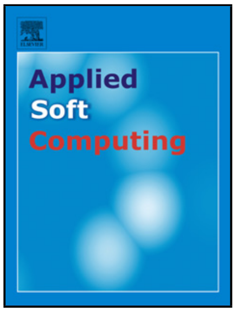
Accepted Manuscript



Title: Stock trend prediction based on a new status box method and AdaBoost probabilistic support vector machine

Author: Xiao-dan Zhang Ang Li Ran Pan

PII: S1568-4946(16)30420-3
DOI: http://dx.doi.org/doi:10.1016/j.asoc.2016.08.026
Reference: ASOC 3768

To appear in: *Applied Soft Computing*

Received date: 15-10-2015
Revised date: 6-7-2016
Accepted date: 12-8-2016

Please cite this article as: Xiao-dan Zhang, Ang Li, Ran Pan, Stock trend prediction based on a new status box method and AdaBoost probabilistic support vector machine, Applied Soft Computing Journal http://dx.doi.org/10.1016/j.asoc.2016.08.026

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Stock trend prediction based on a new status box method and AdaBoost probabilistic support vector machine

Xiao-dan Zhang[a,*], Ang Li[a], Ran Pan[b]

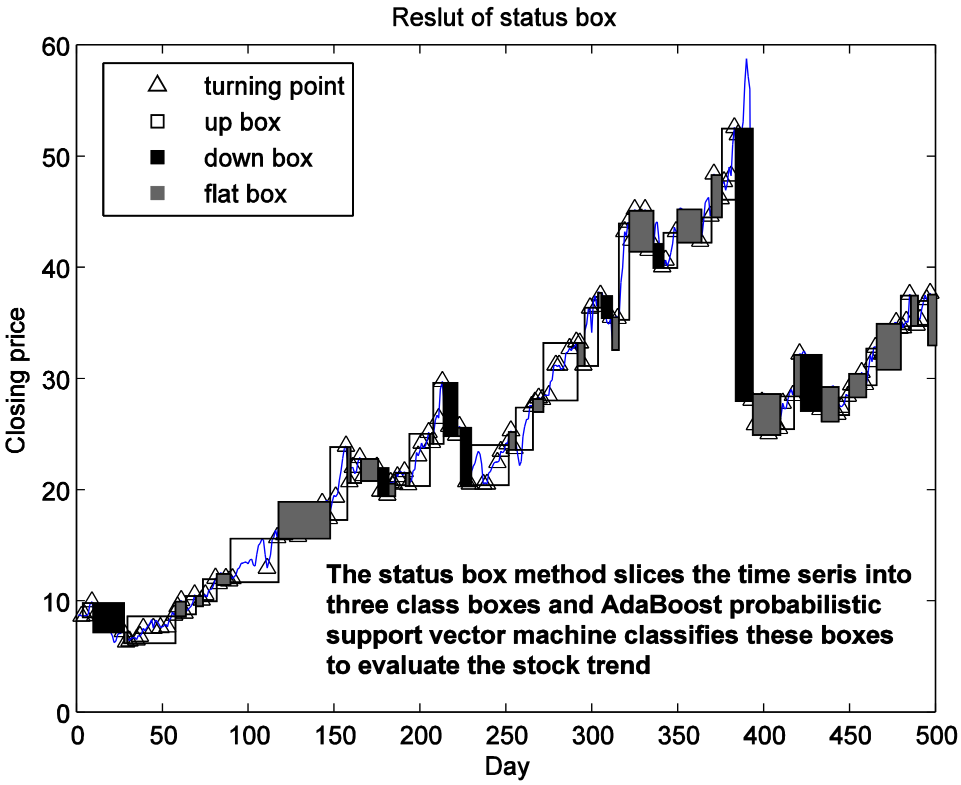
[a.] *School of Mathematics and Physics, University of Science and Technology Beijing, Xueyuan Road No.30 Haidian District, Beijing, China*

[b] *School of mechanical Engineering, Imperial College London, London, UK*

[*] Corresponding author. Tel.: +86 010-62332589.
E-mail addresses: bkdzxd@163.com (Xiao-dan Zhang), siwang744@gmail.com (Ang Li), r.pan13@imperial.ac.uk (Ran Pan)

ACCEPTED MANUSCRIPT

Graphical Abstract



ACCEPTED MANUSCRIPT

Highlights

- A new status box method is proposed to perform the stock trend prediction.
- A new feature construction approach for status box is presented.
- A new hybrid approach that integrates AdaBoost algorithm, genetic algorithm and probabilistic support vector machine is constructed to solve the classification problem of the status boxes.
- The status box method not only have the better classification accuracy but also effectively solve the unbalance problem of the stock turning points classification.

ACCEPTED MANUSCRIPT

Abstract

Stock trend prediction is regarded as one of the most challenging tasks of financial time series prediction. Conventional statistical modeling techniques are not adequate for stock trend forecasting because of the non-stationarity and non-linearity of the stock market. With this regard, many machine learning approaches are used to improve the prediction results. These approaches mainly focus on two aspects: regression problem of the stock price and prediction problem of the turning points of stock price. In this paper, we concentrate on the evaluation of the current trend of stock price and the prediction of the change orientation of the stock price in future. Then, a new approach named *status box method* is proposed. Different from the prediction issue of the turning points, the status box method packages some stock points into three categories of boxes which indicate different stock status. And then, some machine learning techniques are used to classify these boxes so as to measure whether the states of each box coincides with the stock price trend and forecast the stock price trend based on the states of the box. These results would support us to make buying or selling strategies. Comparing with the turning points prediction that only considered the features of one day, each status box contains a certain amount of points which represent the stock price trend in a certain period of time. So, the status box reflects more information of stock market. To solve the classification problem of the status box, a special features construction approach is presented. Moreover, a new ensemble method integrated with the AdaBoost algorithm, probabilistic support vector machine (PSVM), and genetic algorithm (GA) is constructed to perform the status boxes classification. To verify the applicability and superiority of the proposed methods, 20 shares chosen from Shenzhen Stock Exchange (SZSE) and 16 shares from National Association of Securities Dealers Automated Quotations (NASDAQ) are applied to perform stock trend prediction. The results show that the status box method not only have the better classification accuracy but also effectively solve the unbalance problem of the stock turning points classification. In addition, the new ensemble classifier achieves preferable profitability in simulation of stock investment and remarkably improves the classification performance compared with the approach that only uses the PSVM or back-propagation artificial neural network (BPN).

Keywords: Stock trend prediction, Status box method, Piecewise linear representation, AdaBoost, Probabilistic support vector machine

1. Introduction

Stock price prediction is one of the most important subjects of financial markets. Therefore, plenty of research has been done in this field. However, because of the non-linearity and non-stationarity of the stock market, conventional statistical methods used to forecast stock price are not effective enough [1-3]. Thus, many machine learning techniques, such as the back-propagation artificial neural network (BPN) or support vector machine (SVM), are presented to improve the prediction accuracy [4-11]. These techniques mainly focus on two purposes. One is the prediction of the future price based on the historical prices and the technical indicators [2, 12-13], and the other is the forecasting of the turning points of stock price [14-15]. However, the above research is mainly based on the information of single stock data point. While considering the complexity of the stock market and the noisy of stock data, it is very difficult to forecast stock price or position of turning point according to the features of single point.

ACCEPTED MANUSCRIPT

In order to utilize more information of stock points and obtain the reliable trend of stock price according to the history information, a status box method is proposed. Using this method, some points are packaged in some successive status boxes based on the duration and oscillation of the initial turning points computed by piecewise linear representation (PLR). These status boxes are classified into three categories: up box, down box and flat box, which represent the stock quotation being in the rising trend, falling trend and steady state in different time interval respectively. Each status box contains the points of related trend and the adjacent boxes have different status. In this way, the traditional problems of the stock turning points prediction and the stock price prediction are replaced as a classification problem of the status boxes. In addition, the new named box body features and box tail features are presented. In the feature construction of the status box method, the box body features and box tail features are composed by the trend tracking indexes and the trend reversal indexes, which are the main stock market technical analysis indicators, respectively. Further, the minimum redundancy maximum relevance (mRMR) [16] is used to investigate the importance of these new features and determine which features to be chosen.

The advantage of the status box method is reflected in three aspects. Firstly, it considers the stock trend in a period of time rather than the status of a single point so that the prediction result is less sensitive to the single point noise. Secondly, the number of status boxes decreases significantly compared with stock points. Consequently, the unbalanced problem in turning points prediction is avoided effectively. Thirdly, compared with the feature of the single points, the box body features and box tail features are more comprehensive and reasonable for describing stock.

When getting the status boxes, a hybrid approach that integrates AdaBoost algorithm [17], GA and probabilistic SVM is constructed to solve the classification problem of the status boxes. AdaBoost algorithm is applied to improve the performance of SVM classifier and GA is used to select appropriate parameters. In addition, the probabilistic technology has two advantages: decreasing the classification error of multi-class SVM and providing a posteriori probability output which enables people to make better decisions.

The rest of this paper is structured as follows. Section 2 describes the detailed design of the status box method. Section 3 introduces the hybrid method implementation along with the AdaBoost algorithm, GA, and SVM. Section 4 provides the procedure of the feature construction and feature selection. Section 5 describes the implementation of Models .Then, the experiments and the corresponding analysis are shown in Section 6. Section 7 presents some trading strategies based on the status box method and make an investment simulation. Finally, the conclusion is drawn in Section 8.

2. Box construction

2.1. Acquisition of turning points

The first issue in the construction of status boxes is to obtain the turning points. Due to the high cost of manual labeling, some approaches of information extraction, such as discrete Fourier transforms [18], wavelets [19], symbolic mapping [20], and PLR are applied to calculate the turning points of the time series [14,21-23]. Among these methods, PLR is the most suitable one for the examining of turning points because the joint points between adjacent segments calculated by PLR only focus on the change of the trends [14]. The process of turning points acquisition is described as follows.

Given a time series $T=\{y_1, y_2, ..., y_l\}$, the PLR of T represents the piecewise

ACCEPTED MANUSCRIPT

approximation straight lines, which are described by $T_{PLR}=\{L_1(y_1,y_2,\cdots,y_{t_1})$, $L_2(y_{t_1+1},y_{t_1+2},\cdots,y_{t_2}),\ldots$, $L_k(y_{t_{k-1}+1},y_{t_{k-1}+2},\ldots,y_l)\}$, where t_i is the end time of the *i*th segment, $L_i(y_{t_{i-1}+1},y_{t_{i-1}+2},\ldots,y_{t_i}),\ (1\le i\le k)$ indicates the approximation straight line to $y_{t_{i-1}+1},y_{t_{i-1}+2},\ldots,y_{t_i}$,t_i is generally named as turning point because it indicates a transformation of the movement trends. Accordingly, the other points are called non-turning points.

In this paper, the linear regression is applied to generate the approximation line and the Bottom-up algorithm is used to segment the stock data [23]. The goal of segment is to produce the best representation so that the maximum error of any segment would not exceed the given threshold δ .

The segment result of PLR is impacted by the threshold δ remarkably. The number of the turning points is small when δ is large and vice versa [14]. Because the construction process of the turning points is only the pre-produce of status box method, δ is set by a small value to reserve more turning points. During the construction process of boxes, the turning points will be further processed.

2.2. Construction of status boxes

Given a set of turning points $(t_1,y_{t_1}),(t_2,y_{t_2}),\ldots,(t_n,y_{t_n})$ calculated by PLR, where t_i and y_{t_i} are the time point and the stock price of the *i*th turning point respectively. According to the duration and oscillation of these turning points, we construct some successive boxes classified into three categories. These boxes are named as up box, down box, and flat box based on their status respectively. The up box is a set of time points t in which the stock price y_t is increasing. In contrast, the stock price is at a downward trend in the down box. If the stock price is stable during a time duration i.e. the price only makes fluctuation in a small range, the time interval is defined as a flat box. Each status box contains the different number of stock points based on the box duration and the status of adjacent boxes is distinct according to our regulation.

