Pattern Recognition 44 (2011) 1245–1261

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Greedy optimization classifiers ensemble based on diversity

Shasha Mao *, L.C. Jiao, Lin Xiong, Shuiping Gou

Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China, Institute of Intelligent Information Processing, Xidian University, Xi'an, PR China

ARTICLE INFO

Article history:
Received 14 December 2009
Received in revised form
10 November 2010
Accepted 12 November 2010
Available online 4 December 2010

Keywords:
Diversity
Matching pursuit
Greedy optimization
Residual
Selective ensemble
Kappa-error diagram

ABSTRACT

Decreasing the individual error and increasing the diversity among classifiers are two crucial factors for improving ensemble performances. Nevertheless, the "kappa-error" diagram shows that enhancing the diversity is at the expense of reducing individual accuracy. Hence, a new method named Matching Pursuit Optimization Ensemble Classifiers (MPOEC) is proposed in this paper in order to balance the diversity and the individual accuracy. MPOEC method adopts a greedy iterative algorithm of matching pursuit to search for an optimal combination of entire classifiers, and eliminates some similar or poor classifiers by giving zero coefficients. In MPOEC approach, the coefficient of every classifier is gained by minimizing the residual between the target function and the linear combination of the basis functions, especially, when the basis functions are similar, their coefficients will be close to zeros in one iteration of the optimization process, which indicates that obtained coefficients of classifiers are based on the diversity among ensemble individuals. Because some classifiers are given zero coefficients, MPOEC approach may be also considered as a selective classifiers ensemble method. Experimental results show that MPOEC improves the performance compared with other methods. Furthermore, the kappa-error diagrams indicate that the diversity is increased by the proposed method compared with standard ensemble strategies and evolutionary ensemble.

1. Introduction

Ensemble learning is one of promising methods for constructing an accurate predictor, its techniques have developed in the field of ensemble classifiers by combining predictions of large numbers of basis classifiers. On the other hand, classifiers ensemble has been also an active area of research in machine learning and pattern recognition, and many studies have been proposed in Refs. [1–5]. Classifiers ensemble is defined that classifiers are combined by an ensemble strategy, and the motivation is to deal with some problems, which the single classifier is difficult to achieve better performance. According to Kuncheva [6], there are four fundamental approaches of ensemble: (a) using different combination strategies; (b) using different classifier models; (c) using different feature subsets; and (d) using different training set. Two classical methods, bagging [1] and boosting [7], are also used most of four approaches above. For example, Yasummura et al. introduced an ensemble method about integration of boosting and bagging in Ref. [8], Canul-Reich et al. used bagging to create an ensemble of fuzzy classifiers in Ref. [9], and so on.

In fact, ensemble classifiers has been initially proposed as neural network ensemble by Hansen and Salamon [10], in order to change a weak neural network into a strong network, which is interesting for pattern recognition due to improve the performance of classification compared with a single classifier. Subsequently, there are many studies about weak classifier ensembles and the improved algorithm for them. For example, decision tree ensemble in Ref. [11], Causal Discovery Based Neural Network Ensemble Method in Ref. [12], bagging-based selective clusterer ensemble in Ref. [13] and so on. On the basis of the analysis of Tumer and Ghosh [14], in an ensemble system, the generalization error is decided by the error of individual classifiers and the diversity among individual classifiers. Many papers based on increasing the diversity of classifiers and decreasing the individual error are proposed, such as [15–17]. Although the weak classifier was initially adopted to ensemble, some strong classifiers are also considered as basis classifiers to ensemble, such as support vector machine [18–20] and kernel matching pursuit [21]. Classifiers ensemble can deal with some problems that are intractable for a single classifier, simultaneously, compared with the single classifier, it also improves the performance in Refs. [22–26].

However, the study of Margineantu and Dietterich in Ref. [27] indicated that the error of individual and the diversity among individuals are affected by each other, in other word, a conflict occurred between the two factors. In that paper, the kappa-error diagram showed that they augmented the diversity at the expense of the accuracy of individual classifier. Exhilaratingly, Rodriguez and Kuncheva [28] proposed a successful ensemble classifiers method named Rotation Forest, in which the training set of each individual classifier was made by applying PCA to transform separately for different feature subsets of the original samples. Compared with the previous ensemble strategies, such as bagging,

* Corresponding author.
E-mail address: skymss0828@gmail.com (S. Mao).

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.11.007

AdaBoost [29,30] and random forest [31], Rotation Forest is more robust because it can encourage simultaneously individual accuracy and diversity within the ensemble. Subsequently, many researches concerning Rotation Forest technique have been proposed, such as Liu and Huang [32] classified for the cancer datasets by applying the idea of Rotation Forest and using ICA which can better describe the property of the microarray data instead of PCA; Zhang and Zhang [33] combined Rotation Forest and AdaBoost to construct a novel ensemble RotBoost, and the experiments were demonstrated that RotBoost could gain individual with lower prediction error than Rotation Forest and AdaBoost, it also performed better than bagging and MultiBoost; Nanni and Lumini [34] introduced a method with combining ensemble of classifiers (Rotation Forest and IDE [35]) and feature selection to identify students with learning disabilities. Additionally, in Ref. [36], Nanni and Lumini proposed input decimated ensemble based on neighborhood preserving embedding for spectrogram classification, especially, the combination of SVM and the proposed method obtained the best performance. In short, Rotation Forest has been probably the most performing ensemble method, and recently ensemble strategies are constructed based on the factors of the diversity and the individual accuracy. Noticeably, the most researches obtained the ensemble individual by constructing different training sets for individual classifiers to increasing diversity in terms of training classifiers of ensemble. Nevertheless, in the opinion of combination individual classifiers, whether or not two factors of ensemble could be balanced by selecting some different and better performance individual from entire classifiers in an ensemble system.

In Ref. [3], Zhou et al. proposed a method of selective ensemble neural networks, and the paper theoretically proved that selecting some available neural networks to ensemble could gain better performance than all networks. Zhang et al. [37] introduced selective SVM ensemble driven by immune clonal algorithm, and the experiments shown that selective ensemble is better. Partalas et al. [38] constructed the selective ensemble method based on diversity among classifiers, focused ensemble selection (FES), which abandoned the low *fes*-value classifiers by computing the diversity according to the prediction label of classifiers. According as the above researches, it indicates that if one classifier has strong diversity with others but its performance is lower than others, this classifier could be considered whether it is helpful for ensemble or not. Apparently, selecting classifiers to ensemble is essential for improving the performance of the ensemble system, which is not connected with the training sets. Consequently, it will be a hinge of selective ensemble that how to select some available classifiers to balance the diversity and the individual accuracy.

In this paper, we introduce a new method named Matching Pursuit Optimization Ensemble Classifiers (MPOEC), based on the diversity and accuracy of classifiers. This method searches greedily for an optimal combination of the whole ensemble classifiers based on matching pursuit theory, and obliterates some useless classifiers (similar or poor performance) for ensemble by giving a coefficient to each classifier. In MPOEC approach, every classifier of the ensemble system is considered as a basis function from the basis dictionary and the labels of samples as the target function approximated. The coefficient of each classifier is gained by iteratively minimizing the residual between the target function and the linear combination of classifiers, and then classifiers with nonzero coefficients will be selected to ensemble. Hence, a sparse optimal combination of classifiers is produced by MPOEC method, and the method can achieve several advantages compared with other selective ensemble approaches [38–40]. Firstly, selecting ensemble classifiers is an overall optimization process, and the coefficients of classifiers can be updated automatically by backfitting after several iterative processes, which ensure that the combination is the best approximation to the target function. Secondly, we can gain a better sparse combination of classifiers because of the sparsity of matching pursuit thought, and the speed of combining classifiers may be decreased compared with other selective ensemble methods. Finally, it is the most important point that the optimization process of the MPOEC approach is actually based on the diversity between a pair of classifiers according to the analysis for diversity among classifiers, namely, the proposed method selects classifiers to ensemble based on the diversity, which will be explicated detailedly in following section. The detailed algorithm of MPOEC is introduced in the following segment, and the experimental results indicate MPOEC can improve the performance of classification of the ensemble system.

The reminder of this paper is organized as follows. Section 2 introduces the general frame of classifiers ensemble system. Section 3 introduces the basic matching pursuit algorithm, gives the theoretic analysis of MPOEC, and shows how to search for the optimal combination of ensemble classifiers and the detailed introduction of the proposed algorithm. In Section 4, the analysis of the diversity of classifiers selected by MPOEC is shown, and it indicated that selecting classifiers is based on the diversity between a pair of classifiers. Our experiments demonstrate that the proposed method can gain better accuracy than before in Section 5. Finally, Section 6 offers the conclusion as well as the future works of this paper.

2. Classifiers ensemble system

As generally speaking, classifiers ensemble system is such a classification system based on combining same or different classifier models that are trained on different data subsets or feature subsets, in order to improve the classification accuracy of learning systems compared with the single classifier. Therefore, many methods of constructing ensemble classifier systems have been proposed, such as bagging [1] and boosting [7] which are methods based different training samples, random subspace [58] based different feature subsets, Rotation Forest [28] with transforming the feature subsets by PCA and so on. In addition, in recent years, it is proposed that using artificial immune algorithm to ensemble classifiers in Refs. [37,5]. Although there are many different proposed methods of ensemble classifiers, all of them are constructed to combine more than two different classifiers based on the diversity and individual error of classifiers. In general, a classifiers ensemble system is composed of two parts: training a number of component classifiers and then combining the component predictions, and the frame of ensemble classifiers is shown in Fig. 1. In following, an ensemble classifiers system is simply introduced.

Given an original training set $S_{train}=\{(x_i,y_i)|i=1, 2, \ldots, N_{train}\}$ and a testing set $S_{test}=\{x_j|j=1, 2, \ldots, N_{test}\}$, there into, $y_i \in \{1, 2, \ldots, K\}$ is the class label of the training sample x_i. At first, L classifiers $\{C_1, C_2, \ldots, C_L\}$ are produced based the different subsets of the original samples by adopting an ensemble strategy, such as bagging, random subspace, random forest and so on. Secondly, each classifier $C_l (l=1, 2, \ldots, L)$ can gain a prediction output label f_{jl} for each testing sample by learning. So $\{f_{j1}, f_{j2}, \ldots, f_{jL}\}$ is a combination of the prediction label of a testing sample x_j by L classifiers. Finally, the combination is integrated by a combinatorial method, such as majority voting [11] for classification, simple averaging [41], weighted averaging [42] for regression, and the best prediction label f_{jbest} is made according to the form

$$f_{jbest} = sign\left(\sum_{l=1}^{L} w_{jl} f_{jl}\right) \quad (1)$$

Many papers have indicated that ensemble classifiers can deal with some problems better than a single classifier, such as Giacinto

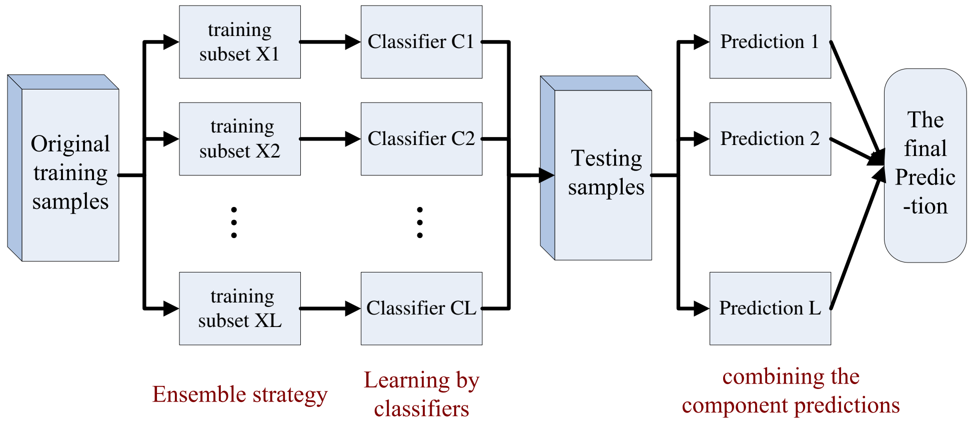

Fig. 1. The basic frame of ensemble classifiers system.

and Roli [43] proposed neural network ensembles for image classification, Pang et al. [44] used SVM ensemble to deal with face classification, Li et al. [45] introduced fault diagnosis based on SVM ensemble and so on. Although the advantages of classifiers ensemble have been shown, the shortages also emerged gradually along with the development of ensemble classifiers. For instance, the instability of the performance is useless for practical problems. In addition, not every classifier is beneficial for the classification performance, and even someone could harm the performance of an ensemble system. Hence, how to improve these shortages has been important for improving the performance.

3. Classifiers ensemble based on matching pursuit

3.1. Basic matching pursuit

In this section, we first introduce the basic matching pursuit algorithm [26,46], which was shown by Mallat and Zhang in Ref. [46] from a machine learning perspective.

