M. Ranging patterns of hamadryas baboons: Random walk analyses. Animal Behaviour, 2010, 80(1): 75–87.

[49] Da Luz M, Buldyrev S V, Havlin S, et al. Improvements in the statistical approach to random Lévy flight searches. Physica A: Statistical Mechanics and its Applications, 2001, 295(1): 89–92.

[50] Reynolds A, Rhodes C. The Lévy flight paradigm: Random search patterns and mechanisms. Ecology, 2009, 90(4): 877–887.

[51] Rajabioun R. Cuckoo optimization algorithm. Applied Soft Computing, 2011, 11(8): 5508–5518.

[52] Kim K J. Financial time series forecasting using support vector machines. Neurocomputing, 2003, 55(1): 307–319.

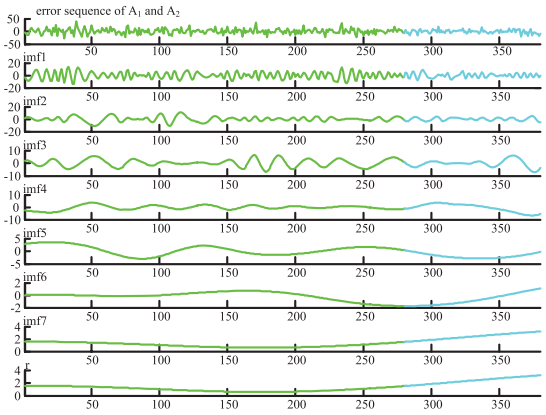
[53] Achelis S B. Technical analysis from A to Z. McGraw Hill New York, 2001.

[54] Chang J, Jung Y, Yeon K, et al. Technical indicators and analysis methods. Seoul: Jinritamgu Publishing, 1996.

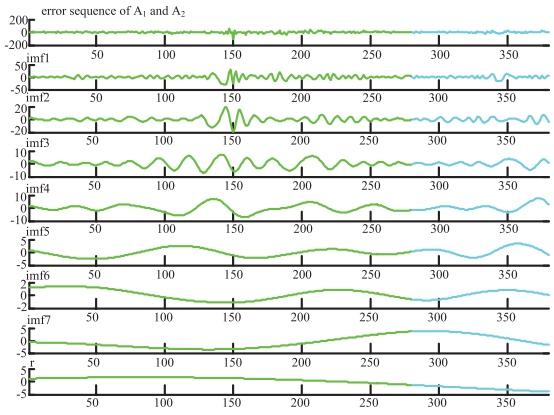
[55] Lee S H, Lim J S. Kospi time series analysis using neural network with weighted fuzzy membership functions. Agent and Multi-Agent Systems: Technologies and Applications, Springer, 2008: 53–62.

[56] Murphy J J. Technical analysis of the financial markets: A comprehensive guide to trading methods and applications. Prentice Hall Press, 1999.

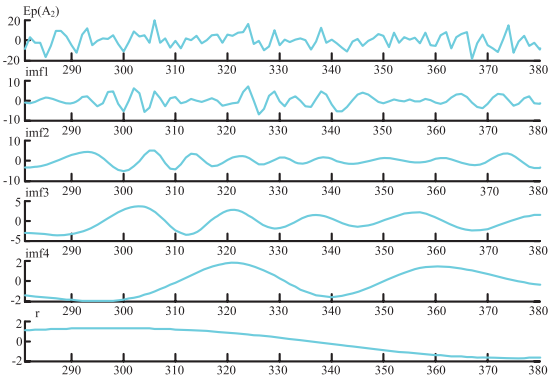
Appendix EMD results of the two indexes



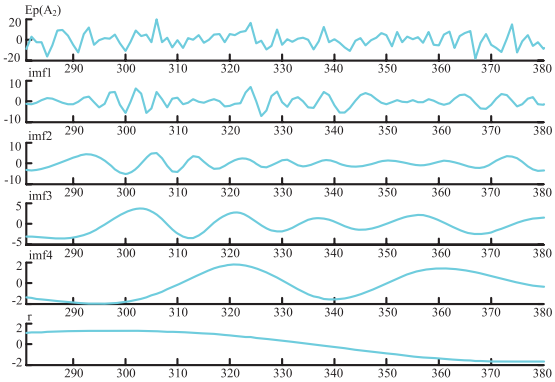
(a) The EMD of error sequence in A_1 and A_2 dataset (SSEC)



(c) The EMD of error sequence in A_1 and A_2 dataset (NASDAQ)



(b) The EMD of $Ep(A_2)$ (SSEC)



(d) The EMD of $Ep(A_2)$ (NASDAQ)

Figure 14 EMD results of the two indexes