The status boxes, their corresponding status, and their duration are defined as B_k , S_k , and $D_k\ (k=1,2,\ldots,K)$, respectively, where *K* is the number of boxes and D_k is the width of the box B_k . The status S_k of up, down, and flat box corresponds to 1,-1, and 0 respectively. The maximum time duration and minimum price oscillation, which are used to control the width and height of boxes, are set by $T_{\max}$ and P , respectively. The detailed process of status box construction is shown in the **Fig.1.** After getting the status boxes, we will choose some appropriate classifiers to perform classification for these boxes.

2.3. Comparison between status boxes and turning points

From the process of the status box construction, it can be found that some adjacent turning

ACCEPTED MANUSCRIPT

points and the stock points between them are put into the same box so that each status box represents a long-term and unambiguous stock trend. Thus, the distinction between the different status boxes is more significant than the difference between the turning points and the non-turning points. Hence, the difficulty of correctly classifying the status boxes could be far less than that of classifying turning points. Moreover, since the number of status boxes is smaller than that of the turning points, the computational complexity would decrease significantly.

ACCEPTED MANUSCRIPT

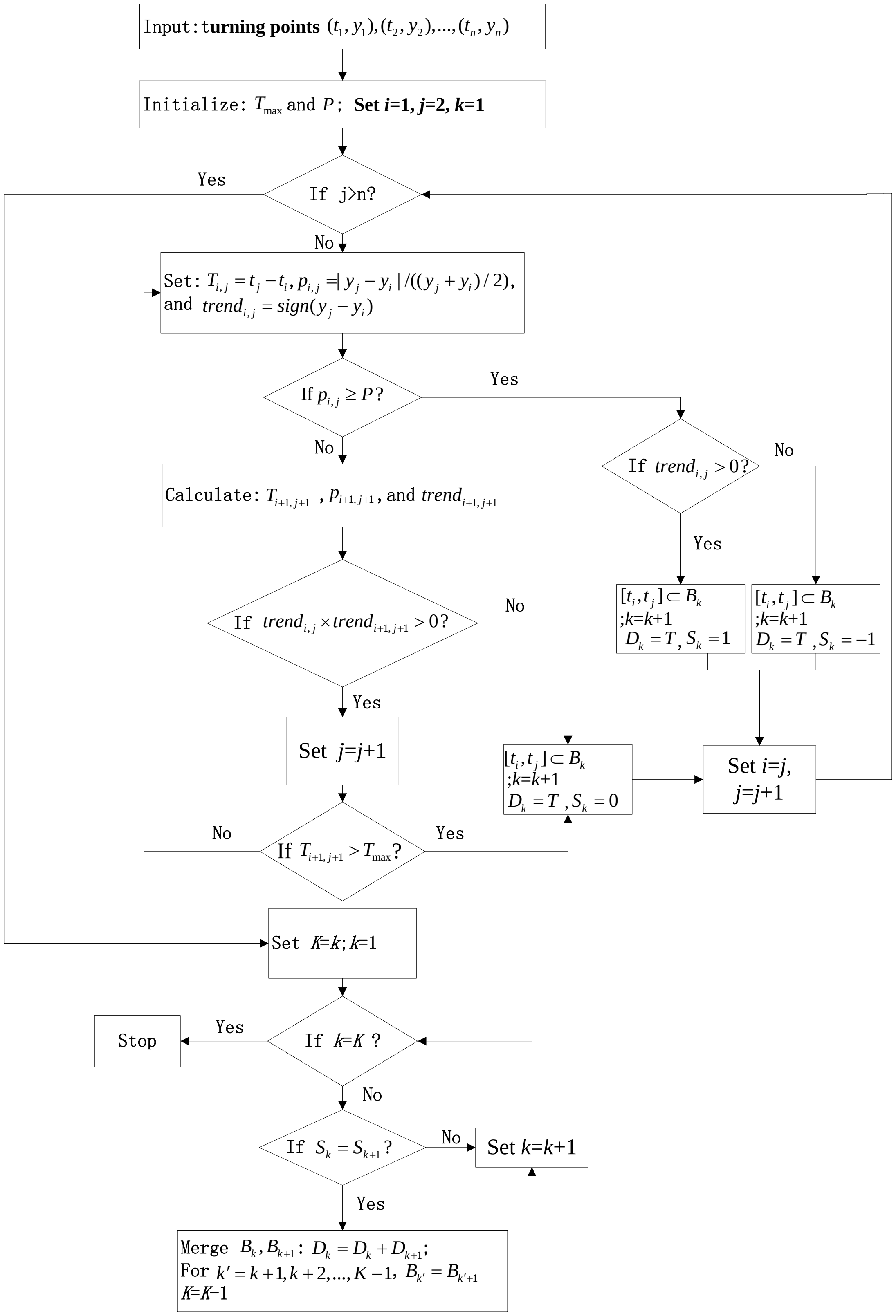

Input: turning points $(t_1, y_1), (t_2, y_2), ..., (t_n, y_n)$
Initialize: $T_{\max}$ and P; Set i=1, j=2, k=1
If j>n?
Yes
No
Set: $T_{i,j} = t_j - t_i, p_{i,j} = |y_j - y_i| / ((y_j + y_i)/2)$,
and $trend_{i,j} = sign(y_j - y_i)$
If $p_{i,j} \geq P$?
Yes
No
If $trend_{i,j} > 0$?
No
Yes
Calculate: $T_{i+1,j+1}$, $p_{i+1,j+1}$, and $trend_{i+1,j+1}$
$[t_i, t_j] \subset B_k$;k=k+1 $D_k = T, S_k = 1$
$[t_i, t_j] \subset B_k$;k=k+1 $D_k = T, S_k = -1$
If $trend_{i,j} \times trend_{i+1,j+1} > 0$?
No
Yes
Set j=j+1
$[t_i, t_j] \subset B_k$;k=k+1 $D_k = T, S_k = 0$
Set i=j, j=j+1
No
If $T_{i+1,j+1} > T_{\max}$?
Yes
Set K=k; k=1
Stop
Yes
If k=K ?
No
If $S_k = S_{k+1}$?
No
Set k=k+1
Yes
Merge B_k, B_{k+1}: $D_k = D_k + D_{k+1}$;
For $k' = k+1, k+2, ..., K-1$, $B_{k'} = B_{k'+1}$
K=K-1

Fig. 1. Process of status box construction

3. AdaBoost probabilistic weight support vector machine

3.1. Multi-class support vector machine

ACCEPTED MANUSCRIPT

The support vector machines (SVM) classifier, proposed by Vapnik in the early 1990s, is one of the most effective machine learning algorithms for classification problems [24,25]. It is based on the structural risk minimization principle and statistical learning theory. The basic idea of SVM is to transform the data into a higher dimensional space and find a classification hyper-plane that separates the data with the maximum margin [14]. The standard SVM model is as follows:

$$\begin{aligned} &\min \ \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l}\xi_i \\ &s.t. \ y_i((\omega\cdot\phi(x_i))+b) \geq 1-\xi_i \\ &\xi_i \geq 0, i=1,2,\cdots,l \end{aligned} \tag{1}$$

where $x_i \in R^n$ and $y_i \in \{-1,+1\}$ are the training samples and the corresponding class label respectively, ϕ is a nonlinear map that transforms the data to the high dimensional feature space, ω is the normal vector to the bounding plane, b is a bias value, $\xi_i (i=1,2,...,l)$ are the slack variables, and C is a penalty parameter. Instead of solving this optimization problem, it is easier to solve the dual problem:

$$\begin{aligned} &\min_{\alpha} \quad \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j k(x_i \cdot x_j) - \sum_{j=1}^{l}\alpha_j \\ &s.t. \quad \sum_{i=1}^{l} y_i\alpha_i = 0, \\ &\qquad 0 \leq \alpha_i \leq C, \ i=1,2,...,l \end{aligned} \tag{2}$$

where $k(x_i \cdot x_j) = \phi(x_i)\cdot\phi(x_j)$ is called kernel function. The binary classifier $g(x)$ and the decision function $f(x)$ could be calculated as follows:

$$f(x) = \mathrm{sign}(g(\mathrm{x})) = \mathrm{sign}(\sum_{x_i \in SV} y_i\alpha_i^* k(x, x_i) + b^*) \, , \tag{3}$$

where α_i^* is an optimal solution of the problem (2), *SV* is the set of support vectors, $b^* = y_j - \sum_{x_i \in SV} y_i \alpha_i^* k(x, x_i)$, if $0 < \alpha_i^* < C$.

The SVM can be extended to the multi-class problem. The common way is to decompose the multi-class issue into a series of two-class problems. The main two decomposition approaches are the one-versus-one (1-v-1) strategy and the one-versus-rest (1-v-r) strategy [26]. Although these methods sometimes obtain a preferable result, the classification accuracy is low when the samples are overlap [27].

In such a case, we map the binary classifier $g(x)$ to the probability for "soft" the classification margin so that the unclassified area can be reduced [28-29]. For 1-v-1 method with *k*-class problem, we must construct $k(k-1)/2$ binary classifiers g_{mn} $(m=1,...,k-1,$

ACCEPTED MANUSCRIPT

$n = m+1,...,k)$ that separates the classes *m* and *n*. For each classifier $g_{mn}(x)$, in order to estimate the probability of an unlabeled input *x* belonging to the class *m*, the probability output, $P(y=m\,|\,\boldsymbol{x})$, is calculated as follow:

$$P(y=m\,|\,x) = P_{A,B}(g_{mn}(x)) = 1/[1+\exp(A\times g_{mn}(x)+B)] \tag{4}$$

and

$$P(y=n\,|\,x) = 1 - P(y=m\,|\,x) \tag{5}$$

where the parameters *A* and *B* are obtained by solving the following optimization problem , based on Platt's criteria [30]:

$$\begin{aligned} &\min H(A,B) = -\sum_{i=1}^{l}(t_i\log(p_i)+(1-t_i)\log(1-p_i)) \\ &s.t.\ \ t_i = \begin{cases} (N_m+1)/(N_m+2) & if\ \ y_i = m \\ 1/(N_n+2) & if\ \ y_i = n \end{cases} \quad , \\ &p_i = P_{A,B}(g_{mn}(x_i)),\ i=1,2,...,l \end{aligned} \tag{6}$$

where N_m and N_n represent the number of class *m* and class *n* in training set respectively. Then, the decision function $f(x)$ and multi-classifier $g(x)$ for *k*-class problem are obtained as follows:

$$f(x) = \arg\max_{1\le m\le k} g(x) = \arg\max_{1\le m\le k} \frac{1}{k-1}\sum_{j=1,\,j\ne m}^{k} P_{A,B}(F_j)\,, \tag{7}$$

where $F_j = \begin{cases} g_{jm}(x), j<m \\ g_{mj}(x), j>m \end{cases}$. Specially, the probabilistic output is

$$P(y=m\,|\,x) = \frac{1}{k-1}\sum_{j=1,\,j\ne m}^{k} P_{A,B}(F_j), m=1,2,..,k \tag{8}$$