It is given that $\{x_1, x_2, \ldots, x_I\}$ and $\{y_1, y_2, \ldots, y_I\}$ separately expresses S observational samples and these observational values, and a finite dictionary $D=\{d_1, d_2, \ldots, d_M\}$ including M functions is given in a Hilbet space H. Hypothetically, with a target function $q \in H$, the observational value $y_i (i=1, \ldots, I)$ can correspond to a sample $x_i (i=1, \ldots, I)$. Hence the motivation of matching pursuit algorithm is that the sparse approximations of q can be found out in the dictionary D, which is expansion of the form

$$q_N = \sum_{n=1}^{N} \alpha_n g_n \tag{2}$$

where, N is the number of the basis functions in the expansion, $\{g_1, g_2, \ldots, g_N\} \subset D$ shall be called the basis of the expansion, $\{\alpha_1, \alpha_2, \ldots, \alpha_N\} \in R^N$ is the set of corresponding coefficients of the expansion, q_N designs an approximation of q that uses exactly N distinct basis function taken from the dictionary. Notice that there is a correspondence between the dictionary function$\{d_1, \ldots, d_M\}$and the particular ones$\{g_1, \ldots, g_N\}$, which can be represented by the equivalent $g_n = d_{\gamma_n} (n=1, \ldots, N)$ with $\gamma_n \in \{1, \ldots, M\}$. Then the reconstruction error is defined in following form:

$$||\vec{R}_N||^2 = ||\vec{y} - \overrightarrow{q_N}||^2 = \sum_{i=1}^{I} (y_i - q_N(x_i))^2 \tag{3}$$

where, $\vec{R}_N = \vec{y} - \overrightarrow{q_N}$ is the residual, $\overrightarrow{q_N}$ corresponds to the evaluation of q on I training samples in form

$$\overrightarrow{q_N} = (q(x_1), q(x_2), \ldots, q(x_I)) \tag{4}$$

In matching pursuit algorithm, the basis function$\{g_1, \ldots, g_N\}$ and these corresponding coefficients $\{\alpha_1, \ldots, \alpha_N\}$ are selected by applying the iterative greedy method, in the interest of minimizing the reconstruction error $\vec{R}_N$. The form of the coefficient α is shown in following:

$$\alpha = \langle \vec{g}, \vec{R}_n \rangle / ||\vec{g}||^2 \tag{5}$$

According to the description given above, the matching pursuit algorithm is a method that searches for a set of the basis functions that can correspond to the maximized approximations of observational values with minimizing the reconstruction error by using a greedy method.

3.2. Matching Pursuit Optimization Ensemble Classifiers

Initially, in order to change a weak classifier into a strong one and made a good classification performance, classifiers ensemble is proposed. Hansen and Salamon [10] put forward a hypothesis that combining models was the most effective, when the individual model of the ensemble system was independence with each other. Subsequently, Tumer and Ghosh [14] indicated the relation of the generalization error and every classifier and gave an Eq. (6) to express it, which was shown in following:

$$\hat{Error} = \frac{1+\rho(N-1)}{N} Error + Error_{OptimalBayes} \tag{6}$$

where, ρ indicates the agreement among the error of classifiers, $Error_{Optimal\ Bayes}$ indicates the error of recognition gained with using the Bayes rule in the case of all conditional probability given. When $\rho=0$, the error of ensemble system decreases in proportion with the number of classifiers increasing; when $\rho=1$, the error of ensemble system is equal to error of single classifier. In 2000, Zhou et al. [47] gave Eq. (7), which indicated that the diversity among classifiers and the error of individual classifier affected the generalization error of ensemble system

$$Error = \overline{Error} - \overline{D} \tag{7}$$

where, $Error$ indicates the generalization error of ensemble system, $\overline{Error}$expresses the error of individual classifier, $\overline{D}$ shows the diversity among all classifiers. Hence, according to the Eqs. (6) and (7), it is the expectation that the classifiers should be as differential as possible and the error of individual should be small at the same time, then the generalization performance of ensemble system will be enhanced. Presently, increasing the diversity $\overline{D}$ is important to construct ensemble classifiers.

However, the high diversity may bring some poor performance classifiers, which should affect the performance of ensemble system according to the form (7). Consequently, it is crucial that how to choose some beneficial classifiers for ensemble with the diversity preserved.

In this paper, a new method that Matching Pursuit Optimization Ensemble Classifiers (MPOEC) is proposed in order to eliminate some useless or similar classifiers from the whole ensemble classifiers to improve the performance of ensemble without decreasing the diversity among classifiers. The MPOEC approach adopts a greedy iterative approach to search for an optimal combination of ensemble classifiers, and this combination must satisfy with minimizing the residual R between the target function and the optimal combination. In this method, the set of all ensemble classifiers $\{C_l(x)\}(l=1, \ldots, L)$ is regarded as the basis function dictionary D, and every classifier can obtain a coefficient α_l by minimizing R. When $\alpha_l \neq 0$, the classifier corresponding to α_l is selected to ensemble, and when $\alpha_l=0$, the classifier should be eliminated. The description of MPOEC algorithm is shown in Algorithm 1.

In MPOEC algorithm, at first, L training sets $\{X_1, \ldots, X_L\}$ are gained from original training set by using an ensemble strategy, and component classifiers $\{C_1(x), \ldots, C_L(x)\}$ are obtained by training, respectively, $\{X_1, \ldots, X_L\}$. Secondly, the coefficients $\{\alpha_1, \ldots, \alpha_L\}$ corresponding to L ensemble classifiers are given by minimizing the residues R, and classifiers with nonzero coefficients α_l are selected to ensemble, then the optimal combination f_{opt} is obtained. Finally, the optimal prediction of ensemble classifiers is given by the following form:

$$h_{opt} = sign(f_{opt}) = sign\left(\sum_{l=1}^{L_{opt}} \alpha_l f_l^*\right), \quad l=1,\ldots,L_{opt} \tag{8}$$

In addition, the error e_t is computed by form (9) in each iterative process

$$e_t = \sum_{i=1}^{N}\left[y_i - sign\left(\sum_{j}^{loop} \alpha_j f_{ji}\right)\right] \Big/ N \tag{9}$$

Algorithm 1. Matching Pursuit Optimization Ensemble Classifiers algorithm (MPOEC algorithm)

Input: Original training set $X_{training}=\{(x_1,y_1), \ldots, (x_N,y_N)\}$, testing set $X_{testing}=\{x_1, \ldots, x_{N^*}\}$, the number of classifiers L, the iterative number T, a threshold error λ.

Procedure:

- **For** $l=1$ to L
  - X_l=gained from $X_{training}$ by an ensemble strategy
  - $C_l(x)$=the classifier learning for X_l
  - $f_l(x)$=the prediction labels for $X_{training}$ by $C_l(x)$
  - $f_l^*(x)$=the prediction labels for $X_{testing}$ by $C_l(x)$

  End for
- Search for the optimization combination of L classifiers
  - * Prediction labels $\{f_1(x), \ldots, f_L(x)\}$ are considered to be the basic functions$\{g_1, \ldots, g_M\}$;
  - * The initial residue R_0 is equal to the class labels $\{y_1, y_2, \ldots, y_N\}$;
  - * **While** (loop $<=T$ & error $>\lambda$)
    - – Obtain initial α_{ini} with the formula:$\alpha = \left\langle \vec{f}, \vec{R_t} \right\rangle / \|\vec{f}\|^2$;
    - – $\alpha_{max} = \alpha_{ini}$;
    - – **For** $i=1, \ldots, L$;

      Compute α_i by the above formula;

      If $(||\alpha_i|| > ||\alpha_{max}||)$

      Note α_i and update $\alpha_{max} : \alpha_{max} = \alpha_i$;

      Else

      Give a zero value to $\alpha_i : \alpha_i=0$;

      End if

      End for
    - – Note $\alpha_t : \alpha_t = \alpha_{max}$, and compute the residue $R_t^* : R_t{}^* = R_t - \alpha_t f_t$;
    - – Update the residue $R_t : R_t = R_t^*$;
    - – Update the number of iteration: $loop=t$;
    - – Compute the error e_t;

    End while
  - * Gain the coefficients of ensemble classifiers $\{\alpha_1, \ldots, \alpha_L\}$ and select the classifiers of the nonzero α_l to ensemble.

Output: The combination of the optimal ensemble classifiers:

$$f_{opt} = \sum_{l}^{L_{opt}} \alpha_l f_l^*, \quad (l=1,\ldots,L_{opt}, \quad L_{opt} \le L)$$

Where, y_i is the label of a training sample, $loop$ is the number of the iteration, N is the number of training samples, α_j is the coefficient gained in jth iteration, f_{ji} is the prediction label given by the classifier with α_j coefficient for the ith training sample in jth iteration.

In accordance with the description of the proposed algorithm, it is found that the MPOEC approach actually selects classifiers to ensemble based on the diversity of classifiers and the performance of classifiers, which is the reason that when a pair of classifiers are similar to each other, the coefficient of one classifier will be close to zero, even be equal to zero. The detailed analysis is shown in the section of the diversity analysis. Hence, the agreement ρ in the Eq. (6) can be decreased by MPOEC, and then the error of ensemble can be decreased. For MPOEC algorithm, the initialization of the residue R_0 is the class labels of samples, so in first iterative process ($t=1$) the obtained coefficient α is satisfied to maximize the form

$$\alpha_{max} = \| \langle \vec{f}, \vec{R}_t \rangle / || \vec{f} ||^2 \| \tag{10}$$

When the prediction function f is more similar to the class labels of samples, the coefficient α is bigger, so the classifier corresponding to the coefficient α_{max} may be the most similar to the labels Y. All latter classifiers minimize the difference between the idea labels Y and the combination q_t of anterior classifiers

$$q_t = \sum_{i=1}^{t-1} \alpha_i f_i + \alpha_t f_t \tag{11}$$

It indicates that the MPOEC approach can retain some classifiers that have good performance of classification. However, the anterior gained coefficients can be adaptively modulated by back fitting method in MPOEC algorithm, so the basis function f that is the most similar to the ideal label Y may not always gain the max value of the coefficient α.

Especially, we find that the coefficients of the classifiers may be given negative values in the MPOEC algorithm, when the basis function f has a negative correlation to the residue R. For a classification problem, when the target function Y is $y_i \in \{-1, 1\}$ $(i=1, \ldots, N)$, the negative correlation means that the prediction function is very disagree with the idea class label for a classification problem. Hence, we also proposed an approach OPMPOEC that selecting the classifiers of the positive coefficients to ensemble. The OPMPOEC approach adds a restriction $\alpha > 0$ to the initial condition $||\alpha_i|| > ||\alpha_{max}||$ for noting α and updating α_{max} in each iterative process.

3.3. MPOEC based on SVM ensemble with bagging algorithm

Specially, in this part, SVM ensemble with bagging strategy by applying MPOEC algorithm is introduced in order to demonstrate the proposed method favorably, but which does not predicate that MPOEC method is not adaptive to other strategies and other classifier models. Then bagging algorithm is shown in Algorithm 2.

As follows, the MPOEC approach is introduced from the view of classifier models.

Algorithm 2. Bagging algorithm

Input: The training samples $(x_1, y_1), \ldots, (x_m, y_m)$, where $x_i \in X$, $y_i \in Y$, $Y = \{-1, +1\}$

For $l = 1, \ldots, L$

* Bootstrap resample from the training sample and gain a set X_l;

* Learn for the samples X_l by the classifier model and gain a classifier $C_l(x)$;

End

Output: The final prediction $H(x) = \arg\max \sum_{l=1}^{L} C_l(x)$.

For an SVM classifier [20], the testing samples are classified by the discriminant function $f(x)$ in form (12), which is the expression of the hyperplane H_{opt} given by the support vectors in fact, as shown in Fig. 4(b)

$$f(x) = wx + bias \tag{12}$$

where, w is normal vector to the hyperplane, $|bias|/||w||$ is the perpendicular distance from the hyperplane to the origin. Hence, in an SVM ensemble system, every classifier gains an optimal hyperplane with $(w_l, bias_l)$ $(l = 1, \ldots, L)$ by learning for its training samples. In MPOEC method, we consider the all hyperplanes $\{(w_1, bias_1), (w_2, bias_2), \ldots, (w_L, bias_L)\}$as basis function $\{g_n\}$. By the greedy iterative method, helpful hyperplanes $\{(w_1, bias_1), \ldots, (w_{L_{opt}}, bias_{L_{opt}})\}$for ensemble classifiers are searched and the nonzero coefficients $\{\alpha_1 \ldots, \alpha_{L_{opt}}\}$ are obtained. Then the optimal combination is shown in following form:

$$f_{opt} = \sum_{t=1}^{L_{opt}} \alpha_t f_t, \quad t = 1, \ldots, L_{opt}, \; L_{opt} \le L \tag{13}$$

where, L_{opt} is the number of classifiers that are searched for nonzero coefficients. According as the form (12), it is gained

$$f_{opt} = \sum_{t=1}^{L_{opt}} \alpha_t (w_t x + bias_t) = \sum_{t=1}^{L_{opt}} \alpha_t w_t x + \sum_{t=1}^{L_{opt}} \alpha_t bias_t \tag{14}$$

Suppose we give $w_{opt} = \sum_{t=1}^{L_{opt}} \alpha_t w_t$ and $bias_{opt} = \sum_{t=1}^{L_{opt}} w_t bias_t$ for above forms, well then f_{opt} is decided by the form

$$f_{opt} = w_{opt} x + bias_{opt} \tag{15}$$

According to the form (15), it is shown that $(w_{opt}, bias_{opt})$ expresses a new hyperplane of classification searched by the proposed algorithm. In Fig. 2, (b) shows several hyperplane $\{(w_l, bias_l)\}$ of ensemble individual classifier, and an optimal hyperplane $(w_{opt}, bias_{opt})$ is shown in (c) by MPOEC algorithm. Note that the final prediction label is obtained by

$$h_{opt} = sign(f_{opt}) = sign(w_{opt} x + bias_{opt}) \tag{16}$$

However, for the traditional ensemble, the final prediction function h_{fin} is given by the voting rule during the procedure of combining classifiers, as shown in following:

$$h_{fin} = sign\left(\sum_{l}^{L} sign(f_l)\right) = sign\left(\sum_{l}^{L} sign(w_l x + bias_l)\right) \tag{17}$$

where, L is the number of ensemble classifiers. According to the above form, it has been found that for a testing sample, when the number of classifiers gaining the correct label for a testing sample is more than $L/2$, ensemble classifiers will obtain a right label. Contrarily, it will obtain a wrong label. Whereas, in the MPOEC approach, obtaining a right label or a wrong one is decided by the optimal $(w_{opt}, bias_{opt})$, this can improve the disadvantage that every classifier has same probability to affect the performance of ensemble.