3.3 Hybrid algorithm integrated with AdaBoost and probabilistic multi-class SVM

To improve the performance of the probabilistic multi-class SVM, the AdaBoost algorithm is introduced. The AdaBoost (adaptive boosting) algorithm is one of the most popular ensemble methods proposed by Yoav Freund and Robert Shapire in 1995 [17]. It creates a collection of moderate classifiers by maintaining a set of weights over training data and adjusting these weights after each learning cycle adaptively. The weights of the training samples which are correctly classified by current classifier will decrease while the weights of the samples which are misclassified will increase [31-32]. Since our work need to use multi probabilistic SVM (PSVM) as component classifier in AdaBoost algorithm, the standard SVM (1) should be extended to the weight SVM (WSVM) for which each training sample has different weights [14]. In WSVM, the models (1) is transformed to

ACCEPTED MANUSCRIPT

$$\begin{aligned}&\min\ \frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{l}\mu_i\xi_i\\&s.t.\ \ y_i((\omega\cdot\phi(x_i))+b)\geq 1-\xi_i\ ,\\&\xi_i\geq 0, i=1,2,\cdots,l\end{aligned}\tag{9}$$

where $\mu_i(i=1,2,...,l)$ indicates the weight of the sample x_i. And the dual problem is as follow:

$$\begin{aligned}&\min_{\alpha} &&\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}y_iy_j\alpha_i\alpha_jk(x_i\cdot x_j)-\sum_{j=1}^{l}\alpha_j\\&s.t. &&\sum_{i=1}^{l}y_i\alpha_i=0,\\& &&0\leq\alpha_i\leq C\mu_i,\ i=1,2,...,l\end{aligned}\tag{10}$$

Likewise, the probabilistic method is used in the WSVM, then the probabilistic WSVM (PWSVM) is obtained. Except for the weights of samples, the classification performance of PWSVM is equally affected by its model parameters, for example, the penalty parameter C and the kernel parameters. During the AdaBoost iterations, if parameters make the classification accuracy of WSVM less than 50%, the requirement on a component classifier in AdaBoost cannot be satisfied. In contrast, if the accuracy is too high, boosting classifiers may become inefficient because the errors of these component classifiers are highly correlated [25]. Hence, how to select appropriate model parameters is important. There are many evolutionary algorithms for searching the suitable solution in real-valued spaces. With the advantages consisting of parallel search, solving complex problems, and large search space, the genetic algorithm (GA) is applied to perform the model parameters selection [33] in the k-fold cross-validation set. However, the process is time consuming and may cause the overfitting. Therefore, we adjust the model selection procedure so that the GA could be stopped when the cross-validation accuracy is over 0.5. In this result, the component classifiers conform to the condition of AdaBoost and the computational time could be saved. Moreover, the randomness of result produced by GA would decrease significantly after many times of Adaboost iterations. As a result, the outcomes corresponding to several independent runs of the hybrid method are similar to each other. So, the process of the hybrid algorithm is relatively stable. The detail process of AdaBoost-GA-PWSVM is given as follows:

AdaBoost-GA-PWSVM algorithm

1) **Input:** Training samples $\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),...,(\boldsymbol{x}_n,y_n)\}$, $\boldsymbol{x}_i\in R^u$ and $y_i\in\{1,...,K\}$; Moderate classifier multi-PWSVM; The number of classes *K*; The total number of the iterations T.

2) **Initialize:** The weights of training samples: $w_i^1=1/n, i=1,...,n$; The GA parameters include size of population N; Maximum number of generations $MaxI$; Length of chromosome of C and kernel parameters l; Crossover rate p_c and mutation rate p_m.

3) **For** $t=1,2,...,T$

ACCEPTED MANUSCRIPT

a) Select appropriate parameters

  i) Encode the parameter C and kernel parameters as an *l*-bit string which consists of l_1 bits standing for C and l_2 bits standing for kernel parameters, here $l= l_1+ l_2$; Generate an initial population consisted of N strings of binary bit. To avoid trapping into same local optimum in GA process, the parameters are estimated starting from a completely new initial population for each *t*.

  ii) For $j=1,2,...,N$, obtain the *j*th group of parameters by decoding the string j and train a component multi-classifier PWSVM g_t^j using these parameters on the *k*-fold cross-validation data set according to the equation (7).

  iii) Calculate the average cross-validation error: $E_t = \frac{1}{N}\frac{1}{n}\sum_{j=1}^{N}(\sum_{i=1}^{n} I(y_i, g_t^j(\boldsymbol{x}_i)))$, where the indicator function *I* produces 0 if the arguments are equal and 1 if they are different..

  iv) If $E_t < 0.5$, do step v) else the parameters satisfy the requirement of AdaBoost and do b)

  v) Perform reproduction: selection, crossover and mutation. Then, Generate new offspring population. If the number of generation exceeds $MaxI$, it means that the *t*-th moderate classifier is invalid, then stop the GA iteration, and do 3) and could not consider this classifier in b)-e). Otherwise, do ii)-iv). Here, the fitness function $Fit_j = \sum_{i=1}^{n} \mathrm{err}(y_i, g_t^j(\boldsymbol{x}_i))/n$.

b) Train a multi-classifier PWSVM G_t using the suitable parameters from a) and obtain a probabilistic output vector $\boldsymbol{P}_t(\boldsymbol{x}_i) = (P_t^1,...,P_t^K)$, where $P_t^m = P(y=m \mid \boldsymbol{x}_i), m=1,..,K$ is calculated by (8).

c) Compute the training error of G_t: $\varepsilon_t = \sum_{i=1}^{n} w_i^t \cdot I(y_i, G_t(\boldsymbol{x}_i))$.

d) Set weight for the current classifier G_t: $\alpha_t = 0.5\ln(\frac{1-\varepsilon_t}{\varepsilon_t})$

e) Update the weights: $w_i^{t+1} = w_i^t \exp\{-\alpha_t y_i G_t(\boldsymbol{x}_i)\}$

4) **Output**: If the number of invalid classifier reaches *T*, the hybrid algorithm is failed. Otherwise, $g(\boldsymbol{x}) = \underset{1\le m\le K}{\arg\max}((\sum_t \alpha_t P_t^1,...,\sum_t \alpha_t P_t^m,...,\sum_t \alpha_t P_t^K))$

4. Feature construction and selection

4.1. Input features construction

In technical analysis, there are many factors that have impact on the trading signal. In addition to the basic information of stock such as open price, the main two technical indicators are

ACCEPTED MANUSCRIPT

trend tracking index and trend reversal index [34-36]. They represent different states of stock. If we directly use these indicators for the single point classification, the delay or predictability of different indexes may affect the classification result. With this regard, the feature construction method for status box method is developed. Instead of using features for a single point, the features of a box are divided into the box body features and box tail features. The box body contains all points of the box and the box tail contains 30% samples at the right end of the box. The features of box body are mainly constituted by trend tracking indexes and the value of each feature is computed by the average of the samples belonging to the box body. Accordingly, the features of box tail are consisted of the trend reversal indexes and each feature value is defined by the average of these box tail samples. As a result, the biased information of single point is reduced because the box body features reflect the primary trend of stock price in box and the box tail indicates the reversal situation of the stock. In this paper, the trend tracking features of stock price consists of 12 variables and the trend reversal indexes have 7 attributes. Moreover, these features are scaled to [-1, 1] by $\boldsymbol{x}_{scoled} = 2\left(\boldsymbol{x}-x_{\min}\right)\Big/\left(x_{\max}-x_{\min}\right)-1$. Adding the status of previous box, the used features are listed in **Table 1**.

Table 1
Details of the constructed features

Trend tracking index
Moving average for 5 days (MA.5)
Moving average for 20 days (MA.20)
Cross signal of moving average (MA.cro)
Deviation of the moving average to the closing price (MA.dev)
Moving average convergence and divergence (MACD)
Oscillator signal line (MACD.cro)
Upper bound of the bollinger band (BB.up)
Lower bound of the bollinger band (BB.lo)
Cross of dual moving average (MACD.diff)
Relative strength index (RSI)
Psychological line (PSY)
On balance volume (OBV)
Trend reversal indexes
Williams Overbought/Oversold Index (WMS)
Signal of the WMS (WNS.sig)
Stochastic index K (K)
Stochastic index D (D)
Stochastic index J (J)
Bias (BIAS)
Momentum index (MTM)
Status box indexes
Status of previous box (SOPB)

4.2. Input features selection

To find the importance of the each input feature to the classification of status boxes and

ACCEPTED MANUSCRIPT

reduce the influences of useless features, the minimal-redundancy-maximal-relevance method (mRMR) is used to perform feature selection [16, 37]. The mRMR method can rank the features based on their relevance to the concerned target and the redundancy among the features themselves. The ranked feature with a smaller index represents that it has better trade-off between the maximum relevance and minimum redundancy.

5. Models implementation

The overall procedure of the status box method with AdaBoost-GA-PWSVM is mainly divided in two parts: the status boxes construction and the status boxes classification. The main process is shown in **Fig. 2.**

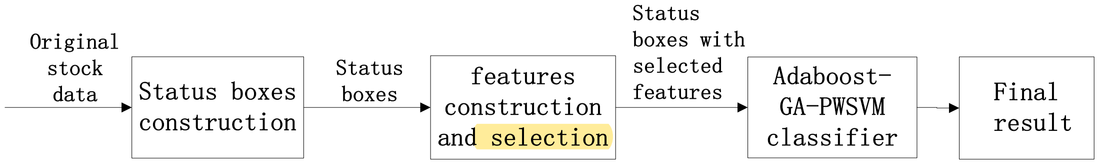

Original stock data
Status boxes construction
Status boxes
features construction and selection
Status boxes with selected features
Adaboost-GA-PWSVM classifier
Final result

Fig. 2. The main process of the stock trend prediction with AdaBoost-GA-PWSVM

5.1. Status boxes construction implementation

In the PLR, the threshold δ is set by 0.01 to guarantee sufficient turning points. After getting the turning points, we could set the thresholds of the maximum time interval $T_{\max}$ and price oscillation. The thresholds of $T_{\max}$ and P have intuitive meaning. Here, since a month is usually considered to be a long stock trader trade and a 5% gain or loss is regards as significant, $T_{\max}$ and P are simply set by 30 and 0.05. The investors could select different thresholds based on their expectations. The **Fig. 3** shows the result of the statue box construction for a share (002250.SZ) from the Shenzhen Stock Exchange. The time span is from 2008.09.10 to 2010.10.15.

ACCEPTED MANUSCRIPT

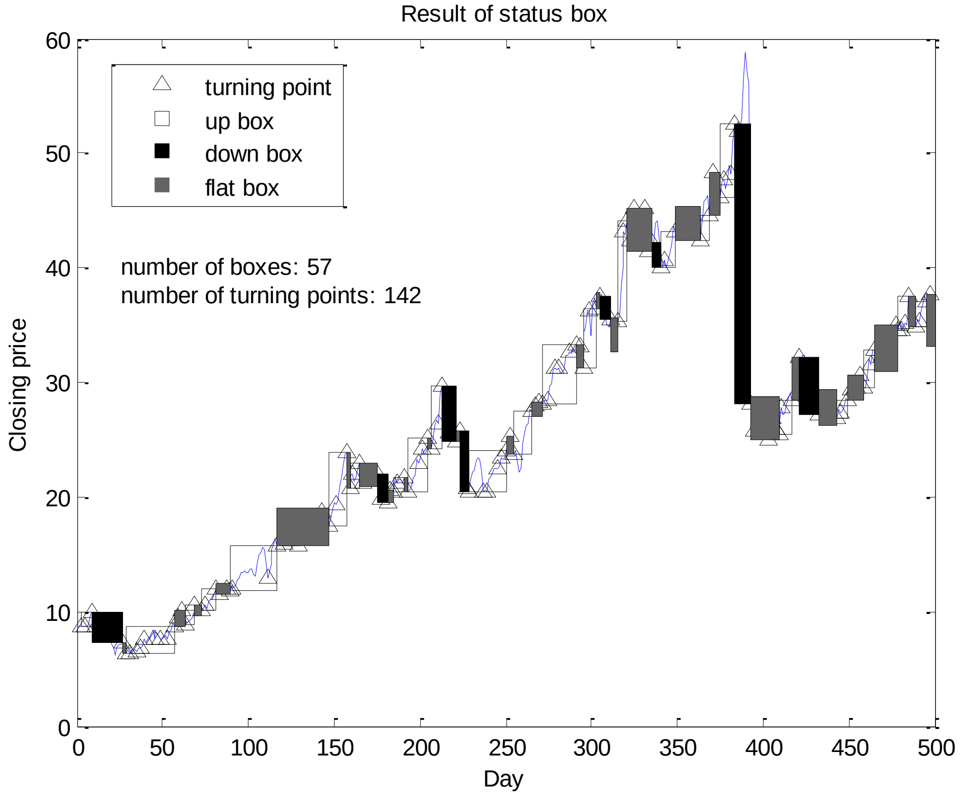


Fig. 3. The result of status boxes on a time series

From the **Fig. 3**, it can be found that the number of the status boxes is significantly less than the original stock points. It means that the computational complexity would be decreased.