Algorithm 3. Random subspace algorithm

Input: The training samples $(x_1, y_1), \ldots, (x_m, y_m)$, where $x_i \in X$, $y_i \in Y$, $Y = \{-1, +1\}$; M is the number of the feature of samples;

For $l = 1, \ldots, L$

* Randomly select a subspace with $m (m < M)$ features from M features, and gain a training subset X_l;

* Learn for the samples X_l by the classifier model and gain a classifier $C_l(x)$;

End

Output: The final prediction $H(x) = \arg\max \sum_{l=1}^{L} C_l(x)$.

Additionally, the simple ensemble strategies, such as bagging and random subspace, are chosen to construct the individual classifiers in experiments, because simpler strategies could illustrate the proposed method favorably, in other word, MPOEC can improve the diversity and accuracy of ensemble connected with selecting classifiers, not with constructing training sets for classifiers. In Algorithm 3, random subspace algorithm is described.

4. The diversity of classifiers for MPOEC algorithm

For the classical ensemble strategies such as bagging and boosting, the diversity of classifiers is gained justly by the different training samples. But the different training samples could not ensure that each classifier is absolutely different from others in an ensemble system. However, the MPOEC algorithm combines the

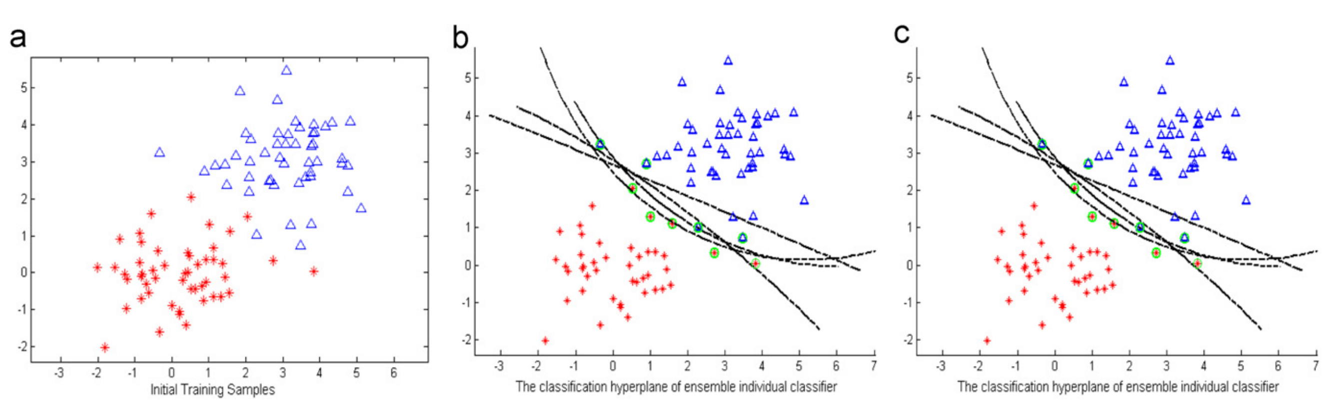

Fig. 2. A Gaussian distribution dataset with 100 training samples in two-dimensional space. Blue triangle stands for the positive data and red dot stands for the negative data. Darkened lines denote the hyperplane of classification. (a) The distribution of initial training samples. (b) The hyperplanes of four ensemble individual classifiers selected randomly. (c) The optimal hyperplane of ensemble classifiers by MPOEC. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

classifiers based on the diversity between a pair of classifiers, and an analysis of the diversity is given in the following description. According to the description of the MPOEC method in Algorithm 1, searching for a classifier is based on minimizing the residual R, and its coefficient α is gained by the form (10). Hence, two forms are given for the tth ($t=1, \ldots, T$) basis function (namely, one classifier of the ensemble system)

$$\alpha_t = \langle f_t, R_{t-1} \rangle / ||f_t||^2 \tag{18}$$

$$R_t = R_{t-1} - \alpha_t f_t \tag{19}$$

where, f_t represents the basis function in the tth iteration, α_t is the coefficient of the basis function f_t and R_t is the residual in the tth iteration. And then the coefficient α_i of each basis function f will be computed by the following form in the $t+1$th iteration

$$\alpha_i = \langle f_i, R_t \rangle / ||f_i||^2, \quad (i=1,\ldots,L) \tag{20}$$

By the form (19), α_{t+1} may be expressed

$$\begin{aligned} \alpha_i &= \frac{\langle f_i, R_t \rangle}{||f_i||^2} = \frac{\langle f_i, R_{t-1} - \alpha_t f_t \rangle}{||f_i||^2} = \frac{\langle f_i, R_{t-1} \rangle - \langle f_i, \alpha_t f_t \rangle}{||f_i||^2} \\ &= \frac{\langle f_i, R_{t-1} \rangle}{||f_i||^2} - \alpha_t \frac{\langle f_i, f_t \rangle}{||f_i||^2} \end{aligned} \tag{21}$$

where, f_t expresses the basis function obtained the coefficient α_t, and $\{f_i\}$ ($i=1, \ldots, L$) expresses the set of all basis functions in the $t+1$th iteration. For the Eq. (21), if the function f_i is similar to f_t, two forms will be obtained

$$\lim_{f_i \to f_t} \frac{\langle f_i, R_{t-1} \rangle}{||f_i||^2} = \alpha_t \tag{22}$$

$$\lim_{f_i \to f_t} \frac{\langle f_i, f_t \rangle}{||f_i||^2} = 1 \tag{23}$$

Then the form (21) may be expressed according as the above forms (22) and (23)

$$\lim_{f_i \to f_t} \alpha_i = 0 \tag{24}$$

Because the function f_{t+1} gains the coefficient α_{t+1}in the $t+1$th iteration, when f_{t+1} is satisfied the following inequation:

$$\| \langle f_i, R_t \rangle / ||f_i||^2 \| \le \| \langle f_{t+1}, R_t \rangle / ||f_{t+1}||^2 \| \tag{25}$$

According as these equations and the inequation, when a function f_i is similar to the function f_t, the probability that the function f_i gains a max coefficient α_i may be close to zero in the $t+1$th iteration, so this classifier will not obtain a nonzero coefficient.

In the MPOEC algorithm, the basis function dictionary is composed of the whole ensemble classifiers, in other words, a basis function represents an individual classifier. Hence, according to the above analysis, it is obvious that a classifier C_l($l=1, \ldots, L$) has gained a nonzero coefficient α_l and its corresponding basis function is f_t, if the basis function f_i($i=1, \ldots, L$)is similar to f_t, the classifier C_m($m=1, \ldots, L$) corresponding to f_i will obtain a coefficient α_i whose value is approach to zero, so this classifier C_m may be not selected to ensemble. On the contrary, obtaining a coefficient which is not close to zero shows that there is the diversity between the former classifier and the latter one. Summarily, the above analysis indicates that the MPOEC approach selects classifiers according to the diversity between a pair of classifiers, and it can neglect the effect of some similar classifiers by giving the zero or close-zero coefficients α to these classifiers C in order to increase the diversity of the ensemble system.

In the optimizing process, the ideal prediction $\{y_1, \ldots, y_N\}$ is considered as the initial residual R_0. Hence, in initial iterative process, the classifier selected by the MPOEC algorithm is the most approximate of the target function, and it means that the classifier may be the best performance of ensemble classifiers, then in following iterative processes, the selected classifiers are added to the anterior combination of classifiers in order to be close to the target label at best. Therefore, the increase in the diversity will not reduce the accuracy of individual classifiers in the optimization process of MPOEC approach.

5. Experiments and analysis

For the comparison of the different models we selected 23 datasets from UCI Machine Learning Repository [48]. The description of the datasets is shown in the following part. In experiments, the available data is randomly partitioned into two disjunctive parts: one is used as training set for each learning algorithm and the other is tested on. The proportion of training set is approximately 6–50% of all samples, which is decided based on the number of dataset. For instance, only the 12% of the whole samples is used as original training set for waveform dataset, the 48% is used as training set for sonar dataset. Especially, in order to elucidate the performance of the MPOEC approach, four artificial datasets are also chosen in experiments. Because the artificial data has low-dimension character, the hyperplanes of classification can be described easily with 2D pictures.

At present, the ensemble classifiers algorithms have been constructed mostly from the part of training individual classifiers to construct the training set of classifiers in the ensemble system, but the MPOEC method balances the diversity and the accuracy from the combination of classifiers. In experiments, we select two ensemble strategies, bagging [1] and random subspace [58], with simpler-constructing training sets, which can be more benefit for exhibiting the performance of our method than other ensemble strategies, especially the diversity among individual classifiers. For each dataset, the radio of the bootstrap resample in bagging algorithm is 50% to gain a new training set, and 60% features are selected randomly to construct a new training set in random subspace algorithm.

It is clear that selecting different classifier model may have the dissimilar effects on the results of the proposed algorithm. Thus, we select two different basis classifier models: the C4.5 algorithm [49] with back-pruning and a support vector machine [20] with a Gaussian kernel. For C4.5 we use 20% to be the percentage of incorrectly assigned samples at a node, and for the support vector machine model, the penalty factor of each dataset applies to $C=2^{10}$, and the parameter σ of the kernel function is given by the cross-validation.

Furthermore, in order to elucidate the performance of the proposed method for classification problems of pattern recognition as comprehensive as possible, our experiments are composed of three parts. Firstly, the classification performance of the proposed method is shown elementarily by the classification problem of four artificial datasets. Specially, it also is illustrated visually with drawing the pictures of the hyperplane. Secondly, UCI datasets are used to validate the performance of our method with two classical ensemble strategies, and we also compare MPOEC algorithm with several ensemble methods mentioned previously. Ultimately, the kappa-error diagrams of the diversity among classifiers are demonstrated, which testifies that our method improves the diversity of classifiers compared with other ensemble methods.

The algorithms in experiments are coded by matlab R2009b and implemented with matlab R2009b, and our experiments are executed on a workstation with HP xw9400 2.4 GHz AMD Opteron, 32G memory and Windows XP 32 operation system. In addition, the source codes are freely available upon request to the authors by e-mail.

5.1. Artificial datasets

Because the low dimension datasets can be visualized, we employ originally four typical artificial datasets to illustrate in experiments, which obey, respectively, the Gaussian distribution, the semicircular distribution, the hyperbolic distribution and the spiral distribution. The distributions of four datasets are shown in Fig. 3. Especially, we apply bagging strategy and SVM classifier model to the MPOEC method. The hyperplane of classification can be drawn in 2D datasets for the SVM classifier, and bagging algorithm can reduce the computation of each individual classifier by selecting a small portion of the whole training set, which is beneficial to learn for SVM. For each artificial dataset, we produce randomly 5000 training samples and 2000 testing samples (or 2828 samples). Note that the training set of each individual classifier contains 500 samples, which selected randomly from original training samples. The parameters σ of SVM are 8, 8, 2 and 4 for four datasets, respectively. In addition, we utilize 30 SVM classifiers to ensemble, and the results of classification are shown in Table 1.

In Table 1, 'Acc' denotes the accuracy rate of classification, 'NSC' denotes the number of classifiers selected to ensemble, and 'NC' indicates the number of the whole ensemble classifiers. According to the results in Table 1, it is seen that our method may outperform bagging ensemble and single SVM classifier appreciably, especially, MPOEC justly selects less than five classifiers to ensemble and gains better performance than others for the front three datasets. However, only the spiral dataset, the accuracy is lower than the method of ensemble with bagging. We give some pictures about the hyperplane of classifier to demonstrate in Fig. 4, and the three hyperplanes are arbitrarily chosen from all hyperplanes of each artificial dataset in an ensemble process. It is shown that the proposed method is actually a process that the hyperplane of ensemble classifiers is improved.

In an ensemble system, each classifier produces a hyperplane, obviously, but there are not only some good (useful) ones but also some poor (useless so much as baneful) ones as shown in Fig. 2(b) in all hyperplanes. The MPOEC method should make an advisable selection for those hyperplanes. The useful individual classifiers are reserved, at the same time, the useless or harmful ones could be eliminated. The proposed method can achieve improvement of ensemble performance.

5.2. UCI datasets

In this section, UCI datasets are utilized to validate the validity of the proposed method. We also make a detailed comparison with the standard methods. A summary of the datasets is shown in Table 2. In this experiment, we select two classifier models as the

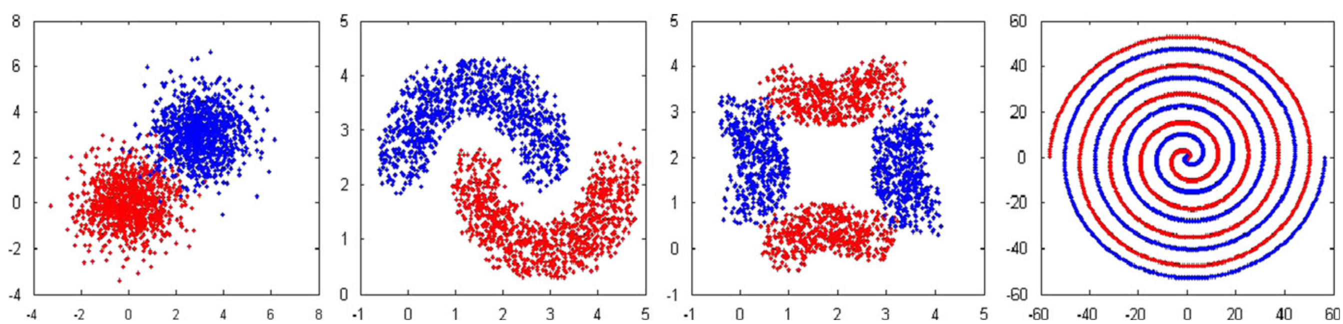

Fig. 3. The distributions of four artificial datasets.

Table 1
Results of four artificial datasets recognized by 30 SVM classifiers ensemble.

Datasets distribution	Number of samples		MPOEC		OPMPOEC		Ensemble with bagging		Single SVM
	EC-Trn	Testing	Acc	NSC	Acc	NSC	Acc	NC	
Gaussian	500	2000	**0.9855**	4	0.985	3	0.9845	30	0.981
Semicircular	500	2000	**0.9695**	4	0.9655	3	0.963	30	0.962
Hyperbolic	500	2000	**0.995**	4	**0.995**	4	**0.995**	30	0.98
Spiral	500	2828	0.8851	11	0.8851	11	**0.8854**	30	0.8221

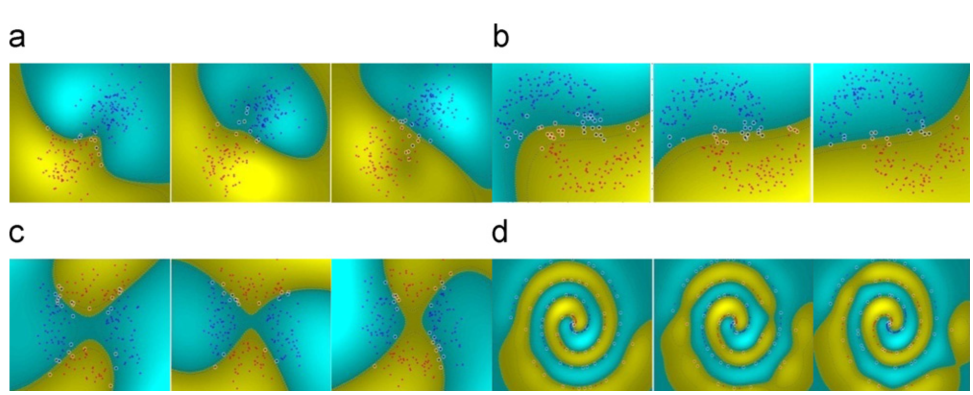

Fig. 4. Three arbitrary hyperlanes for four artificial datasets: (a) Gaussian distribution. (b) Semicircular distribution. (c) Hyperbola distribution, and (d) Spirals distribution.

ensemble individuals to learn and classify for UCI datasets, respectively. One is a weak classifier model applied popularly in ensemble learning, decision tree C4.5, the other is a strong classifier model, support vector machine, which has been recently used to deal with classification problem in the fields of not only pattern recognition but also ensemble learning. Additionally, bagging and random subspace algorithms are applied to construct ensemble individuals.

Table 2
Summary of characteristics of the datasets used in experiments.

Datasets	Cases	Features			Input	Classes
		Continue	Binary	Nominal		
Balance	625	4	–	–	4	3
Breast	277	9	–	–	9	2
Chart	600	60	–	–	60	6
Clean	476	166	–	–	166	2
Glass17	99	9	–	–	9	2
Glass25	90	9	–	–	9	2
Glass	214	9	–	–	9	7
Heart_c	303	6	1	6	13	2
Ionosphere	351	33	1	–	34	2
Iris	150	4	–	–	4	3
Liver	345	6	–	–	6	2
Musk	6598	166	–	–	166	2
Pima	768	8	–	–	8	2
Sat	6435	36	–	–	36	6
Shuttle	14500	–	–	9	9	7
Sonar	208	60	–		60	2
Spambase	4601	57	–		57	2
Vehicle	846	18	–		18	4
Vot	435	–	16		16	2
Waveform	5000	40	–	–	40	3
Wdbc	569	30	–	–	30	2
Wine	178	13	–	–	13	3
Wpbc	198	33	–	–	33	2

This experiment distinguishes from the former, and the numbers of ensemble classifiers are 100 instead of 30, which aims at revealing the sparsity of the proposed method. Our method will select a few classifiers to ensemble. For each dataset, the parameter σ is selected respectively by the 10 fold cross-validation for the original training set. It will be used in each individual classifier in order to ensure the diversity among classifiers that is irrelevant with the parameter and simple the process of producing classifiers. For optimal parameters, the threshold error is decided according to the accuracy of training samples of each dataset (note that the most cases are $\lambda = 0.003$), and the iterative numbers T are 80 corresponding to the numbers of ensemble classifiers 100. In addition, for each dataset and each method, the results of the experiment are the average of 50 times. As follows, the tables show the results of all experiments.

Tables 3 and 4 show the experiments results of C4.5 tree and SVM ensemble with 100 individual classifiers, respectively. In tables, 'MPOEC' and 'OPMPOEC' denote the proposed methods, 'Standard' denotes the standard ensemble method, 'Single Classifier' denotes the accuracy of original training samples by single basic classifier algorithm, 'Accuracy' is the accuracy rate and bias of classification for datasets and 'NSC' indicates the number of classifiers which obtained nonzero coefficients, in the other word, it is the number of classifiers selected from 100 individual by our method in ensemble. The tables also show the results using a single classifier to classify for datasets.

In Tables 3 and 4, it is obvious that our algorithm exceeds the others at the accuracy of classification for the most datasets, such as balance, breast, chart, iris, liver, sat, vehicle, vot, waveform, wdbc, wine, wpbc and so on, and the results are bold. For C4.5 classifier, the proposed method improves the performance of standard methods for bagging and RSM more than 2% for most datasets, especially, like balance, heart_c, vehicle and waveform. The results are even better for SVM as base classifier. According as the results of 23 UCI datasets, the MPOEC method obtains higher accuracy for 18 datasets and 16 datasets than standard bagging and standard RSM

Table 3
The results for ensemble 100 classifiers using a C4.5 tree as the basis learner.

Datasets	Bagging strategy					Random subspace strategy					Single classifier
	MPOEC		OPMPOEC		Standard	MPOEC			OPMPOEC	Standard	
	Accuracy	NSC	Accuracy	NSC	Accuracy	Accuracy	NSC	Accuracy	NSC	Accuracy	Accuracy
Balance	**0.7835 ± 0.0577**	19	0.7184 ± 0.0445	12	0.6808 ± 0.0096	**0.8104 ± 0**	4	0.7126 ± 0	4	0.7713 ± 0.0489	0.7126
Breast	**0.7331 ± 0.033**	55	0.7292 ± 0.033	30	0.727 ± 0.0076	0.7252 ± 0.0152	27	**0.7264 ± 0.0076**	11	0.7204 ± 0.0076	0.7208
Chart	**0.9821 ± 0.006**	3	0.9767 ± 0.005	2	0.9801 ± 0.004	0.9821 ± 0.0167	21	0.9833 ± 0.0127	13	**0.9914 ± 0.0033**	0.9813
Clean	**0.8186 ± 0.0483**	26	0.797 ± 0.0682	22	0.8032 ± 0.0256	0.8028 ± 0.0455	24	0.795 ± 0.0881	17	**0.8385 ± 0.0199**	0.8011
Glass17	0.8531 ± 0.0306	1	0.8531 ± 0.0306	1	**0.8584 ± 0.0102**	**0.858 ± 0.1429**	1	0.858 ± 0.1429	1	0.8571 ± 0	0.8571
Glass25	0.8414 ± 0.044	1	0.8414 ± 0.0114	1	**0.8459 ± 0.0341**	**0.8623 ± 0.0227**	1	0.8309 ± 0.0341	1	0.8309 ± 0.0341	0.8409
Glass	**0.8600 ± 0.0451**	3	0.8568 ± 0.0516	2	0.8383 ± 0.0449	0.8487 ± 0.1182	4	0.8507 ± 0.1125	3	**0.8632 ± 0.0268**	**0.8641**
Heart_c	**0.7938 ± 0.0493**	31	0.7639 ± 0.0542	18	0.7655 ± 0.0172	**0.8098 ± 0.032**	16	0.7776 ± 0.0369	10	0.7726 ± 0.0029	0.7635
Ionosphere	**0.9319 ± 0.0629**	29	0.903 ± 0.0762	16	0.8674 ± 0.053	0.9017 ± 0.0066	25	0.9034 ± 0.0265	16	**0.9081 ± 0.0099**	0.9078
Iris	0.9746 ± 0.011	6	**0.9802 ± 0.0083**	4	0.9788 ± 0.0056	**0.9778 ± 0**	3	0.9778 ± 0	2	0.9287 ± 0.0389	0.9778
Liver	**0.6408 ± 0.0378**	43	0.6381 ± 0.0844	29	0.6332 ± 0.0533	0.6267 ± 0	8	0.6267 ± 0	2	0.6267 ± 0	0.6267
Musk	0.9397 ± 0	1	0.9397 ± 0	1	0.9397 ± 0	0.9397 ± 0	1	0.9397 ± 0	1	0.9397 ± 0	0.9397
Pima	0.7379 ± 0.0238	56	0.7286 ± 0.0299	17	**0.7413 ± 0.0053**	**0.7413 ± 0.0035**	32	0.7423 ± 0.0053	9	0.7065 ± 0.0405	0.7038
Sat	**0.9655 ± 0.0025**	21	0.9592 ± 0.0046	16	0.9501 ± 0.0006	**0.9634 ± 0.0037**	18	0.9578 ± 0.0076	14	0.9495 ± 0.0006	0.9458
Shuttle	**0.9992 ± 0**	2	0.9992 ± 0	2	0.9807 ± 0.0002	**0.994 ± 0.0043**	4	0.9904 ± 0.0058	2	0.9817 ± 0.0012	0.9794
Sonar	**0.698 ± 0.0602**	37	0.6802 ± 0.0602	20	0.6506 ± 0.0324	0.6856 ± 0.1065	9	0.6856 ± 0.1065	8	**0.6937 ± 0.0324**	0.6204
Spambase	**0.8816 ± 0.0181**	39	0.8721 ± 0.0522	32	0.879 ± 0.0109	**0.8697 ± 0.025**	31	0.8599 ± 0.0512	23	0.8674 ± 0.0184	0.8257
Vehicle	**0.833 ± 0.0238**	34	0.8246 ± 0.0304	20	0.8132 ± 0.0151	**0.8349 ± 0.0145**	23	0.8295 ± 0.0111	12	0.8251 ± 0.0093	0.8169
Vot	**0.9359 ± 0.0191**	13	0.9354 ± 0.0191	9	0.9237 ± 0.0021	0.9212 ± 0.0043	14	**0.9234 ± 0**	10	0.9195 ± 0.0085	0.9234
Waveform	**0.8497 ± 0.0099**	33	0.8377 ± 0.0162	21	0.8149 ± 0.0066	**0.8291 ± 0.0137**	24	0.816 ± 0.0183	15	0.8073 ± 0.0043	0.8046
Wdbc	**0.9114 ± 0.0285**	16	0.8882 ± 0.065	11	0.9095 ± 0.0312	**0.9175 ± 0.019**	7	0.9103 ± 0.0244	6	0.8982 ± 0.0257	0.8997
Wine	**0.9591 ± 0.0251**	11	0.9449 ± 0.0251	7	0.952 ± 0.0502	**0.9782 ± 0.0137**	4	0.9782 ± 0.0137	3	0.928 ± 0.0137	0.9087
Wpbc	**0.6404 ± 0.0816**	29	0.6351 ± 0.0714	16	0.6357 ± 0.0255	**0.6767 ± 0.1071**	10	0.6745 ± 0.102	7	0.6539 ± 0.0918	**0.6837**
Average	**0.8507 ± 0.0312**	22	0.8349 ± 0.0366	13	0.8334 ± 0.0194	**0.8503 ± 0.0311**	13.5	0.8413 ± 0.0351	8.3	0.8382 ± 0.0191	0.8307

Table 4
the results for ensemble 100 classifiers using SVM as the basis learner.