5.2. mRMR implementation

The mRMR method is implemented by the MATLAB package written by Hanchuan Peng [16, 38]. The mutual information difference (MID) is used as the selection method and the number of initial selected features is set by the number of all features. According to the ranking result of the features computed by mRMR, we remove the feature from the last one of the order in turn and use the rest features to train model by AdaBoost-GA-PWSVM, PSVM or BPN. Based on the performance of cross-validation, the influence of the number of features can be found. Finally, the most important set of features are selected to do the independent testing. The feature importance rank of the status box and turning points are shown in appendix respectively.

5.3. AdaBoost probabilistic weight support vector machine implementation

The maximum number of AdaBoost iterations is set by 100. If the variation of the cross validation result is less than 5% in successive 10 iterations, we terminate the iteration process. This early stopping method could reduce the overfitting effectively. At the each iteration of AdaBoost iterations, the multi-class probabilistic PWSVM is used as component classifier which is trained by LibSvm developed by Lin [39]. Here, the RBF kernel is chosen as follows:

$$K(x_i, x_j) = \exp(-\frac{\left\| x_i - x_j \right\|}{2g^2}),$$

where g is an adjustable parameter. Thus, two parameters C and g need to be chosen in PWSVM algorithm. The GA in 5-fold cross-validation set is applied to select appropriate parameters of model for ensuring the effective classification result. The parameters of GA are

ACCEPTED MANUSCRIPT

listed in **Table 2**. These parameters are chosen based on the previous classification experimental results in order to ensure that the optimal solution would be found probably in the given search range.

Table 2

GA parameters selection

Parameters	Value or value range
Size of population (N)	20
Maximum number of generations ($MaxI$)	30
Penalty parameter C	[0-100]
Kernel parameter g	[0-100]
Crossover rate (P_c)	0.9
Mutation rate (P_m)	0.01

6. Numerical experiments

6.1. Experiment design

In order to better validate the results we apply the proposed methods to the entire market. Twenty shares come from different areas in the Second-board Market of Shenzhen Stock Exchange (SZSE) and sixteen shares chosen from National Association of Securities Dealers Automated Quotations (NASDAQ) are used to investigate the performance of the status boxes classification. The time span of these shares is from 2010.01.04 to 2015.05.22. These stock data are from the Wind database and other open sources. The fields of these shares include finance, medical industry, internet, retail industry, and transportation etc. To establish the trading strategies, the stock points would be classified into three categories: buy points, sell points and other points, where the buy points are local minimal point and the definition of sell points is on the contrary. To reduce the cost of tagging, the all buy points and sell points are from the turning points calculated by PLR. Accordingly, the other points are the non-turning points. Thus, the problem of turning points prediction is equivalent to a three classification problem. In the following, the turning points method and the status boxes method will be applied respectively to perform numerical experiments for above stock data Except for the variable of the previous box status, the used features of both the status boxes and the original stock points are same. Each dataset is divided into two portions: training set and testing set. The 80% random selected samples are the training data and the rest are testing data no matter what original stock point data or statues boxes are considered. The training set and testing set are mutually independent. The training samples are used to perform feature importance estimation and feature selection with 5-fold cross-validation. During 5-fold cross validation, the training samples are randomly partitioned into 5 equal sized subsets. Among the 5 subsets, a single subset is used to evaluate the influence of the different features, and the remaining 4 subsamples are used to train the model with the hybrid methods. In order to evaluate the superiority of the proposed status box method, the numerical experiments are divided into two aspects. Part one: compare classification performance of the different methods in features selection procedure using training samples. Part two: study performance of the different methods in independent evaluation using testing samples. In the features selection process, the mRMR method is used to select optimal features for both the status boxes and stock points about

ACCEPTED MANUSCRIPT

each share respectively. Both multi-class PSVM and BPN are applied to classify the stock points and the status boxes. Because of the requirement of the component classifier in ensemble method and low performance of single turning points certified by experiments, the AdaBoost-GA-PWSVM is only used to perform classification for the status boxes. For the BPN, we use the neural networks which contains three types of layers and set the number of the neurons of hidden layer as $\sqrt{NI \times (NO+1)}$ based on the rule of thumb, here NI is the dimension of input features and NO is the dimension of the output. Then, in the process of feature importance evaluation, the cross-validation results are set as the metric of the performance of status box method and turning points method. In the independent testing, the PSVM, BPN and AdaBoost-GA-PWSVM are applied to train the model based on the best features computed by the process of features selection for each share and predict the test samples.

6.2. Performance evaluation

The correct rate of the classification served as the accuracy is a usual evaluation criterion for classification problems which reflects the overall classification performance of classifier. It is calculated as follow: $Acc = T / n$, where T is the number of correct classification samples and n is the total number of the samples. However, the accuracy will be invalid when the samples are unbalanced. The geometric means (g-means) metric, which are proposed by Kubat and Matwin [40], is applied, here $g - means = \sqrt[K]{\prod Acc_i}$, Acc_i is the classification accuracy of the class i and the K is the number of classes. For unbalanced datasets, the g-means can better measure the performance of classifier than the accuracy. In this case, both the cross-validation and independent testing use the g-means as the metric. Specifically, the accuracy evaluation criterion is also applied to the independent testing to show the serious unbalance problem of turning points classification.

6.3. Experimental results

Table 3 exhibits the result of sell points, buy points and other points from turning points calculated by PLR. **Table 4** shows the result of the status boxes construction.

Table 3

The number of the buy points, sell points, and others.

Code of share	Number of buy points	Number of sell points	Number of other points
300002.SZ	155	147	907
300003.SZ	139	158	942
300004.SZ	158	145	942
300005.SZ	156	146	1005
300006.SZ	144	133	922
300007.SZ	153	148	934
300008.SZ	146	139	958
300009.SZ	146	138	915
300010.SZ	143	135	868
300011.SZ	146	149	906
300012.SZ	144	161	955
300013.SZ	132	149	932
300014.SZ	155	165	963
300015.SZ	147	159	1013
300016.SZ	160	138	1001

ACCEPTED MANUSCRIPT

300017.SZ	166	158	983
300018.SZ	158	161	938
300019.SZ	155	184	980
300020.SZ	156	160	922
300021.SZ	143	165	1002
AAL	171	184	976
ABAX	175	162	994
ACAS	146	180	1005
ACHN	140	164	1027
ACTA	162	162	1007
ADI	179	175	977
ADP	181	181	969
AGII	156	185	990
AKAM	180	175	976
AMAT	174	170	987
AMKR	162	174	995
CAR	159	146	1026
CBRL	176	190	965
CERN	150	180	1001
COKE	159	182	990
FAST	144	158	1029

Table 4

The number of the three status boxes

Code of share	Number of up box	Number of down box	Number of flat box
300002.SZ	43	43	59
300003.SZ	42	36	55
300004.SZ	41	40	54
300005.SZ	43	38	53
300006.SZ	42	36	55
300007.SZ	36	35	52
300008.SZ	36	35	54
300009.SZ	34	36	46
300010.SZ	35	34	47
300011.SZ	37	39	55
300012.SZ	37	33	47
300013.SZ	39	34	48
300014.SZ	42	32	57
300015.SZ	38	35	58
300016.SZ	40	31	50
300017.SZ	44	38	58
300018.SZ	37	37	54
300019.SZ	44	35	56
300020.SZ	41	33	55
300021.SZ	37	36	52
AAL	50	33	60
ABAX	40	31	56
ACAS	34	28	44
ACHN	44	37	50

ACCEPTED MANUSCRIPT

ACTA	33	32	45
ADI	31	15	35
ADP	22	9	27
AGII	29	18	40
AKAM	41	25	50
AMAT	31	26	43
AMKR	38	31	49
CAR	46	33	55
CBRL	31	13	41
CERN	34	14	42
COKE	21	16	30
FAST	27	22	37

From **Table 3**, both the numbers of buy points and sell points are in a minority of the samples, resulting in a serious imbalance problem. **Table 4** shows that the proportions of the three status boxes are similar. It implies that the imbalance problem is eliminated by the status box method. To show the process of features selection clearly, we use the average g-means of 20 shares to represent the performance with different number of features. It is emphasized that the importance order of the features of each share is different. In addition, the number of the optimal features calculated by different classification approaches is also various. **Fig. 4** shows the change of the average g-means of 20 shares from SZSE. The g-means of each share is calculated by the status box method with different classification approaches in the features selection process. For the same process, fig 5 indicates the average g-means of 20 shares via the single stock point method under various classification approaches. Similarly**,** fig.6 and fig.7 present the average g-means of 16 shares from NASDAQ. Here, the parameters C , g are set as 1 and 0.07.

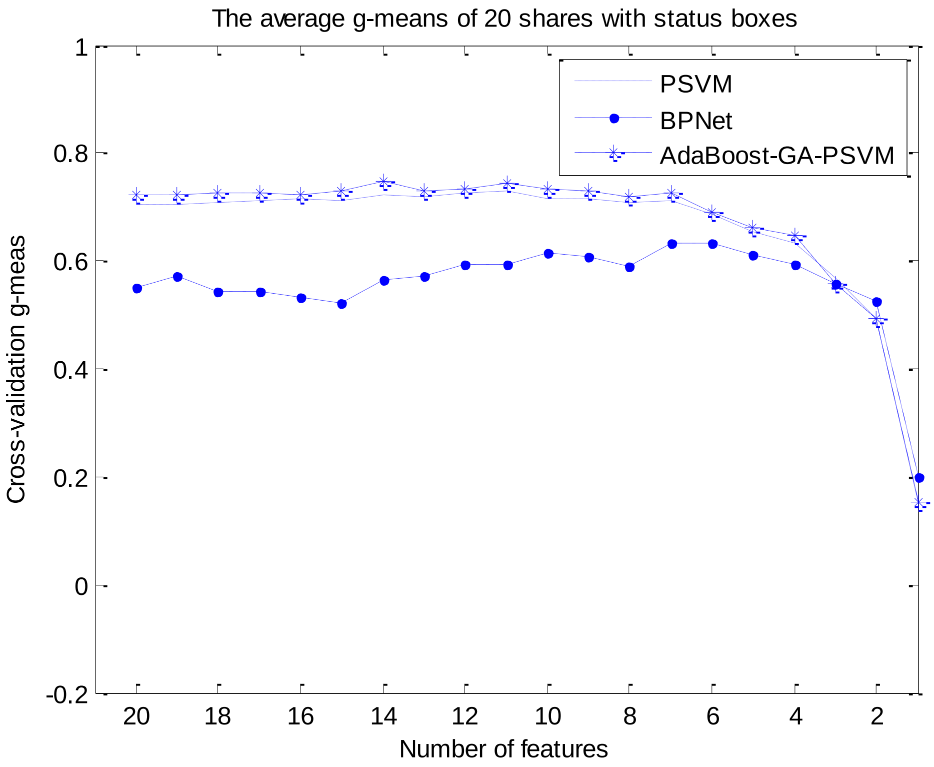


Fig. 4. The result of the features selection for status box method with PSVM and BPN in

ACCEPTED MANUSCRIPT

cross-validation set from SZSE .

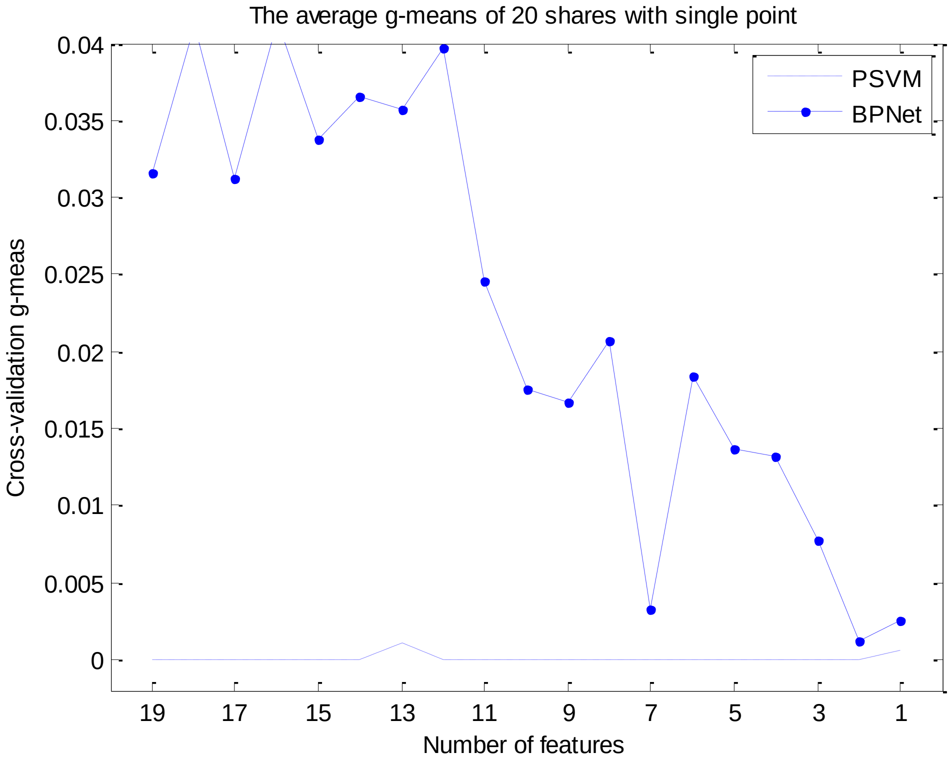


Fig. 5. The result of the features selection for stock points with PSVM and BPN in cross-validation set from SZSE.

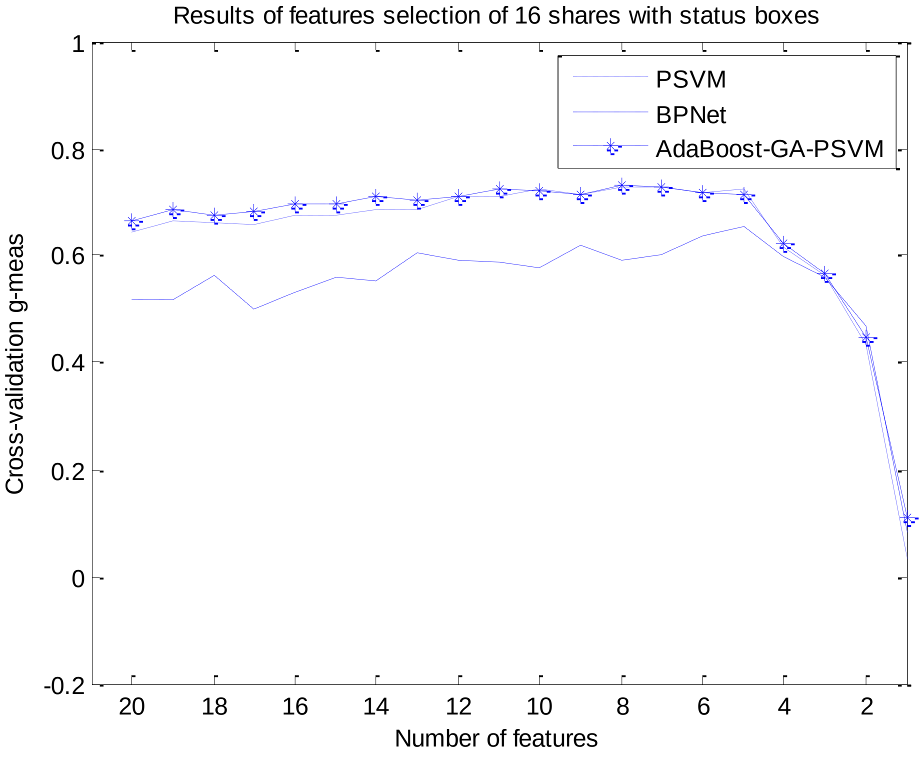


Fig. 6. The result of the features selection for status box method with PSVM and BPN in

ACCEPTED MANUSCRIPT

cross-validation set from NASDAQ.

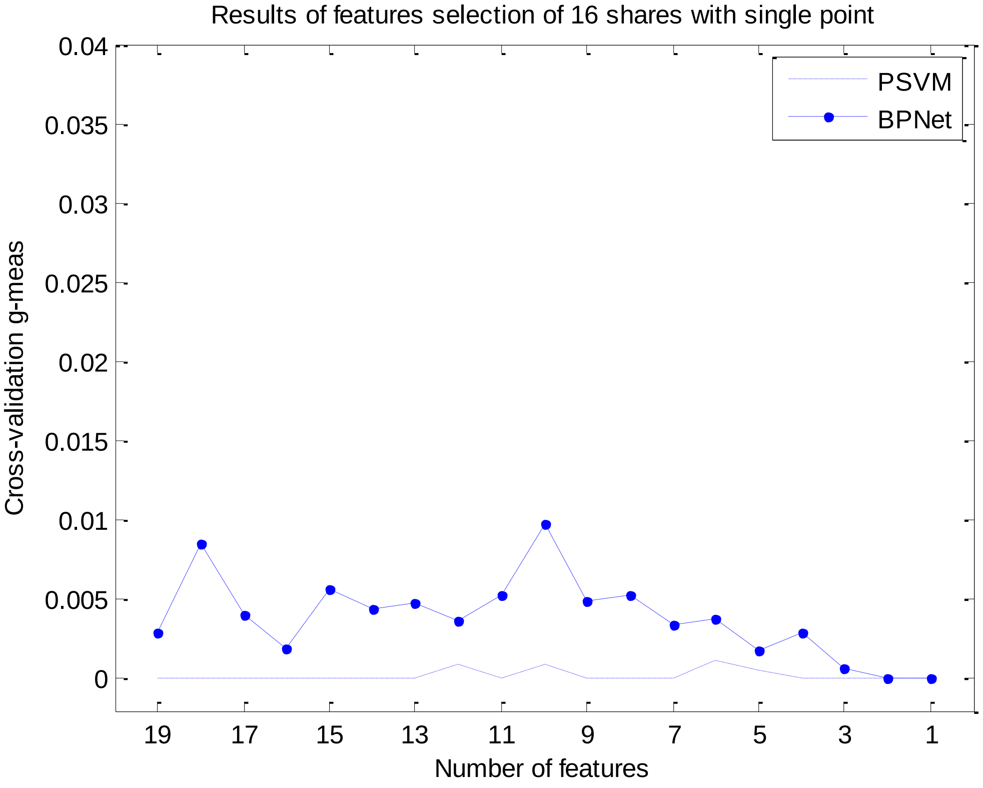


Fig. 7. The result of the features selection for stock points with PSVM and BPN in cross-validation set from NASDAQ.

From the **Fig. 4, 5, 6, and 7**, we can see that the performance of the status box method is much better than that of the standard turning points prediction approach whatever the PSVM or BPN is used. It shows that the cross-validation g-means of the status box method is maintained at high level until the number of attributes decrease to 6 or less. It means that we could use less features to obtain a good result, which would reduce the computational expense greatly. In contrast, the cross-validation g-means of each standard turning points prediction is zero or approximate zero for all shares no matter what the attributes are. It shows that the standard turning points prediction method cannot perform correct classification because of the serious imbalance problem and the unobvious classification features. We now choose the features which correspond to the highest cross-validation g-means as the input features to train the models and classify the testing samples. The number of features of different shares and the corresponding classification methods are shown in **Table 5**. It emphasizes that the best g-means of most shares are same (zero) in stock turning points method. Hence, the g-means is insensitive to the number of features for the stock turning points method, we choose the whole 20 features to train the model. The testing results, which are similar to the cross validation outcomes, are displayed in **Table 6**. From the **Table 6**, for all shares, we can see that both accuracy and g-means of the status box method for different classifiers are high and the difference of these two metric is small. Furthermore, although the testing accuracy of stock points classification for some shares are high as well, the g-means is zero. It means that this approach has serious unbalance problem which performs the false prediction for samples that belong to the minority classes. In particular, the status box method with AdaBoost-GA-PWSVM

ACCEPTED MANUSCRIPT

has the best g-means comparing with other approaches in the majority of shares. A few shares could give a better result via PSVM or BPN method, the main reason is that the model parameters selection causes a slight over-fitting in hybrid method. Although the g-means of the hybrid method is not the best for some shares, its stability is best in overall datasets. For the shares from SZSE market, the average of g-means of the AdaBoost-GA-PWSVM is 0.7552 which is better than that of PSVM (0.0214) and BPN (0.0519). Moreover, the variance of g-means of the AdaBoost-GA-PWSVM is 0.085402 which is much smaller than that of PSVM (0.142587) and BPN (0.22202) methods in Status boxes classification. And, for the shares from NASDAQ market, the average of g-means of the AdaBoost-GA-PWSVM, PSVM and BPN are 0.856724, 0.724306 and 0.723361 respectively. The variance of g-means of these three algorithms are 0.100495, 0.119771 and 0.178845 respectively. The proposed method achieves the best results in precision and stability. That means the performance of the AdaBoost–GA-PWSVM is very preferable and stable. Moreover, the g-means and the accuracy of the hybrid algorithm are more close to each other compared with the other methods. Finally, the average training time and testing time of all shares with different box status methods are listed in Table 7. (The CPU is Core i5-5200U and memory is 8G). It could be found that the AdaBoost-GA-PWSVM spends more training time than other methods, but same testing time. The reason is that GA performing the multiple classifiers' parameters selection spends more time on training process and all testing samples are simply classified based on voting. Hence we use more training time to obtain reliable model and r-elatively less time to test unclassified samples that satisfy the pattern of popular algorithm such as deep learning.

Table 5

The selected features number of status boxes and stock points about all shares with different classification methods.