Datasets	Bagging strategy					Random subspace strategy					Single classifier
	MPOEC		OPMPOEC		Standard	MPOEC		OPMPOEC		Standard	
	Accuracy	NSC	Accuracy	NSC	Accuracy	Accuracy	NSC	Accuracy	NSC	Accuracy	Accuracy
Balance	**0.9521 ± 0.0052**	10	0.9455 ± 0.013	8	0.9492 ± 0.0024	0.9289 ± 0	3	**0.9526 ± 0**	3	0.8314 ± 0.0489	**0.9541**
Breast	**0.7509 ± 0.0054**	12	0.7383 ± 0.0148	7	0.718 ± 0.0033	**0.7599 ± 0.0076**	4	0.7148 ± 0.0025	2	0.7107 ± 0	0.7157
Chart	0.9434 ± 0.02	2	**0.9441 ± 0.0217**	2	0.9367 ± 0.0054	1 ± 0	1	1 ± 0	1	1 ± 0	0.936
Clean	**0.8119 ± 0.015**	7	0.7857 ± 0.0234	6	0.8005 ± 0.0227	0.8267 ± 0.054	1	0.8267 ± 0.054	1	**0.8314 ± 0.0142**	0.7906
Glass17	**0.9971 ± 0.0071**	1	0.9971 ± 0.0071	1	0.8477 ± 0.0969	**0.9996 ± 0.0102**	11	0.9922 ± 0.0102	6	0.9102 ± 0.0102	0.7143
Glass25	**0.9082 ± 0.0121**	14	0.8582 ± 0.0542	8	0.8577 ± 0.0969	0.8841 ± 0.0341	10	0.8605 ± 0.0455	6	**0.9205 ± 0.0114**	0.8864
Glass	**0.9026 ± 0.0264**	10	0.8905 ± 0.0351	6	0.8786 ± 0.0121	**0.9033 ± 0.0174**	4.7	0.8926 ± 0.0167	3	0.8878 ± 0.0198	0.8653
Heart_c	0.8299 ± 0.0066	17	0.8169 ± 0.0194	11	**0.8332 ± 0.0043**	**0.835 ± 0.0146**	19	0.752 ± 0.034	11	0.8289 ± 0.0194	0.8252
Ionosphere	**0.9433 ± 0.0072**	15	0.9307 ± 0.0087	9	0.9193 ± 0.0023	0.955 ± 0.0199	24	0.9682 ± 0.0166	12	**0.975 ± 0.0066**	0.9404
Iris	**0.9741 ± 0.0089**	1	0.9737 ± 0.0098	1	0.9691 ± 0.0032	**0.9889 ± 0**	1	0.9889 ± 0	1	0.9841 ± 0.0083	0.9667
Liver	**0.7154 ± 0.017**	14	0.706 ± 0.0255	9	0.6923 ± 0.0128	**0.7243 ± 0.0044**	14	0.6666 ± 0.0022	10	0.708 ± 0.0267	0.6327
Musk	**0.9401 ± 0.0002**	47	0.9399 ± 0.0002	30	0.9397 ± 0	0.9372 ± 0.005	1	0.9322 ± 0.005	1	**0.9414 ± 0.003**	0.9386
Pima	**0.7563 ± 0.0071**	29	0.7441 ± 0.0084	17	0.7411 ± 0.0042	**0.7768 ± 0.0035**	15	0.7621 ± 0.0035	9	0.763 ± 0.007	0.7588
Sat	**0.9624 ± 0.0037**	1	0.9622 ± 0.0039	1	**0.9691 ± 0.0011**	**0.9405 ± 0.0036**	48	0.9388 ± 0.0037	26	0.9381 ± 0.0011	0.9602
Shuttle	**0.9579 ± 0.0023**	27	0.9508 ± 0.0041	22	0.9313 ± 0.0016	0.9453 ± 0.0131	59	0.9409 ± 0.013	42	**0.9537 ± 0.0012**	0.9234
Sonar	0.8111 ± 0.0273	4	0.7978 ± 0.0289	3	**0.8222 ± 0.0131**	0.8141 ± 0.0463	1	0.8141 ± 0.0463	1	**0.8169 ± 0.0231**	0.8148
Spambase	**0.7679 ± 0.0062**	47	0.7639 ± 0.0062	34	0.7539 ± 0.0037	**0.7678 ± 0.0147**	5	0.7516 ± 0.02	5	0.7607 ± 0.0081	0.7245
Vehicle	**0.8722 ± 0.006**	14	0.8579 ± 0.0088	8	0.8576 ± 0.0031	**0.8825 ± 0.0083**	15	0.8633 ± 0.0129	9	0.8605 ± 0.0048	0.8647
Vot	**0.9351 ± 0.0031**	18	0.9308 ± 0.0048	12	0.9351 ± 0.0039	**0.9277 ± 0**	15	0.9277 ± 0	8	0.9275 ± 0.0043	0.9191
Waveform	**0.9106 ± 0.0025**	31	0.9073 ± 0.0036	26	0.9094 ± 0.0023	**0.8875 ± 0.0227**	64	0.8727 ± 0.0186	32	0.8635 ± 0.004	0.884
Wdbc	**0.9467 ± 0.0063**	12	0.931 ± 0.0104	7	0.9143 ± 0.0034	**0.9408 ± 0.0136**	14	0.914 ± 0.0298	10	0.9277 ± 0.0095	0.9268
Wine	**0.9862 ± 0.0054**	1	0.9637 ± 0.0245	1	0.9639 ± 0.0245	**0.9831 ± 0.0205**	7	0.9678 ± 0.0936	5	0.9829 ± 0.0137	0.9817
Wpbc	**0.7371 ± 0.0092**	15	0.7363 ± 0.0115	12	0.7258 ± 0.0039	**0.6894 ± 0.1327**	1	0.6894 ± 0.1327	1	0.6727 ± 0.0408	0.6633
Average	**0.8832 ± 0.0091**	15.2	0.8740 ± 0.0151	10.5	0.8637 ± 0.0142	**0.8823 ± 0.0194**	14.7	0.8691 ± 0.0243	8.9	0.8694 ± 0.0124	0.8516

with C4.5 learner, respectively. When the SVM classifier model is applied, the MPOEC outperforms respectively two standard methods in 18 datasets and 14 datasets. Compared with the single classifier, our method can improve the performance greatly. Furthermore, the average of results of all datasets is shown on the bottom of each table, which illustrates obviously that the proposed method improves accuracy from 1.73% to 3.16% compared with standard methods and single classifier. As a summary, the results indicate that MPOEC method have the advantage of the standard methods by greedy optimization selective ensemble classifiers.

From the aspect of NSC, intuitively, the number of selected classifiers for most datasets is less than 50% of entire number of ensemble classifiers, and even no more than 15%. It indicates that fewer classifiers are selected by the proposed method and are combined obtain better performance than the whole classifiers ensemble. Because of the character of sparse solution of MPOEC, only less than 10 classifiers are selected to ensemble for some datasets in 100 classifiers ensemble, such as chart, glass, iris and shuttle. In a word, the results indicate that the proposed methods can improve the performance of ensemble classifiers, and decrease the complexity of ensemble by gaining a sparse combination of classifiers.

For the sake of expressing the results of the proposed method better than tables, we illustrate the results in Fig. 5. The figures represent the testing accuracy of the standard method and ours for each point. The horizontal-axis is accuracy of the proposed method, and vertical-axis is accuracy of standard ensemble method. Evidently, points below the diagonal line, which are more than ones above the diagonal line, show a better performance of MPOEC method, and points above the diagonal line show a better performance of standard ensemble strategies. Hence, the points below the diagonal line are more than above points, which indicates that our method obtains better performance than standard methods.

Based on the theoretic analysis, one important reason is that MPOEC algorithm actually can optimize the combination of ensemble classifiers, which is acquired the best similar predictions to the labels of samples. In optimization, MPOEC selects some classifiers with good performance and discrepancy to ensemble, and the selecting process of ensemble classifiers is shown in Fig. 6 for the sake of explanation about its advantage. Furthermore, in the procedure of combining classifiers, many classifiers obtain zero coefficients, so the MPOEC method is also considered to be a selective ensemble classifiers method.

The process of selective classifiers for Breast dataset is shown in Fig. 6, and it is the results of ensemble 50 SVM classifiers with bagging and MPOEC. The upper chart is the accuracy of every classifier for the original training set, the middle chart is the coefficients gained by MPOEC, and the lower chart is the accuracy of the testing set by individual classifiers. In this ensemble, the accuracy of ensemble classifiers is 75.13% by MPOEC method, and the accuracy is 72.59% by standard bagging strategy. In the upper and lower charts, the points with 'blue circle' indicates that these classifiers gained the positive coefficients, the points with 'red triangle' indicates that these classifiers gained the negative coefficients, and the rest points are the classifiers gained zero coefficients in ensemble. From the results of the three charts, only 7 classifiers are selected by MPOEC approach from 50 classifiers to ensemble, and they are the 4th, 5th, 6th, 21st, 27th, 29th and 46th classifiers, respectively. From the middle chart, the 21st and 27th classifiers gained greater coefficients than others, because they obtain better performance for the training set than others. Hence, the effect of good classifiers (21st and 27th classifiers) for ensemble performance is boosted up by given higher coefficients. According to the form (26) (when the classifiers of zero coefficients are eliminated, the form (26) is equal to the form (13)), the result of ensemble is equal to the combination of all classifiers with their coefficient α_l, so the selected classifiers are helpful for increasing the accuracy of the ensemble system. In the lower chart, the purple dashed line denotes that the accuracy for the testing set is 73%, and it is easily found that there are more than 25 classifiers (50%) which are lower than 73%. If the classifiers are combined by the general

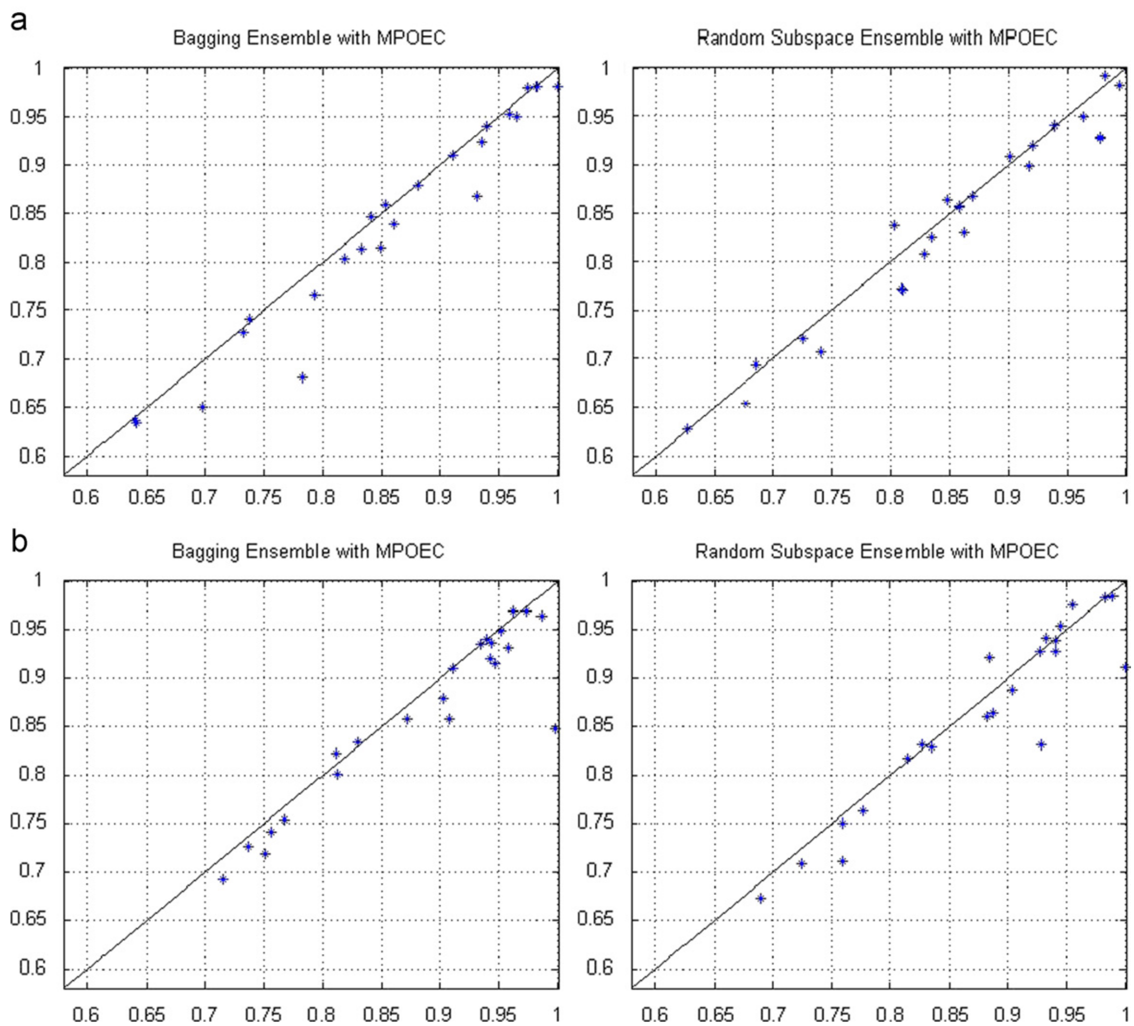

Fig. 5. Accuracy of testing samples compared the proposed method with standard ensemble method. (a) C4.5 as the basic classifier and (b) SVM as the basic classifier.

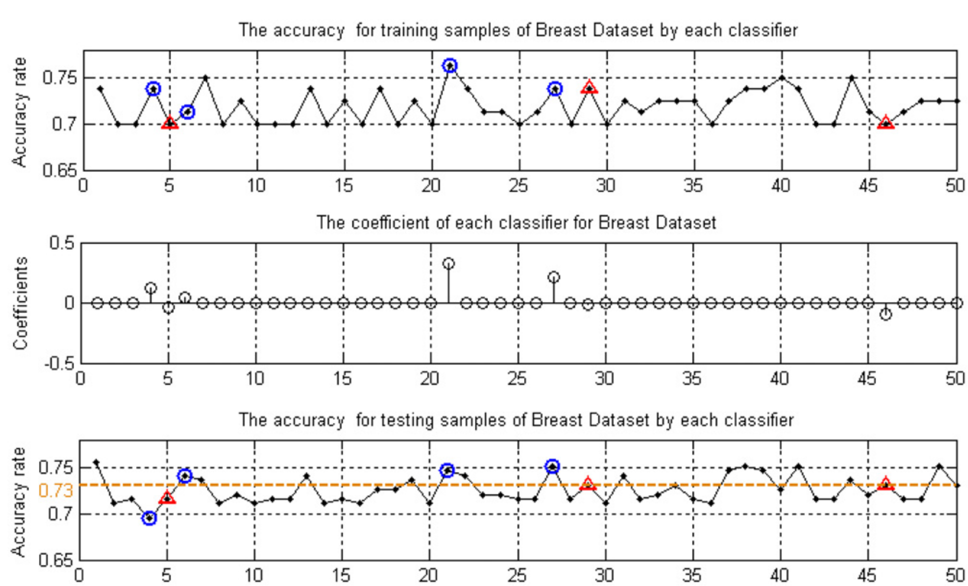

Fig. 6. Diagram of accuracy of the training set and the testing set by each classifier and the coefficients of ensemble classifiers for Breast Dataset.

voting rule, the performance of ensemble must be lower than 73%. But MPOEC method eliminates some poor classifiers by giving zero coefficients α_l to them. Especially, some similar performance classifiers are erased based on the analysis of diversity in former section of our paper, and a few poor classifiers are still reserved, which also satisfies the need of diversity of an ensemble system, for instance, two classifiers that are lower than 73% are selected in Fig. 6. Hence, MPOEC approach improves the performance compared with the standard ensemble strategies

$$f_{opt} = \sum_{l=1}^{L} \alpha_l f_l, \quad l = 1,\ldots,L \tag{26}$$

In following part, it is exhibited that the obtained coefficients of all classifiers and accuracy of classification by each classifier in ensemble system for some UCI Datasets are demonstrated in Fig. 7. The first and third rows show the accuracy of classifiers for datasets, the second and forth rows illustrate the gained optimal coefficient of each classifier. The results are gained by MPOEC ensemble and standard bagging with 100 classifiers (C4.5). The classifier with good performance is given a higher coefficient, especially, when some classifiers have similar performances, they are eliminated by zero coefficients, which not only increases the diversity but also reduces the complexity of combining classifiers. However, according to the results of the pictures, it is found that