Code	Status boxes			Stock points	
	AdaBoost–GA-PWSVM	PSVM	BPN	PSVM	BPN
300002.SZ	20	13	12	19	9
300003.SZ	9	17	8	19	7
300004.SZ	10	14	7	19	16
300005.SZ	13	16	5	19	11
300006.SZ	13	20	6	19	12
300007.SZ	4	11	8	19	19
300008.SZ	6	15	6	19	15
300009.SZ	11	18	12	19	8
300010.SZ	7	19	7	19	16
300011.SZ	14	11	9	19	6
300012.SZ	10	18	11	19	10
300013.SZ	9	20	8	19	18
300014.SZ	8	16	9	19	12
300015.SZ	20	12	7	19	11
300016.SZ	7	11	4	19	13
300017.SZ	7	9	8	19	16
300018.SZ	6	20	4	19	13
300019.SZ	15	17	11	19	13

ACCEPTED MANUSCRIPT

300020.SZ	5	15	4	19	7
300021.SZ	11	10	6	19	17
AAL	12	10	8	19	10
ABAX	6	8	7	19	12
ACAS	6	5	6	19	6
ACHN	14	15	5	19	19
ACTA	10	11	12	19	9
ADI	6	5	4	19	11
ADP	7	5	7	19	15
AGII	4	3	7	19	14
AKAM	7	14	11	19	13
AMAT	10	11	5	19	17
AMKR	11	10	10	19	18
CAR	10	12	19	19	13
CBRL	6	5	6	19	16
CERN	5	4	4	19	14
COKE	8	7	9	19	19
FAST	12	5	5	19	17

Table 6

The g-means and the accuracy of the three models for the shares of SZSE and NASDAQ in testing set.

Code	Criterion	Status boxes			Stock points	
		AdaBoost–GA-PWSVM	PSVM	BPN	PSVM	BPN
300002.SZ	accuracy	0.862069	0.724137	0.931034	0.764463	0.764462
	g-means	0.768881	0.648499	**0.830454**	0	0
300003.SZ	accuracy	0.777778	0.740740	0.481481	0.754032	0.754032
	g-means	**0.782974**	0.752828	0.493242	0	0
300004.SZ	accuracy	0.777778	0.740740	0.740740	0.726908	0.726907
	g-means	0.821227	0.789609	**0.884639**	0	0
300005.SZ	accuracy	0.777778	0.814814	0.407407	0.774809	0.774809
	g-means	0.711379	**0.801188**	0.400594	0	0
300006.SZ	accuracy	0.666667	0.703703	0.777777	0.754167	0.775000
	g-means	0.704730	0.704729	**0.741887**	0	0
300007.SZ	accuracy	0.760000	0.760000	0.520000	0.753036	0.372469
	g-means	**0.713766**	0.662601	0.586765	0	0
300008.SZ	accuracy	0.880000	0.880000	0.880000	0.755020	0.755020
	g-means	**0.900069**	0.804574	0.655185	0	0
300009.SZ	accuracy	0.833333	0.916666	0.625000	0.762500	0.762500
	g-means	0.822071	**0.919641**	0.854987	0	0
300010.SZ	accuracy	0.666667	0.666666	0.500000	0.734783	0.643478
	g-means	**0.640290**	0.629960	0.396850	0	0
300011.SZ	accuracy	0.555556	0.407407	0.222222	0.771784	0.771784
	g-means	**0.572357**	0.364959	0.436790	0	0
300012.SZ	accuracy	0.750000	0.750000	0.583333	0.793651	0.793650

ACCEPTED MANUSCRIPT

	g-means	**0.704730**	0.531329	0.443952	0	0
300013.SZ	accuracy	0.800000	0.800000	0.800000	0.786008	0.732510
	g-means	**0.843433**	0.814325	0.600000	0	0
300014.SZ	accuracy	0.777778	0.740741	0.629629	0.739300	0.735408
	g-means	**0.772071**	0.629629	0.678604	0	0
300015.SZ	accuracy	0.814815	0.703703	0.740740	0.795455	0.784090
	g-means	0.727479	0.621446	**0.870658**	0	0
300016.SZ	accuracy	0.960000	0.960000	0.760000	0.784615	0.665384
	g-means	**0.908560**	0.908560	0.754518	0	0
300017.SZ	accuracy	0.750000	0.785714	0.428571	0.782443	0.713740
	g-means	0.753947	**0.797877**	0.729919	0	0
300018.SZ	accuracy	0.653846	0.538461	0.692307	0.753968	0.583333
	g-means	**0.629961**	0.500000	0.381571	0	0
300019.SZ	accuracy	0.814815	0.814815	0.296296	0.731061	0.727272
	g-means	0.854988	0.885548	**0.908560**	0	0
300020.SZ	accuracy	0.807692	0.730769	0.730769	0.754032	0.754032
	g-means	**0.758489**	0.700432	0	0	0
300021.SZ	accuracy	0.760000	0.600000	0.640000	0.774809	0.774809
	g-means	**0.713766**	0.541853	0.623531	0	0
AAL	accuracy	0.896551	0.896552	0.793103	0.753968	0.756554
	g-means	0.913195	0.913196	**0.916260**	0	0
ABAX	accuracy	0.961538	0.884615	0.923077	0.750000	0.737828
	g-means	**0.961500**	0.867061	0.893351	0	0
ACAS	accuracy	0.863636	0.727273	0.818182	0.763052	0.756554
	g-means	**0.881822**	0.736806	0.843433	0	0
ACHN	accuracy	0.777778	0.777778	0.592593	0.783333	0.764045
	g-means	0.764855	**0.781488**	0.341346	0	0
ACTA	accuracy	0.727273	0.727273	0.727273	0.753968	0.805243
	g-means	0.662602	0.662602	**0.834826**	0	0
ADI	accuracy	0.941176	0.823529	0.941176	0.750000	0.726592
	g-means	**0.956466**	0.854988	0	0	0
ADP	accuracy	0.916667	0.750000	0.833333	0.763052	0.715356
	g-means	**0.941036**	0	0.941036	0	0
AGII	accuracy	0.833333	0.722222	0.833333	0.737373	0.741573
	g-means	0.658634	**0.709492**	0.629961	0	0
AKAM	accuracy	0.833333	0.791667	0.791667	0.783333	0.689139
	g-means	0.742346	0.713766	**0.849775**	0	0
AMAT	accuracy	0.950000	0.750000	0.900000	0.783333	0.711610
	g-means	**0.965489**	0.683990	0.861774	0	0
AMKR	accuracy	0.875000	0.791667	0.708333	0.783333	0.767790
	g-means	**0.865951**	0.786768	0.785601	0	0
CAR	accuracy	0.925926	0.851852	0.777778	0.783333	0.771536
	g-means	**0.920210**	0.854247	0.785601	0	0
CBRL	accuracy	1.000000	0.882353	0.882353	0.783333	0.704120

ACCEPTED MANUSCRIPT

	g-means	**1.000000**	0.919641	0.746901	0	0
CERN	accuracy	1.000000	0.944444	0.888889	0.783333	0.704120
	g-means	**1.000000**	0.961500	1.000000	0	0
COKE	accuracy	0.785714	0.428571	0.642857	0.783333	0.767790
	g-means	**0.829827**	0.362460	0.658634	0	0
FAST	accuracy	0.722222	0.777778	0.500000	0.783333	0.749064
	g-means	0.643660	**0.780897**	0.485286	0	0

Table 7
The average training time and testing time of all shares with different box status methods

	AdaBoost-GA-PWSVM	PSVM	BPN
Training time (s)	**88.0815**	**1.8500**	**30.7815**
Testing time (s)	**0.0132**	**0.0079**	**0.0141**

7. Trading strategy of the status box method

After getting the reliable classification result based on our proposed approach, we would make the decision about buying or selling strategies. Different from the approach of turning point prediction which performs classification by single point, we would use the stock points of recent days (e.g. five days from now) to compose a new box or put the latest data into the last trained box and use it as the new box. Then, we classify the new box to obtain its status which represents the recent stock price trend. Here, there are two criterions to help investors to make decision. First, if the stock price is rising at high level, meanwhile the new box is classified as a flat box, down box, or an up box with lower output probability, that means the future price maybe decrease. Then, the investors are suggested to sell the stock, and vice versa. Second, if the latest data is added to the last box, and as a consequence the status of the new box changes or the probabilistic output of the original status decreases significantly, it indicates that the stock trend would reverses. The investors should make the related measures i.e. buy or sell stocks, based on the change of stock trend. In practice, the parameters of the status box method can be adjusted according to the investment strategy of investor. For example, the long-term investors could choose the bigger price oscillation P to increase the width of the box. It is noticeable that this status box method only provides a preferable suggestion in stock investments, the investors could combine the suggestion with other strategies to make more accurate decision. To show the profitability superiority of the proposed method, a comparative simulation between the investment strategies based on the status box with AdaBoost–GA-PWSVM and the one based on the turning points with PSVM has been done. We choose the above 16 shares in section 6 as the datasets as well (The shares whose code are 300005.SZ, 300008.SZ, 300017.SZ and 300019.SZ were suspended from 2015.05.22, so these shares are ignored). Following the date of the training set and the testing set, we newly select stock points of 20 days as the samples. For the status box method, the 20 days points are divided into 4 boxes. The detail about the buy-sell strategy of status boxes is shown in **Fig.8**. For the turning stock points, we use the strategy which buy shares in buy points, sell all holding shares in sell points and hold shares in other points. Supposed both initial funds and stock holdings are 10000 and set the profitability after 20 days as the evaluation criterion. The results of 16 shares are

ACCEPTED MANUSCRIPT

presented in **Fig.9**. In the Fig.7, it can be seen that the proposed status box method has higher profitability in 14 shares compared to the stock turning points method. Moreover, the majority of profitability is greater than 0 which means the proposed approach could generate more profit.

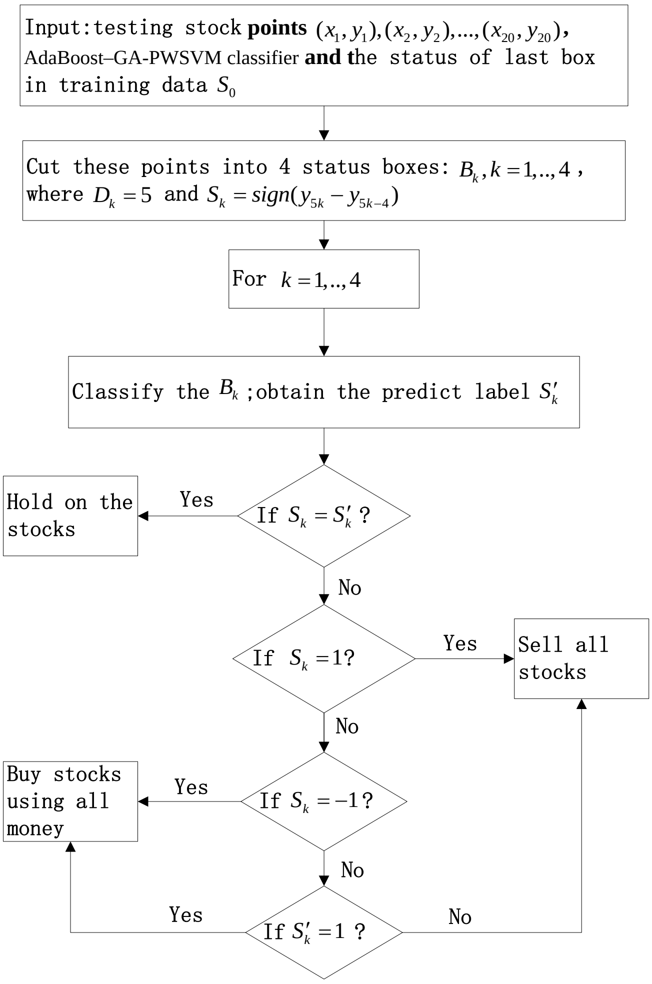

Input:testing stock points $(x_1,y_1),(x_2,y_2),...,(x_{20},y_{20})$,
AdaBoost–GA-PWSVM classifier and the status of last box in training data S_0
Cut these points into 4 status boxes: $B_k, k=1,..,4$, where $D_k=5$ and $S_k=sign(y_{5k}-y_{5k-4})$
For $k=1,..,4$
Classify the B_k;obtain the predict label S'_k
If $S_k=S'_k$?
Yes
Hold on the stocks
No
If $S_k=1$?
Yes
Sell all stocks
No
If $S_k=-1$?
Yes
Buy stocks using all money
No
If $S'_k=1$?
Yes
No

Fig. 8. The detail buy-sell strategy of status boxes.