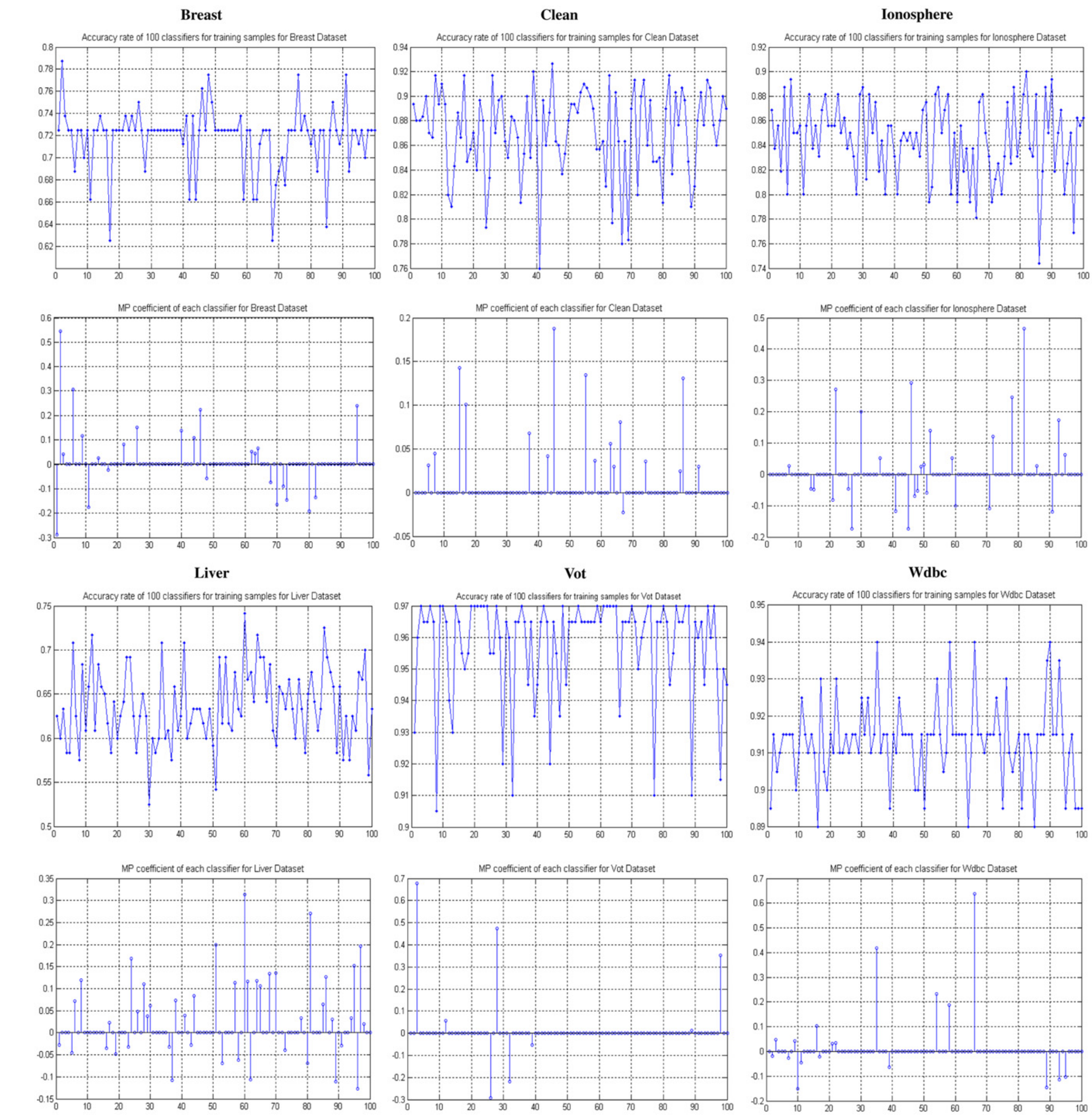

Fig. 7. Accuracy of each classifier and its coefficient in MPOEC. Note that accuracy for six datasets are given as follow, respectively: ((Dataset) MPOEC and standard method) (Breast) 75.63% and 72.59%, (Clean) 80.11% and 78.98%, (Ionosphere) 92.72% and 90.07%, (Liver) 65.78% and 64.89%, (Vot) 93.19% and 92.34%, (Wdbc) 93.77% and 91.60%.

some poor classifiers are also retained in ensemble system, and this indicates that the diversity among ensemble classifiers can be increased not be reduced by MPOEC. In conclusion, the accuracy and coefficient diagrams indicate that a majority of similar and harmful classifiers can be weakened by giving lower or zero coefficients to them in the MPOEC algorithm, and the ensemble diversity is increased by selecting some available classifiers, consequently, our method can improve the performance of ensemble classifiers.

However, in experiments, we discover that the MPOEC approach was poor for some datasets compared with standard ensemble methods. According to the analysis of experiments, there is an important reason that the over-fitting may occur in the process of searching the basis function of the MPOEC algorithm, which may lead to the better performance made by standard ensemble strategies than MPOEC algorithm. In MPOEC approach, the basis functions are the predictions obtained by classifier algorithms for the training set, and the target function of optimization is the true label of training samples, so it is indicated that the coefficients of individual classifiers is given based on original training set. If a selected classifier has a good performance for original training set but poor for the testing set, this classifier will damage the performance of ensemble for testing set. Hence, the over-fitting will be produced in the MPOEC approach. As follows, the over-fitting problem is shown in a picture of the 50 classifiers ensemble by MPOEC algorithm for Heart Dataset in Fig. 8.

In Fig. 8, the 'blue circle' denotes classifiers gained the positive coefficient, and the 'red triangle' indicates classifiers gained the negative coefficient. Three charts show the accuracy of the training set, the coefficients and testing set of 50 classifiers for Heart Dataset. In this ensemble, MPOEC method gains 81.55% accuracy and bagging strategy gains 83.49%. From the upper and middle charts, the classifiers, like the 6th, 18th, 24th, 30th and 39th classifiers, have good classification performances for training set and gain the higher coefficients than others. But these classifiers have poor performance for testing samples in the lower chart. It indicates that the MPOEC approach will not obtain a good performance, comparing with the general ensemble methods, because high coefficients are given to the poor classifiers for the testing set by MPOEC algorithm, which are selected to ensemble.

Based on the experimental results, it is also found that the accuracy obtained by single classifier algorithms is higher than standard ensemble and MPOEC methods for several datasets in Tables 3 and 4. According to the analysis of experiment, two influential factors could lead to this problem in experiments. Firstly, the parameter of classifier may affect the accuracy. In our experiments, the parameter of every individual is given based on the original training set, which is the same as the parameter of single classifier, in

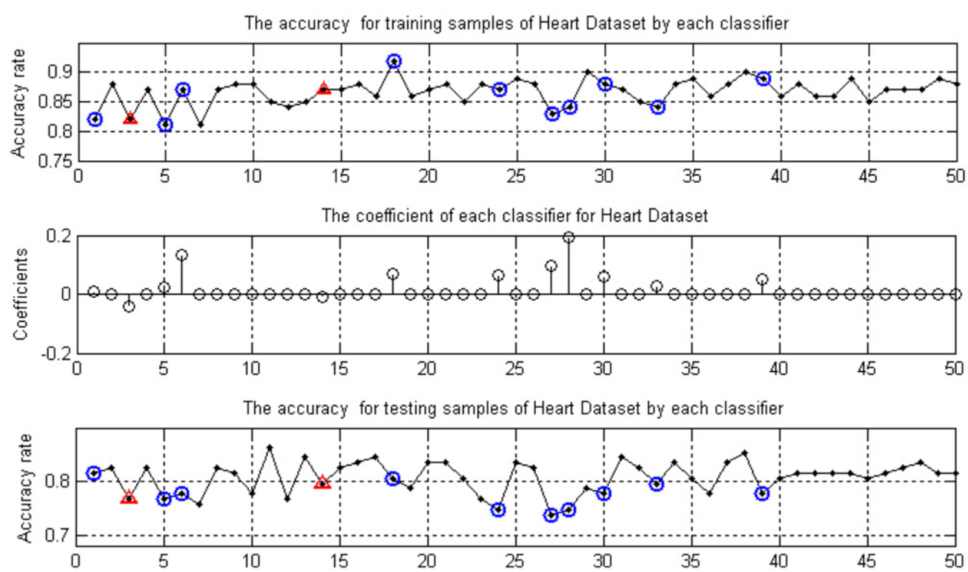

Fig. 8. Diagrams of accuracy of the training set and the testing set by each classifier and the coefficients of ensemble classifiers for Heart Dataset.

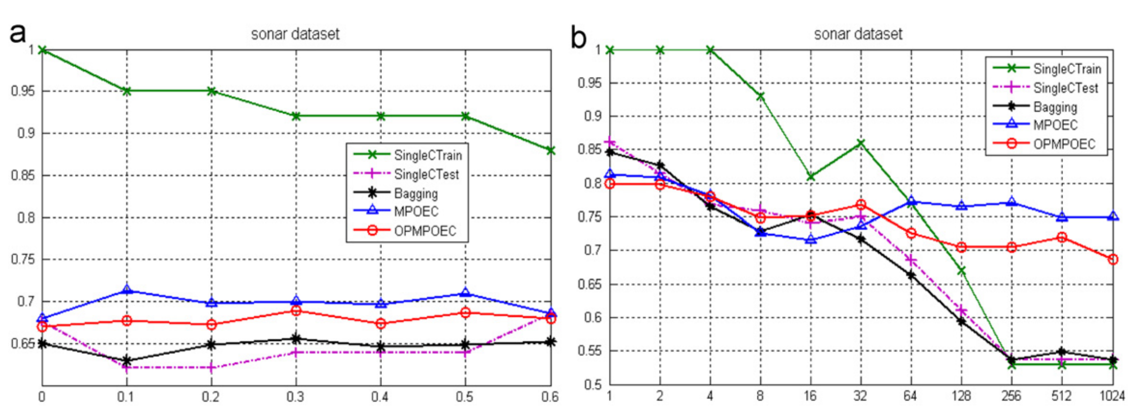

Fig. 9. The results of changing parameters for sonar dataset: (a) Bagging and MPOEC with C4.5 and (b) Bagging and MPOEC with SVM.

order to reduce the complexity of producing individual classifiers. For instance, for C4.5 decision tree, all parameters are 20% as the percentage of incorrectly assigned samples at a node. For SVM, the parameter of every classifier is the one by which single classifier can obtain better or best performance. The results of various parameters are shown in Fig. 9. Secondly, it may be relation to the number of

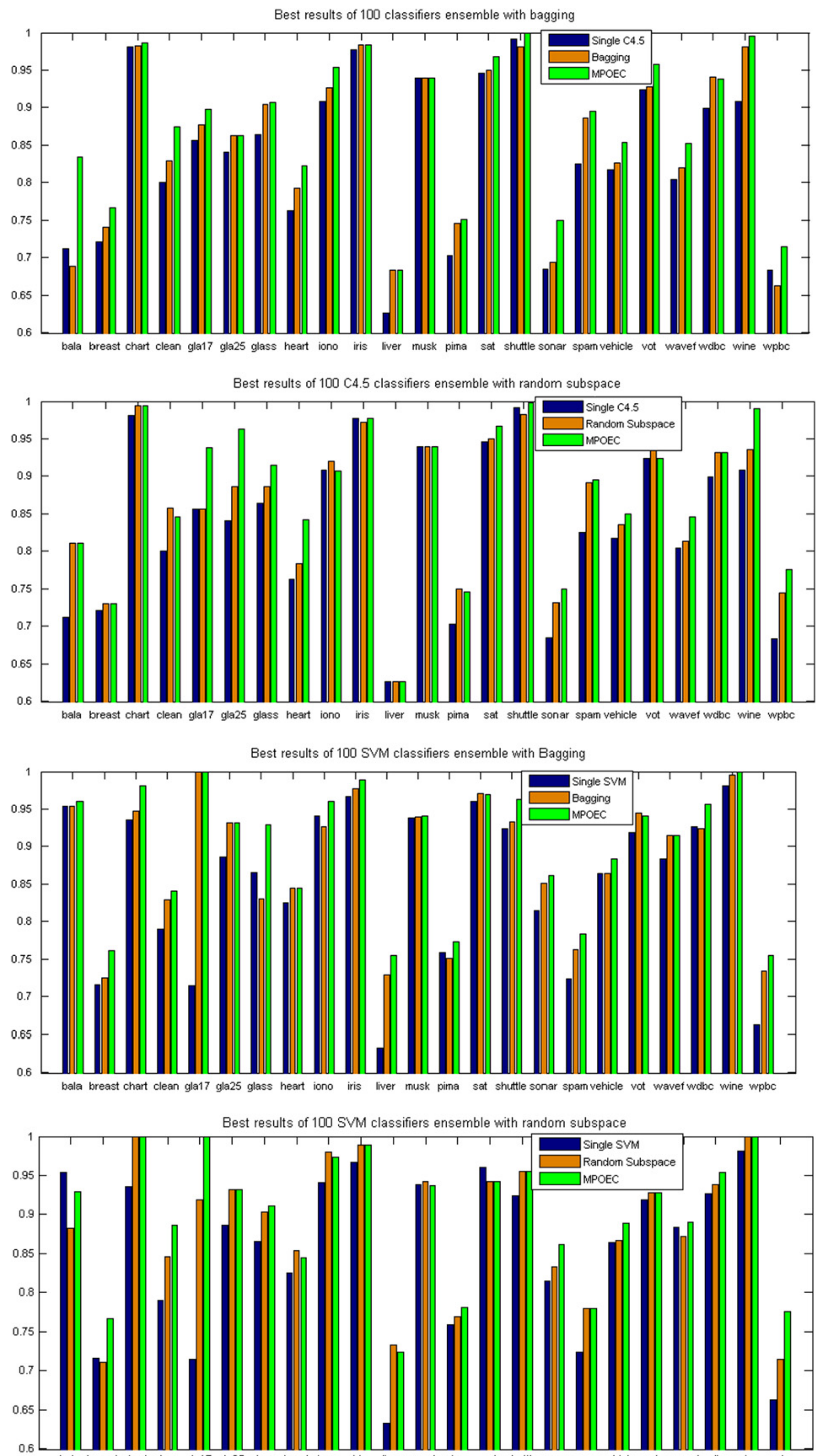

Fig. 10. Best accuracy of three methods.

Table 5
Results of our method compared with previous ensemble methods using the same datasets.

Datasets	MPOEC				Ref.						
	Bagging		RSM		[1]	[2]	[3]	[4]	[5]	[6]	[7]
	C4.5	SVM	C4.5	SVM							
Balance	0.7835	0.9521	0.8104	0.9289	–	–	0.9098	0.9487	0.9033	–	0.9056
Breast	0.7331	0.7509	0.7252	0.7599	–	–	0.6821	0.7352	0.7266	–	0.7374
Chart	0.9821	0.9434	0.8523	1	–	–	–	–	–	–	–
Clean	0.8186	0.8119	0.8028	0.8267	–	–	–	–	–	–	–
Glass17	0.8531	0.9971	0.858	0.9996	–	–	–	–	–	–	–
Glass25	0.8414	0.9082	0.8623	0.8841	–	–	–	–		–	–
Glass	0.86	0.9026	0.8487	0.9033	0.686	0.651	0.7234	0.7736	0.7427	–	–
Heart_c	0.7938	0.8299	0.8098	0.835	–	0.809	0.7751	0.875	0.8225	–	0.8288
Ionosphere	0.9319	0.9433	0.9017	0.955	–	–	0.916	0.9149	0.9388	–	0.9315
Iris	0.9746	0.9741	0.9778	0.9889	–	–	–	–	0.9573	–	0.9533
Liver	0.6408	0.7154	0.6267	0.7243	–	0.626	–	0.7	–	–	–
Musk	0.9397	0.9401	0.9397	0.9322	–	–	–	–	–	–	–
Pima	0.7379	0.7563	0.7413	0.7768	–	0.745	0.7552	0.7625	0.7648	–	0.7811
Sat	0.9655	0.9624	0.9634	0.9405	–	–	–	–	–	–	–
Shuttle	0.9992	0.9579	0.994	0.9453	–	–	–	–	–	–	–
Sonar	0.698	0.8111	0.6683	0.8141	0.8423	0.753	0.8221	0.8221	0.8356	–	–
Spambase	0.8816	0.7679	0.8697	0.76	–	–	–	–	–	–	–
Vehicle	0.833	0.8722	0.8349	0.8825	–	–	0.7542	0.8123	0.7805	0.811	–
Vot	0.9359	0.9351	0.9212	0.9277	–	–	0.9548	0.9333	0.9653	–	0.9654
Waveform	0.8497	0.9106	0.8291	0.8875	0.85	–	0.9092	0.8616	0.8393	0.86	0.8668
Wdbc	0.9114	0.9467	0.9175	0.9408	–	–	–	–	–	–	–
Wine	0.9591	0.9862	0.9782	0.9831	–	–	–	–	–	–	0.9944
Wpbc	0.6404	0.7371	0.6767	0.6894	–	–	–	–	–	–	–

[1] (Zhang Xiangrong, 2005, in Ref. [37]) [2] (Ting and Zheng, 2003, in Ref. [50]) [3] (Melville and Mooney, 2005, in Ref. [51]) [4] (Nicolas Garcia-Pedrajas, 2008, in Ref. [5]) [5] (Juan and Ludmila, 2006, in Ref. [28]) [6] (Partalas et al., 2008, in Ref. [38]) [7] (Tsoumakas et al., 2005, in Ref. [39], note that its accuracy is the result of only one fold cross-validation).

training subset of individual. For bagging strategy, every subset is only 50% of original training samples instead of normal 75% or 80%. When training samples are only 50%, the individual classifier may obtain poor performance in the parameter that is used to obtain good performance for single classifier (for original training samples), so the ensemble may be inferior to single classifier.

In Fig. 9(a), it is obvious that when the parameter changed from 0% to 60%, MPOEC method (blue line) obtained better accuracy than others. Specially, bagging strategy (black line) is not better than single C4.5 (purple broken line) in all parameters. In Fig. 9(b), single SVM algorithm is better than MPOEC and standard strategies for several parameters, such as $\sigma=1$, 8. Note that the parameter σ of SVM is equal to $2^n(n=0, 1, 2, \ldots, 10)$. However, it is obvious that the proposed method can still obtain a good performance, when single SVM and standard strategies have inferior accuracy. The change of accuracy of MPOEC is not dramatic but robust while the parameters are changed. According to the results, the performance of ensemble is better than single classifier for original training samples in not all parameters, but the proposed method improve the sensitivity of parameter compared with standard ensemble strategies.

In order to illustrate visually the superiority of the proposed method, Fig. 10 shows the best performances of three methods, and they are respectively MPOEC, standard ensemble strategies (bagging and random subspace) and single classifier. In charts, x-axis denotes the UCI datasets and their names are listed at bottom of each chart. Noticeably, several datasets use partial names.y-axis denotes the best accuracy in 50 times of ensemble corresponding to the datasets. 'Blue bar' is the accuracy of the single classifier, 'Carmine bar' is the accuracy of the standard ensemble strategy and 'Green bar' is the accuracy of the MPOEC approach. Apparently, the results show that the proposed method can improve ensemble performance.

In Table 5, it is shown that the results of the proposed method compared with several previous ensemble methods and selective ensemble methods. According to the accuracy of the datasets, the performance of our method is slightly higher than other methods for the same datasets, and the number of winning is listed in the foot of the Table 5. In the comparison methods, original training sets are, respectively, 50%, 75%, 40% and 90% (10 fold cross-validations) of the whole datasets, but the training set is smaller than them in our experiments. Especially the large datasets, such as waveform, shuttle, sat and so on. The results indicate that the proposed method obtains better performance than others, when the training samples are smaller. In the light of the analysis of the algorithms, our method can update the anterior obtained coefficients of classifiers by the posterior basis function selected in n iteration, comparing with the other ensemble methods.

5.3. Kappa-error diagram of the diversity

The diversity of ensemble classifiers that is an important factor for affecting the performance of ensemble system is studied extensively. Generally speaking, it is not easy to measure the diversity among more than two classifiers [52,53]. So the kappa statistic κ measuring the diversity between the pairwise classifiers is used widely [54,28], and it is introduced to a pruning method for AdaBoost by Margineantu and Dietterich [27]. It is defined as follows.

Given two classifiers C_a and C_b, and a training set $X_{training}$ which has N samples. The κ statistic is defined in

$$\kappa = \frac{p_1 - p_2}{1 - p_2} \tag{27}$$

where, p_1 is the probability that two classifiers agree, which is computed by form (28), p_2 is the probability that two classifiers agree by chance, which is gained by form (29)

$$p_1 = \sum_i^L D_{ii} \bigg/ N \tag{28}$$

$$p_2 = \sum_{i=1}^L \left(\sum_{j=1}^L \frac{D_{ij}}{N} \cdot \sum_j^L \frac{D_{ji}}{N} \right) \tag{29}$$

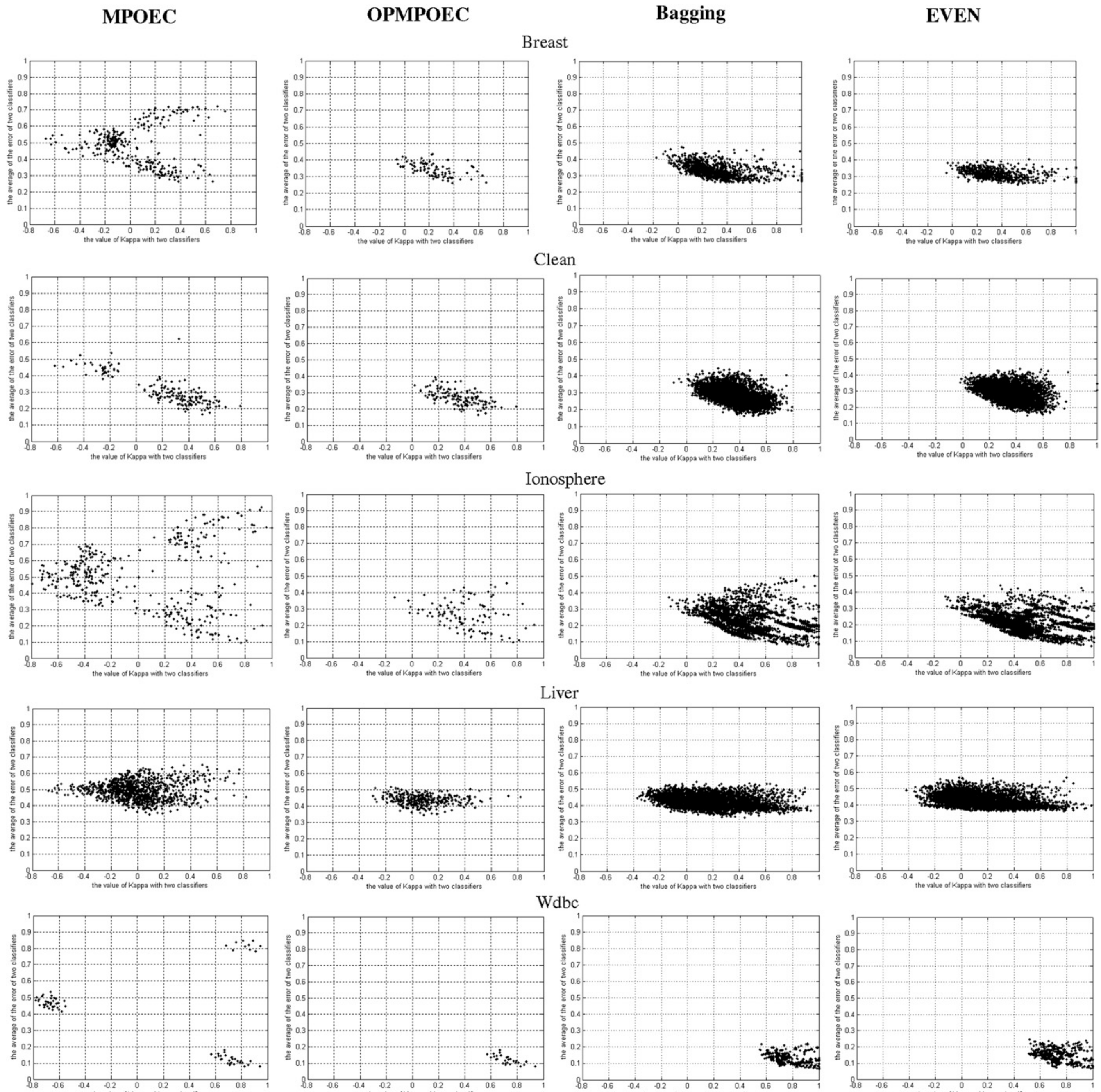

Fig. 11. Kappa-error diagrams of four ensemble methods.

Table 6
The accuracy and number of dots of five datasets for four methods.

Methods	Breast		Clean		Ionophere		Liver		Wdbc	
	Accuracy	Number	Accuracy	Number	Accuracy	Number	Accuracy	Number	Accuracy	Number
MPOEC	0.6701	300	0.8125	171	0.9139	406	0.6933	1035	0.9106	78
OPMPOEC	0.731	91	0.75	136	0.9007	105	0.6533	351	0.905	28
Bagging	0.7259	4950	0.8182	4950	0.8411	4950	0.6711	4950	0.897	4950
EVEN	0.7259	4950	0.7841	4950	0.8675	4950	0.6311	4950	0.8943	4950

where, L is the number of class of dataset, $i \in \{1, 2, \ldots, L\}$, D_{ij} is the number of training samples for which $C_a(x)=i$ and $C_b(x)=j$. The kappa-error diagram is a scatter plot where each point corresponds to a pair of classifiers [28]. In kappa-error diagram, it is shown that the diversity between two classifiers and the average error of two classifiers, the x axis is the value of κ and the y axis is the average error.

In following, the experiments are aimed at comparing the diversity of the proposed method against the standard ensemble methods and evolutionary ensemble method [55,56]. Evolutionary ensemble (EVEN) applies a framework of genetic algorithm to weight the contribution of each classifier via an appropriate fitness function with the positive class *f-measure* [57] for unbalance datasets. In experiments, 100 classifiers (C4.5) are ensemble with bagging strategy by our method, and the parameters of MPOEC are as same as the Section 5.2. For EVEN method, the number of generations is $G=1000$, the population size p is equal to 250, the number of classifiers is 100, and three parameters of genetic algorithm (crossover probability, mutation probability and survivor probability) are, respectively, 80%, 1% and 19%.

General speaking [28], the kappa value $\kappa=0$ indicates that the agreement of the two classifiers equals that expected by chance, and $\kappa=1$ indicates the two classifiers agree on every sample. Especially, negative value $\kappa<0$ indicates that agreement is weaker than expected by chance. Fig. 11 shows the kappa-error diagrams for five datasets. According to these charts, it is easily seen that the kappa κ of diagrams of standard bagging and EVEN method focus on the bottom right corner. However, the kappa κ of our method's diagrams is divided into three parts for wdbc dataset: the right corner segment, the left middle segment and the right upper segment.