ACCEPTED MANUSCRIPT

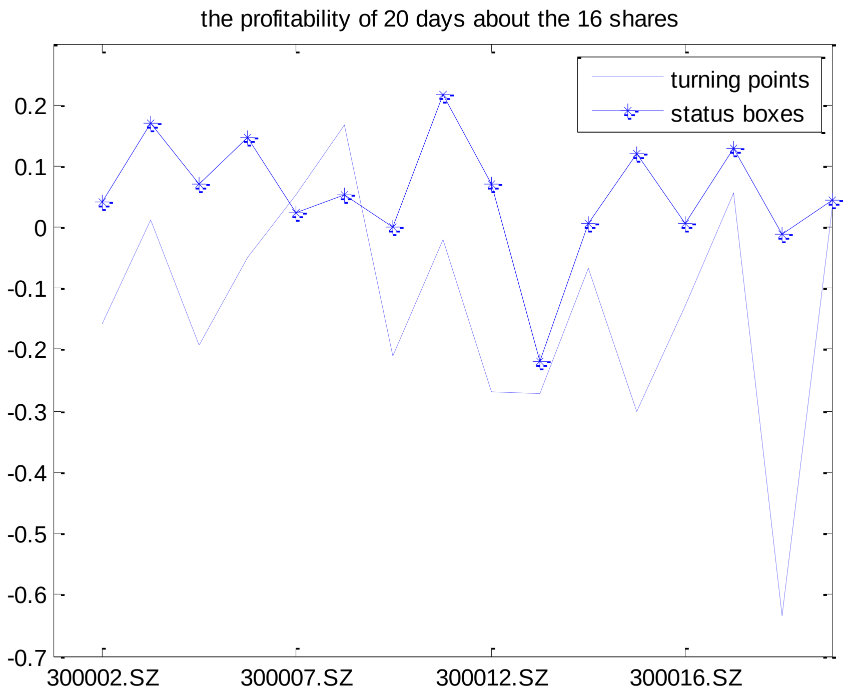


Fig. 9. The profitability of 20 days about the 16 shares with status boxes and turning points.

8. Conclusion

In this paper, a status box method is proposed to appraise the states of current stock price and predict the future stock trend so as to help investors make a reasonable investment decision. Compared with the traditional methods which predict stock price or turning point directly, the status box method makes improvements in many aspects. Firstly, the stock points are packaged into some status box based on the initial turning points computed by PLR. An effective and long-term turning trend could be obtained according to the junction of neighboring status boxes. Secondly, because of the uniform distribution of the status boxes of different categories, the unbalance problem about the classification issues of turning points can be avoided effectively. Thirdly, the proposal of the box body feature and the box tail feature makes the feature construction more reasonable. Fourthly, a new AdaBoost-GA-PWSVM algorithm has been applied to solve the classification problem and consequently the stronger generalization performance has been presented. The experimental results about 20 shares from Shenzhen Stock Exchange in China and 16 shares from NASDAQ show that the proposed status box method with AdaBoost-GA-PWSVM obtains the best g-means compared with the corresponding results from other models. Based on the simulation about buy or sell strategies through different approaches, the proposed hybrid algorithm with status boxes achieves a better profitability than that of applying turning points method in majority of testing shares. It supports that the new method could bring more profit for investors. Thus, according to the status and the probabilistic output obtained by the status box method, we can determine that what the stock future trend could be and adopt more reasonable investment strategies.

Acknowledgement

This work was supported by the Fundamental Research Funds for the Central Universities of China under Grant FRF-BR-12-021.

References

ACCEPTED MANUSCRIPT

[1] H. Takayasu, Practical fruits of econophysics, Springer, 2006.

[2] A. Kazem, E. Sharifi, F.K. Hussain, M. Saberi, O.K. Hussain, Support vector regression with chaos-based firefly algorithm for stock market price forecasting, Applied Soft Computing, 13 (2013) 947-958.

[3] Y. Zuo, E. Kita, Stock price forecast using Bayesian network, Expert Systems with Applications, 39 (2012) 6729-6737.

[4] K.-J. Kim, I. Han, Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index, Expert Systems with Application, 19 (2000) 125-132.

[5] M. Qi, G.P. Zhang, Trend time–series modeling and forecasting with neural networks, IEEE Transactions on Neural Networks, 19 (2008) 808-816.

[6] L. Yu, S. Wang, K.K. Lai, A neural-network-based nonlinear metamodeling approach to financial time series forecasting, Applied Soft Computing, 9 (2009) 563-574.

[7] M. Pulido, P. Melin, O. Castillo, Particle swarm optimization of ensemble neural networks with fuzzy aggregation for time series prediction of the Mexican Stock Exchange, Information Sciences, 280 (2014) 188-204.

[8] H. Yu, R. Chen, G. Zhang, A SVM stock selection model within PCA, Procedia Computer Science, 31 (2014) 406-412.

[9] K.-R. Müller, A.J. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, V. Vapnik, Predicting time series with support vector machines, Proceedings of 7th International Conference on Artificial Neural Networks, Springer, (1997) 999-1004. DOI: 10.1007/BFb0020283

[10] S. Choudhury, S. Ghosh, A. Bhattacharya, K.J. Fernandes, M.K. Tiwari, A real time clustering and SVM based price-volatility prediction for optimal trading strategy, Neuro-computing, 131 (2014) 419-426.

[11] E. Hajizadeh, A. Seifi, M.F. Zarandi, I. Turksen, A hybrid modeling approach for forecasting the volatility of S&P 500 index return, Expert Systems with Applications, 39 (2012) 431-436.

[12] L. Cao, F. E. Tay, Financial forecasting using support vector machines, Neural Computing & Applications, 10 (2001) 184-192.

[13] Q. Wen, Z. Yang, Y. Song, P. Jia, Automatic stock decision support system based on box theory and SVM algorithm, Expert Systems with Applications, 37 (2010) 1015-1022.

[14] L. Luo, X. Chen, Integrating piecewise linear representation and weighted support vector machine for stock trading signal prediction, Applied Soft Computing, 13 (2013) 806-816.

[15] Y.S. Abu-Mostafa, A.F. Atiya, Introduction to financial forecasting, Applied Intelligence, 6 (1996) 205-213.

[16] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Transactions on Pattern Analysis and Machine Intelligence, 27 (2005) 1226-1238.

[17] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, Journal of Computer and System Sciences, 55 (1997) 119-139.

[18] R. Agrawal, C. Faloutsos, A. Swami, Efficient similarity search in sequence databases, Springer, 1993.

[19] K.-P. Chan, A.W.-C. Fu, Efficient time series matching by wavelets, Proceedings of the 15th International Conference on Data Engineering, (1999) 126-133. DOI: 10.1109/ICDE.1999.754915

ACCEPTED MANUSCRIPT

[20] G. Das, K.-I. Lin, H. Mannila, G. Renganathan, P. Smyth, Rule Discovery from Time Series, Proceedings of the 3rd International Conference on Data Engineering, (1998) 16-22.

[21] P.-C. Chang, C.-Y. Fan, C.-H. Liu, Integrating a piecewise linear representation method and a neural network model for stock trading points prediction, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 39 (2009) 80-92.

[22] P.-C. Chang, T.W. Liao, J.-J. Lin, C.-Y. Fan, A dynamic threshold decision system for stock trading signal detection, Applied Soft Computing, 11 (2011) 3998-4010.

[23] E. Keogh, S. Chu, D. Hart, M. Pazzani, An online algorithm for segmenting time series, Proceedings IEEE International Conference on Data Mining, (2001) 289-296. DOI: 10.1109/ICDM.2001.989531

[24] V. Vapnik, The nature of statistical learning theory, Springer-Verlag New York, 1995.

[25] V. N. Vapnik, V. Vapnik, Statistical learning theory, Wiley New York, 1998.

[26] J. Weston, C. Watkins, Support vector machines for multi-class pattern recognition, Proceedings of the 7th European Symposium on Artificial Neural Networks, (1999) 219-224.

[27] H. Guo, W. Wang, An active learning-based SVM multi-class classification model, Pattern Recognition, 48 (2015) 1577-1597.

[28] F. Acevedo, S. Maldonado, E. Dominguez, A. Narvaez, F. Lopez, Probabilistic support vector machines for multi-class alcohol identification, Sensors and Actuators B: Chemical, 122 (2007) 227-235.

[29] T.-F. Wu, C.-J. Lin, R.C. Weng, Probability estimates for multi-class classification by pairwise coupling, The Journal of Machine Learning Research, 5 (2004) 975-1005.

[30] J. Platt, Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, Advances in large margin classifiers, 10 (1999) 61-74.

[31] R. Wang, AdaBoost for feature selection, classification and its relation with SVM, a review, Physics Procedia, 25 (2012) 800-807.

[32] X. Li, L. Wang, E. Sung, AdaBoost with SVM-based component classifiers, Engineering Applications of Artificial Intelligence, 21 (2008) 785-795.

[33] S.-H. Min, J. Lee, I. Han, Hybrid genetic algorithms and support vector machines for bankruptcy prediction, Expert systems with applications, 31 (2006) 652-660.

[34] P.J. Kaufman, Trading systems and methods, Wiley, 1998.

[35] C.-W. Chen, C.-S. Huang, H.-W. Lai, The impact of data snooping on the testing of technical analysis: An empirical study of Asian stock markets, Journal of Asian Economics, 20 (2009) 580-591.

[36] T.R.C.C. da Costa, R.T. Nazário, G.S.Z. Bergo, V.A. Sobreiro, H. Kimura, Trading System based on the use of technical analysis: A computational experiment, Journal of Behavioral and Experimental Finance, 6 (2015) 42-55.

[37] B.-Q. Li, L.-L. Hu, S. Niu, Y.-D. Cai, K.-C. Chou, Predict and analyze S-nitrosylation modification sites with the mRMR and IFS approaches, Journal of Proteomics, 75 (2012) 1654-1665.

[38] C. Ding, H. Peng, Minimum redundancy feature selection from microarray gene expression data, Journal of bioinformatics and computational biology, 3 (2005) 185-205.

[39] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, ACM Transactions on Intelligent Systems and Technology, 2 (2011). DOI: 10.1145/1961189.1961199.

[40] M. Kubat, S. Matwin, Addressing the curse of imbalanced training sets: one-sided selection,

ACCEPTED MANUSCRIPT

Proceedings of the 4th International Conference on Machine Learning, (1997) 179-186. DOI: 10.1.1.43.4487.