The right corner segment of MPOEC is similar with OPMPOEC, because these points are obtained by classifiers with positive coefficients, and it is obvious that the values of diversity are lower than bagging and EVEN. The left middle segment is composed of the dots gained by computing diversities between classifiers with positive coefficients and classifiers with negative coefficients. In the right upper segment, the dots are obtained by classifiers with negative coefficients. Hence, from the results of diagrams of diversity, it is obviously found that the dots in right corner of our method are so smaller, and the values of dots are lower than bagging and EVEN. It indicates that the proposed method eliminates some similar and useless classifiers compared with bagging and EVEN. For instance, the value of breast dataset is lower than 0.8. The accuracy of each method and the number of dots in diagrams are shown in Table 6.

6. Conclusions

In this paper, we have presented a new method to combine classifiers in ensemble system based on the diversity between a pair of classifiers and the performance of classifiers, using a greedy algorithm to search for an optimal combination of ensemble classifiers. Because diversity and accuracy can be balances by the proposed method, a very simple strategy can be used to construct individual classifiers, which simples the process of constructing individual of ensemble. In the optimal process, MPOEC approach can select some classifiers with diversity, which makes the diversity of ensemble instead of constructing different training sets. Furthermore, the experimental results indicate MPOEC improves the performance of ensemble and increase the diversity compared with bagging and random subspace strategies. However, we still found a shortage about over-fitting problem in the proposed method. Hence, our works will focus on improving the over-fitting phenomenon in the future. One direction is to select some testing samples to add to the training samples, and the other is to use the feedback information of the testing performance of each basis function to decide the coefficient together with the training performance.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Nos. 60603019, 60602064, 61003198 and 61003199), the National High Technology Research and Development Program (863 Program) of China (Nos. 2007AA12Z223 and 2009AA12Z210), and the National Basic Research Program (973 Program) of China (No. 2006CB705707).

References

[1] L. Breiman, Bagging predictors, Machine learning 24 (2) (1996) 123–140.
[2] Tin Kam Ho, Complexity of classification problems and comparative advantages of combined classifiers, in: Multiple Classifier Systems, Lecture Notes in Computer Science, vol. 1857, Springer, 2000, pp. 97–106.
[3] Z.H. Zhou, J. Wu, W. Tang, Ensembling neural networks: many could be better than all, Artificial Intelligence 137 (1–2) (2002) 239–263.
[4] L.I. Kuncheva, Evaluation of stability of k-means cluster ensembles with respect to random initialization, IEEE Transaction on Pattern Analysis and Machine Intelligence 28 (11) (2006) 1798–1808.
[5] N. Garcia-Pedrajas, C. Fyfe, Construction of classifier ensembles by means of artificial immune systems, Joural of Heuristics 14 (3) (2008) 285–310.
[6] L.I. Kuncheva, Combining classifiers: soft computing solution, in: S.K.A. Pal (Ed.), Pattern Recognition: From Classical to Modern Approaches, World Scientific, Singapore, 2001, pp. 427–451.
[7] R.E. Schapire, The strength of weak learnability, Machine Learning 5 (2) (1990) 197–227.
[8] Y. Yasummura, N. Kitani, K. Uehara, Integration of bagging and boosting with a new reweighting technique, in: Proceedings of the 2005 International Conference on Computational Intelligence for Modelling, Control and Automation, and International Conference Intelligent Agents, Web Techologies and Internal Commerce, 2005.
[9] J. Canul-Reich, L. Shoemaker, L.O. Hall, Ensembles of fuzzy classifiers, in: IEEE International Fuzzy Systems Conference, pp. 1–6, 2007.
[10] L.K. Hansen, P. Salamon, Neural network ensembles, IEEE Transactions on Pattern Analysis and Machine Intelligence 12 (10) (1990) 993–1001.
[11] Tin Kam Ho, A Data complexity analysis of comparative advantages of decision forest constructors, Pattern Analysis and Applications 5 (2002) 102–112.
[12] J.J. Ling, Z.H. Zhou, Causal discovery based neural network ensemble method, Journal of Software 15 (10) (2004) 1479–1484.
[13] W. Tang, Z.H. Zhou, Bagging-based selective clusterer ensemble, Journal of Software 16 (4) (2005) 496–502.
[14] K. Tumer, J. Ghosh., Error correlation and error reduction in ensemble classifiers, Connection Science, Special Issue on Combining Artificial Neural Networks: Ensemble Approaches 8 (1996) 385–404.
[15] P. Melville, R.J. Mooney, Creating diversity in ensembles using artificial data, Information Fusion 6 (2004) 99–111.
[16] A. Tsymbal, M. Pechenizkiy, P. Cunningham, Diversity in search strategies for ensemble feature selection, Information Fusion 6 (2005) 83–98.
[17] Jeremy Z. Kolter, Marcus A. Maloof, Dynamic weighted majority: a new ensemble method for tracking concept drift. In: Proceedings of the Third IEEE International Conference on Data Mining, 2003, pp. 123–130.
[18] V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1999.
[19] Translated by Zhang Xue-gong. The nature of statistical learning theory. Beijing, Tsinghua University Press, 2000.
[20] Christopher J.C. Burges, A tutorial on support vector machine for pattern recognition, Data Mining and Knowledge Discovery 2 (2) (1998) 121–167.
[21] Pascal Vincent, Yoshua Bengio, Kernel matching pursuit, Machine Learning 48 (1–3) (2002) 165–187.
[22] Tao Ban, Shigeo Abe, SVM ensembles for selecting the relevant feature subsets, in: Proceedings of International Joint Conference on Neural Networks, Montreal, Canada, July 31–August 4, 2005.
[23] Dacheng Tao, Xiaoou Tang, et al., Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (7) (2006) 1088–1099.
[24] Husan-Tien Lin, Ling Li, Support vector machinery for infinite ensemble learning, Journal of Machine Learning Research 9 (2008) 285–312.
[25] Ivor W. Tsang, James T. Kwok, Ensembles of partially trained SVMs with multiplicative updates, in: The 20th International Joint Conference on Artificial Intelligence Hyderabad, India, 2007, pp. 1089–1094.
[26] L.C. Jiao, Q. Li, Kernel matching pursuit classifier ensemble, Pattern Recognition 36 (2006) 587–594.
[27] D.D. Margineantu, T.G. Dietterich, Pruning adaptive boosting, in: Proceedings of the 14th International Conference on Machine Learning, 1997, pp. 211–218.
[28] J.J. Rodriguez, L.I. Kuncheva, Rotation forest: a new classifier ensemble method, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (10) (2006) 1619–1630.
[29] Y. Freund, R.E. Schapire, Experiments with a new boosting algorithm, in: Proceedings of the 13th International Conference on Machine Learn, Morgan Kaufmann, Bari, Italy, 1996, pp. 148–156.
[30] Y. Freund, R.E. Schapire., A decision-theoretic generalization of on-line learning and an application to boosting, Computational Learning Theory, Lecture Notes in Computer Science 904 (1995) 23–37.
[31] L. Breiman, Random forest, Machine Learning 45 (1) (2001) 5–32.

[32] Kun-Hong Liu, De-Shuang Huang, Cancer classification using rotation forest, Computers in Biology and Medicine 38 (2008) 601–610.
[33] Chun-Xia Zhang, Jiang-She Zhang, RotBoost: a technique for combining rotation forest and AdaBoost, Pattern Recognition Letters 29 (2008) 1524–1536.
[34] Loris Nanni, Alessandra Lumini, Ensemble generation and feature selection for the identification of students with learning disabilities, Expert Systems with Applications 36 (2009) 3896–3900.
[35] K. Tumer, N.C. Oza, Input decimated ensembles, Pattern Analysis Application 6 (2003) 65–77.
[36] Loris Nanni, Alessandra Lumini, Input decimated ensemble based on neighborhood preserving embedding for spectrogram classification, Expert Systems with Applications 36 (2009) 11257–11261.
[37] X. Zhang, S. Wang, T. Shan, L.C. Jiao, Selective SVMs ensemble driven by immune clonal algorithm, in: Rothlauf, F. (Ed.) Proceedings of the EvoWorkshops, Springer, Berlin, 2005, pp. 325–333.
[38] Ioannis Partalas, Grigorios Tsoumakas, Ioannis Vlahavas, Focused ensemble selection: a diversity-based method for greedy ensemble selection, in: Proceedings of the 18th International Conference on Artificial Intelligence, 2008, pp. 117–121.
[39] G. Tsoumakas, L. Angelis, I. Vlahavas, Selective fusion of heterogeneous classifiers, Intelligent Data Analysis 9 (6) (2005) 511–525.
[40] G. Martinez-Munoz, A. Suarez, Pruning in ordered bagging ensembles, in: Proceedings of the 23rd International Conference in Machine Learning, ACM Press, 2006, pp. 609–616.
[41] D.W. Opitz, J.W. Shavlik, Actively searching for an effective neural network ensemble, Connection Science 8 (1996) 337–353.
[42] M.P. Perrone, L.N. Cooper, When networks disagree: ensemble method for neural networks, in: R.J. Mam-mone (Ed.), Artificial Neural Networks for Speech and Vision, Chapman & Hall, New York, 1993, pp. 126–142.
[43] Giorgio Giacinto, Fabio Roli, Design of effective neural network ensembles for image classification purposes, Image and Vision Computing 19 (2001) 699–707.
[44] Shaoning Pang, Daijin Kim, Sung Yang Bang, Membership authentication in the dynamic group by face classification using SVM ensemble, Pattern Recognition Letters 24 (2003) 215–255.
[45] Ye Li, Yun Ze Cai, et al., Fault diagnosis based on support vector machine ensemble, in: Proceedings of 2005 International Conference on Machine Learning and Cybernetics, vol. 6, 2005, pp. 3309–3314.
[46] S.G. Mallat, Z.F. Zhang, Matching pursuit with time-frequency dictionaries, IEEE Transaction on Signal Processing 41 (12) (1993) 3397–3415.
[47] Jianxin Wu, Z.H. Zhou, et al., A selective constructing approach to neural network ensemble, Journal of Computer Research and Development (Chinese) 37 (9) (2000) 1039–1044.
[48] S. Hettich, C. Blake, C. Merz, UCI repository of machine learning databases. 〈http://www.ics.uci.edu/mlearn /MLRepository.html〉, 1998.
[49] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, 1993.
[50] K.M. Ting, Z. Zheng, A study of Adaboost with naïve Bayesian classifiers: weakness and improvement, Computational Intelligence 19 (2) (2003) 186–200.
[51] P. Melville, R.J. Mooney, Creating diversity in ensembles using artificial data, Information Fusion 6 (2005) 99–111.
[52] L.I. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms, John Wiley and Sons, 2004.
[53] L.I. Kuncheva, C.J. Whitaker, Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy, Machine Learning 51 (2) (2003) 181–207.
[54] A. Agresti, Categorical Data Analysis, John Wiley and Sons, Inc., 1990.
[55] J. Sylvester, N.V. Chawla, Evolutionary ensemble creation and thinning, in: International Joint Conference on Neural Networks, 2006, pp. 5148–5155.
[56] N.V. Chawla, J. Sylvester, Exploiting diversity in ensembles: improving the performance on unbalanced datasets. multiple classifier systems, Lecture Notes in Computer Science 4472 (2007) 397–406.
[57] C.J. van Rijsbergen, Information Retrieval, 2nd edition, Butterworths, London, 1979.
[58] T.K. Ho, The random subspace method for constructing decision forests, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (8) (1998) 832–844.

Shasha Mao was born in Shaanxi, China, on August 1, 1985. She received the B.S. degree from Xidian University, Xi'an, China, in 2006. Since 2006, she has been working toward the M.S. and Ph.D. degrees in circuit and system at Xidian University. Her research interests include machine learning, pattern recognition and data mining.

Licheng Jiao (M'87–SM'91) was born in Shaanxi, China, on October 15, 1959. He received the B.S. degree from Shanghai Jiaotong University, China, in 1982 and the M.S. and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 1984 and 1990, respectively. From 1984 to 1986, he was an Assistant Professor with the Civil Aviation Institute of China, Tianjing, China. During 1990 and 1991, he was a Postdoctoral Fellow with the Key Lab for Radar Signal Processing, Xidian University, Xi'an, China. He is currently Professor and Dean of the School of Electronic Engineering in Xidian University. His current research interests include signal and image processing, nonlinear circuits and systems theory, learning theory and algorithms, optimization problems, wavelet theory, machine learning and data mining. He is the author of four books: Theory of Neural Network Systems, Theory and Application on Nonlinear Transformation Functions, Neural Computing and Applications and Implementations of Neural Networks (1990, 1992, 1993 and 1996, respectively, Xidian University Press: Xi'an). He is the author or coauthor of more than 150 scientific papers.

Lin Xiong was born in Shaanxi, China, on July 15, 1981. He received the B.S. degree from Shaanxi University of Science & Technology, Xi'an, China, in 2003. He worked for Foxconn Co. after graduation from 2003 to 2005. Since 2006, he has been working toward the M.S. and Ph.D. degrees in pattern recognition and intelligent information at Xidian University. His research interests include pattern recognition, and machine learning, especially active learning and sparse learning.

Shuiping Gou was born in Shannxi, China, on December 16, 1978. She received the B.S. degree and Master degree from Xidian Yniversity in 1999 and 2002. In 2008, she received the Ph.D. degree in Pattern Recognition and Intelligent System from Xidian University, Xi'an, China. She joined the faculty of the Institute of Intelligent Information Processing, Xidian University in 2005, where she is currently an associate professor. Her research domains are machine learning, data mining and SAR image processing.