Appendix

Table 1

Feature importance rank of status boxes

Code	300002.SZ	300003.SZ	300004.SZ	300005.SZ	300006.SZ
MA.5	13	17	14	16	20
MA.20	17	14	6	20	12
MA.cro	2	2	5	2	2
MA.dev	9	18	16	14	7
MACD	7	20	13	15	14
MACD.cro	3	3	2	6	3
BB.up	18	16	17	17	19
BB.lo	19	10	18	7	18
MACD.dif	15	19	19	4	5
RSI	10	15	9	9	11
PSY	6	5	10	13	16
OBV	14	12	15	8	8
WMS	20	6	7	11	13
WNS.sig	4	8	3	3	4
K	16	13	8	12	17
D	11	4	12	18	9
J	5	7	4	10	10
BIAS	12	11	11	5	6
MTM	8	9	20	19	15
SOPB	1	1	1	1	1

Code	300007.SZ	300008.SZ	300009.SZ	300010.SZ	300011.SZ
MA.5	7	15	18	18	11
MA.20	12	19	12	12	19
MA.cro	2	2	2	2	2
MA.dev	6	17	11	11	4
MACD	19	13	6	6	18
MACD.cro	17	5	4	4	5
BB.up	20	12	19	19	13
BB.lo	9	18	15	15	12
MACD.dif	14	8	14	14	16
RSI	15	11	13	13	20
PSY	8	16	10	10	10
OBV	5	7	5	5	15
WMS	10	6	7	7	14
WNS.sig	3	3	3	3	3
K	4	14	9	9	17
D	13	4	16	16	9

ACCEPTED MANUSCRIPT

J	16	10	8	8	6
BIAS	11	9	20	20	7
MTM	18	20	17	17	8
SOPB	1	1	1	1	1

Code	300012.SZ	300013.SZ	300014.SZ	300015.SZ	300016.SZ
MA.5	18	14	16	12	11
MA.20	13	17	20	19	20
MA.cro	2	2	2	2	2
MA.dev	6	10	3	14	9
MACD	19	19	6	8	14
MACD.cro	9	3	4	3	13
BB.up	14	20	17	10	18
BB.lo	3	12	18	15	15
MACD.dif	17	15	15	20	16
RSI	5	13	10	17	8
PSY	15	8	9	6	12
OBV	12	11	13	11	17
WMS	10	9	7	16	5
WNS.sig	4	4	5	4	7
K	16	18	19	13	10
D	20	16	11	5	3
J	11	6	8	9	6
BIAS	7	5	14	7	4
MTM	8	7	12	18	19
SOPB	1	1	1	1	1

Code	300017.SZ	300018.SZ	300019.SZ	300020.SZ	300021.SZ
MA.5	9	20	16	15	10
MA.20	18	9	20	17	20
MA.cro	6	2	2	2	2
MA.dev	3	8	12	10	7
MACD	19	18	10	20	9
MACD.cro	2	12	5	7	4
BB.up	8	17	13	19	12
BB.lo	16	19	18	13	13
MACD.dif	17	14	19	18	3
RSI	20	15	15	9	16
PSY	13	7	7	11	15
OBV	11	5	8	14	8
WMS	15	10	14	4	11
WNS.sig	10	3	6	3	6
K	7	11	9	12	18
D	14	4	17	5	17

ACCEPTED MANUSCRIPT

J	5	6	3	6	5
BIAS	4	13	11	8	14
MTM	12	16	4	16	19
SOPB	1	1	1	1	1
WNS.sig	4	4	3	4	3
K	14	10	5	14	8
D	3	8	12	6	15
J	7	11	11	5	7
BIAS	18	5	16	7	4
MTM	5	6	13	16	5
SOPB	1	1	1	1	1

Code	AAL	ABAX	ACAS	ACHN	ACTA
Feature					
MA.5	7	10	16	16	16
MA.20	19	18	17	19	19
MA.cro	2	5	2	2	2
MA.dev	5	4	7	11	11
MACD	17	8	12	14	14
MACD.cro	3	16	4	3	3
BB.up	13	15	20	17	17
BB.lo	16	20	13	18	18
MACD.dif	20	19	19	20	20
RSI	18	11	9	13	13
PSY	14	9	14	10	10
OBV	15	13	10	7	7
WMS	8	3	5	12	12
WNS.sig	4	2	3	5	5
K	6	12	18	15	15
D	9	17	11	6	6
J	10	7	8	8	8
BIAS	12	6	6	4	4
MTM	11	14	15	9	9
SOPB	1	1	1	1	1

Code	ADI	ADP	AGII	AKAM	AMAT
Feature					
MA.5	14	13	12	19	12
MA.20	17	15	8	11	17
MA.cro	3	3	2	2	2
MA.dev	2	11	19	13	4
MACD	15	9	10	16	8
MACD.cro	20	16	13	4	20

ACCEPTED MANUSCRIPT

BB.up	6	4	16	20	15
BB.lo	18	19	15	10	19
MACD.dif	7	7	18	18	14
RSI	11	14	5	8	7
PSY	19	12	17	15	13
OBV	8	8	11	9	6
WMS	10	17	6	5	9
WNS.sig	9	6	9	6	3
K	5	20	7	17	16
D	16	10	14	14	18
J	13	2	20	7	11
BIAS	4	5	4	12	10
MTM	12	18	3	3	5
SOPB	1	1	1	1	1

Code	AMKR	CAR	CBRL	CERN	COKE
Feature					
MA.5	19	14	9	16	9
MA.20	12	20	14	12	15
MA.cro	2	2	2	2	2
MA.dev	6	15	8	5	3
MACD	14	5	16	14	4
MACD.cro	3	4	4	20	19
BB.up	8	19	10	19	13
BB.lo	20	18	12	11	14
MACD.dif	17	7	20	17	20
RSI	7	17	11	8	8
PSY	13	12	13	15	11
OBV	15	10	7	7	7
WMS	10	13	5	9	6
WNS.sig	4	3	3	10	10
K	18	8	15	13	17
D	5	11	6	6	12
J	9	6	17	18	16
BIAS	16	9	19	4	5
MTM	11	16	18	3	18
SOPB	1	1	1	1	1

Code	FAST
Feature	
MA.5	5
MA.20	12
MA.cro	2

ACCEPTED MANUSCRIPT

MA.dev	3
MACD	7
MACD.cro	20
BB.up	10
BB.lo	19
MACD.dif	17
RSI	14
PSY	9
OBV	16
WMS	6
WNS.sig	4
K	18
D	8
J	15
BIAS	11
MTM	13
SOPB	1

Table 2

Feature importance rank of stock turning points

Code	300002.SZ	300003.SZ	300004.SZ	300005.SZ	300006.SZ
MA.5	13	17	18	17	11
MA.20	5	18	16	14	17
MA.cro	7	8	1	1	10
MA.dev	19	19	19	19	1
MACD	10	14	12	9	6
MACD.cro	12	10	13	5	19
BB.up	14	13	14	8	16
BB.lo	18	4	17	16	15
MACD.dif	9	7	11	12	5
RSI	16	15	4	3	4
PSY	6	12	10	10	14
OBV	3	5	7	7	18
WMS	1	3	3	4	7
WNS.sig	4	2	9	13	12
K	11	6	8	15	9
D	17	11	2	2	3
J	2	1	6	6	2
BIAS	15	16	5	18	8
MTM	8	9	15	11	13

Code	300007.SZ	300008.SZ	300009.SZ	300010.SZ	300011.SZ

ACCEPTED MANUSCRIPT

MA.5	18	17	17	Fea	5
MA.20	17	14	15	1	19
MA.cro	8	1	8	2	4
MA.dev	19	19	19	3	16
MACD	9	3	14	4	8
MACD.cro	1	13	9	5	14
BB.up	12	16	4	6	17
BB.lo	16	18	18	7	18
MACD.dif	3	9	13	8	6
RSI	7	8	2	9	13
PSY	13	12	12	10	10
OBV	14	11	5	11	7
WMS	4	4	7	12	1
WNS.sig	5	5	6	13	3
K	6	6	1	14	12
D	2	2	10	15	9
J	11	7	3	16	2
BIAS	15	15	16	17	11
MTM	10	10	11	18	15

Code	300012.SZ	300013.SZ	300014.SZ	300015.SZ	300016.SZ
MA.5	17	18	15	16	13
MA.20	4	14	16	17	9
MA.cro	1	6	2	6	2
MA.dev	19	19	19	18	19
MACD	15	10	11	5	16
MACD.cro	9	5	7	11	18
BB.up	18	9	17	13	15
BB.lo	11	17	3	19	6
MACD.dif	2	8	10	8	5
RSI	8	3	1	1	4
PSY	16	12	12	10	14
OBV	12	11	8	9	8
WMS	3	4	5	2	3
WNS.sig	14	1	4	15	10
K	7	15	6	7	7
D	13	16	13	3	1
J	5	2	9	4	12
BIAS	6	13	14	12	11
MTM	10	7	18	14	17

Code	300017.SZ	300018.SZ	300019.SZ	300020.SZ	300021.SZ
MA.5	15	13	8	18	18
MA.20	11	17	7	17	17

ACCEPTED MANUSCRIPT

MA.cro	1	3	5	1	1
MA.dev	19	19	19	19	19
MACD	5	14	10	7	8
MACD.cro	18	15	15	6	3
BB.up	16	10	2	15	14
BB.lo	13	18	18	16	9
MACD.dif	12	6	9	12	11
RSI	8	4	6	2	7
PSY	10	12	11	14	15
OBV	4	9	16	11	13
WMS	7	1	1	5	4
WNS.sig	17	7	3	13	12
K	3	8	14	8	6
D	9	11	17	3	5
J	2	2	4	10	10
BIAS	6	16	13	9	16
MTM	14	5	12	4	2

Code	AAL	ABAX	ACAS	ACHN	ACTA
Feature					
MA.5	6	15	18	9	11
MA.20	12	17	17	16	16
MA.cro	8	1	5	14	7
MA.dev	1	19	19	19	19
MACD	10	2	3	11	2
MACD.cro	18	11	11	18	13
BB.up	16	18	14	10	3
BB.lo	15	9	16	17	17
MACD.dif	13	8	8	3	14
RSI	9	7	7	2	18
PSY	4	16	15	15	5
OBV	19	4	2	13	10
WMS	7	13	9	7	1
WNS.sig	14	12	1	1	4
K	3	3	6	4	12
D	11	14	10	6	15
J	5	5	13	5	8
BIAS	2	6	12	8	6
MTM	17	10	4	12	9

Code	ADI	ADP	AGII	AKAM	AMAT
Feature					

ACCEPTED MANUSCRIPT

MA.5	14	11	14	12	14
MA.20	9	17	11	8	17
MA.cro	3	1	10	1	5
MA.dev	4	3	1	19	19
MACD	16	14	17	4	1
MACD.cro	17	10	3	9	15
BB.up	12	19	18	17	18
BB.lo	11	16	12	14	10
MACD.dif	10	18	15	7	9
RSI	5	12	13	18	13
PSY	13	4	9	15	16
OBV	1	9	16	6	3
WMS	2	13	7	2	4
WNS.sig	19	2	19	13	12
K	18	15	6	5	8
D	7	6	2	11	7
J	6	5	8	3	2
BIAS	15	7	4	10	6
MTM	8	8	5	16	11

Code	AMKR	CAR	CBRL	CERN	COKE
Feature					
MA.5	11	12	8	3	17
MA.20	12	13	9	16	15
MA.cro	3	18	1	2	1
MA.dev	19	19	3	8	18
MACD	18	8	16	11	2
MACD.cro	5	6	18	19	3
BB.up	15	16	12	4	11
BB.lo	14	15	11	17	16
MACD.dif	13	14	10	15	13
RSI	2	9	2	12	7
PSY	17	11	17	18	19
OBV	6	10	5	13	4
WMS	1	3	14	10	9
WNS.sig	9	2	19	5	8
K	7	5	4	6	14
D	4	4	6	1	10
J	8	1	7	14	6
BIAS	10	7	15	9	12
MTM	16	17	13	7	5

Code	FAST

ACCEPTED MANUSCRIPT

Feature	
MA.5	19
MA.20	2
MA.cro	1
MA.dev	5
MACD	9
MACD.cro	15
BB.up	18
BB.lo	17
MACD.dif	3
RSI	14
PSY	6
OBV	7
WMS	10
WNS.sig	16
K	12
D	4
J	11
BIAS	13
MTM	8