Front. Electr. Electron. Eng. China 2010, 5(2): 173–178
DOI 10.1007/s11460-010-0004-3

RESEARCH ARTICLE

**Deqiang HAN, Chongzhao HAN, Yi YANG**

# Modified center-based feature line classification approach



**Abstract** A novel classification approach called modified center-based feature line (MCFL) is proposed to reduce the computational cost of the nearest feature line (NFL) and maintain the advantages of NFL. Unlike NFL, MCFL defines a different type of feature line and utilizes both the query point's local information and corresponding class-global information in training set. In experiments provided, the comparisons with the nearest neighbor (NN), NFL, and other NFL-refined approaches show that the computation time of MCFL can be shortened dramatically with less accuracy decreases. MCFL proposed is probably a better choice for the classification application tasks of large-scale dataset.

**Keywords** classification, nearest feature line (NFL), nearest neighbor line (NNL), center-based nearest neighbor (CNN), modified center-based feature line (MCFL)

## 1 Introduction

Classification in machine learning and pattern recognition aims to assign the correct class label to the query sample [1]. Various types of classifiers have emerged. Nearest neighbor (NN) classifier is a nonparametric and instance-based approach, which does not need priori knowledge, such as priori probabilities and conditional probabilities. It has been proved that NN has asymptotic error rate that is at most twice the Bayes error rate [2]. However, NN has its own drawbacks. In NN-based classification, the representational capacity of a database and the error rate depend on how the training samples are chosen to account for possible variations and also how many training samples are available [3]. Then, the nearest feature line (NFL) [3] was proposed to generalize the representational capacity of training sample set. NFL uses the straight lines passing through each pair of the samples from the same class. It has shown wonderful performance in many applications, including face recognition [3], audio retrieval [4], speaker identification [5], image classification [6], etc. Though it is successful in improving the classification ability, NFL still has some drawbacks, which can be concluded into two major aspects:

1) NFL encounters high computational complexity when training set is large scale.

2) NFL may fail due to extrapolation inaccuracy and interpolation inaccuracy.

Some researchers proposed refined NFL algorithms to counteract the drawbacks. Center-based nearest neighbor (CNN) [7] was proposed to lessen the computation burden of the original NFL method by defining another kind of line called center-based line (CL), which connects a training sample and the center of the sample's corresponding class. CNN can achieve competitive performance at the same time. Reference [8] indicated one of the inaccuracies referred above—extrapolation inaccuracy, and proposed a solution called nearest neighbor line (NNL). It uses the query point's two nearest training samples in each class to constitute neighbor line (NL). NNL classification is based on the perpendicular distance from query point to corresponding NL. NNL can also lower the computation cost and achieve competitive performance. There are also several other refined NFL methods, such as tunable nearest neighbor (TNN) [9], nearest feature midpoint (NFM) [10], rectified nearest feature line segment (RNFLS) [11], etc. They are all rational and effective classification approaches which can more or less suppress the drawbacks of NFL.

A new alternative NFL-refined classifier called modified center-based feature line (MCFL) is proposed in this paper to reduce the computational cost and maintain the classification accuracy.

1) To reduce the computational cost, MCFL uses the straight line called modified feature line (MFL) instead of the feature line (FL) in NFL. Each class has a corresponding MFL passing through the class' center and the query sample's NN in the corresponding class. That is, for MCFL, only one MFL is required in each class, so the

Received December 31, 2009; accepted January 29, 2010

Deqiang HAN (✉), Chongzhao HAN, Yi YANG
Institute of Integrated Automation, Xi'an Jiaotong University, Xi'an 710049, China
E-mail: deqhan@gmail.com

computational cost of MCFL is reduced significantly when compared with that of the original NFL.

2) To maintain the classification accuracy, MCFL classification decision is based on the area of the hyper-triangle spanned by query sample, class center, and query sample's NN in corresponding class. MCFL also has relatively competitive classification performance.

Comparisons between MCFL and other classification approaches, such as NN, NNL, NFL, and CNN, are provided in the experiments based on some datasets from University of California, Irvine (UCI), and Olivetti Research Laboratory (ORL) face database of Cambridge. Experimental results verify that MCFL is a simple yet effective classifier, and its major advantage is low computational cost without significant loss of classification accuracies, which is crucial for classification tasks of large-scale dataset.

## 2 NFL classifier

NFL is proposed to overcome the drawbacks of NN classifier, i.e., the performance of NN is limited by the available training samples in each class. NFL algorithm assumes that at least two samples are available for each class, which is usually satisfied, and attempts to generalize the representational capacity of available training samples. NFL approach is briefly reviewed as follows [3].

As illustrated in Fig. 1, $\boldsymbol{x}_i^\theta, \boldsymbol{x}_j^\theta$ are two training samples in the same class $\theta$, $\theta = 1,2,...,M$ ($M$ is the number of the class). Each training sample is according to an $n$-dimensional feature vector. $\overline{\boldsymbol{x}_i^\theta \boldsymbol{x}_j^\theta}$ is an FL of class $\theta$, which is a straight line passing through $\boldsymbol{x}_i^\theta$ and $\boldsymbol{x}_j^\theta$. $\boldsymbol{x}_\mathrm{q}$ is a query point, and $\boldsymbol{x}_{ij}^\theta$ is the projection point of $\boldsymbol{x}_\mathrm{q}$ on FL $\overline{\boldsymbol{x}_i^\theta \boldsymbol{x}_j^\theta}$. $\boldsymbol{x}_{ij}^\theta$ can be calculated as follows:

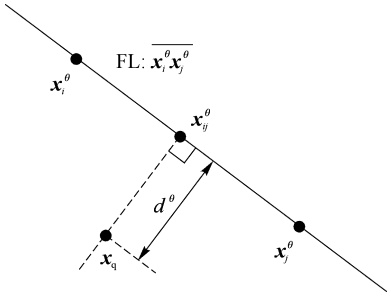


**Fig. 1** Feature line, query point, and distance between them

$$\boldsymbol{x}_{ij}^\theta = (1-\mu)\boldsymbol{x}_i^\theta + \mu \boldsymbol{x}_j^\theta, \tag{1}$$

where $\boldsymbol{x}_\mathrm{q}$, $\boldsymbol{x}_i^\theta$, and $\boldsymbol{x}_j^\theta$ are all column vectors; the position parameter

$$\mu = \frac{(\boldsymbol{x}_\mathrm{q} - \boldsymbol{x}_i^\theta)^\mathrm{T}(\boldsymbol{x}_i^\theta - \boldsymbol{x}_j^\theta)}{(\boldsymbol{x}_i^\theta - \boldsymbol{x}_j^\theta)^\mathrm{T}(\boldsymbol{x}_i^\theta - \boldsymbol{x}_j^\theta)}. \tag{2}$$

The distance from query point $\boldsymbol{x}_\mathrm{q}$ to $\overline{\boldsymbol{x}_i^\theta \boldsymbol{x}_j^\theta}$ can be derived based on the following equation:

$$d^\theta(\boldsymbol{x}_\mathrm{q}, \overline{\boldsymbol{x}_i^\theta \boldsymbol{x}_j^\theta}) = \left\| \boldsymbol{x}_\mathrm{q} - \boldsymbol{x}_{ij}^\theta \right\|, \tag{3}$$

where $\| \cdot \|$ denotes Euclidean norm.

Based on

$$\theta^* = \arg\min_{1 \leqslant \theta \leqslant M} d^\theta, \tag{4}$$

class label $\theta^*$ can be obtained and assigned to query point. $d_\mathrm{NFL} = d^{\theta^*}$ is called NFL distance.

NFL uses line features instead of point features for line features are always representational than point features. For a given training sample set, quantity of line features based on NFL is larger than point features. Thus, the representational capacity of training samples can be generalized.

Though NFL is successful in improving the classification performance, some drawbacks in NFL will limit its further application in practice. Generalized representational capacity of training samples is a double-edged sword. As referred above, the drawbacks mainly include the following:

1) High computational cost

Computational complexity is high, especially when we encounter the large-scale training dataset. For $N_\theta$ training samples belonging to class $\theta$, there will be $N_\theta(N_\theta - 1)/2$ feature lines. Suppose that each sample corresponds to an $n$-dimensional feature vector, for original NFL, there will be $(3n+1)N_\theta(N_\theta - 1)/2$ multiplication operations for each class.

2) Inaccuracy problems [11]

NFL may fail when query point is far away from training samples in NFL. This is called extrapolation inaccuracy, as illustrated in Fig. 2. In Fig. 2, $\boldsymbol{x}_i^\mathrm{c}$ represents the sample belonging to class "Circle", and $\boldsymbol{x}_i^\mathrm{s}$ represents the sample belonging to class "Star".

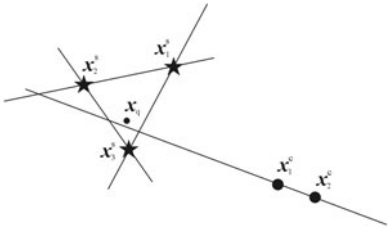

**Fig. 2** Extrapolation inaccuracy

Query sample $\boldsymbol{x}_\mathrm{q}$ is surrounded by training samples belonging to class "Star", but it is classified to class

"Circle" with the criterion of NFL. This classification error is due to the extrapolating part of FL $\overline{\boldsymbol{x}_1^{\mathrm{c}}\boldsymbol{x}_2^{\mathrm{c}}}$.

There is another type of inaccuracy called interpolation inaccuracy, as illustrated in Fig. 3.

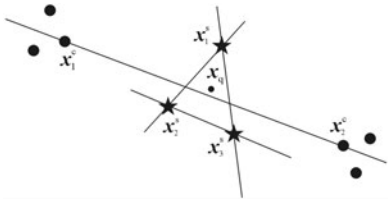

**Fig. 3** Interpolation inaccuracy

Query sample $\boldsymbol{x}_{\mathrm{q}}$ is surrounded by the samples belonging to class "Star", but $\boldsymbol{x}_{\mathrm{q}}$ is labeled "Circle". This is due to the interpolating part of FL $\overline{\boldsymbol{x}_1^{\mathrm{c}}\boldsymbol{x}_2^{\mathrm{c}}}$. The classification errors including extrapolation inaccuracy and interpolation inaccuracy are due to the FLs of one class's trespassing into of other classes' area.

There emerge several refined NFL methods to suppress the drawbacks, as referred in Sect. 1, while they also have their own drawbacks. Motivated by the idea of NNL and CNN, we propose MCFL in this paper, which can reinforce the advantages of both the two refined NFL approach (NNL and CNN) and can counteract their drawbacks at the same time. MCFL will be introduced in the following section.

## 3 MCFL

Unlike NFL, MCFL proposed uses another kind of line feature for classification motivated by NNL and CNN. The main idea of MCFL will be described as follows.

### 3.1 Implementation of MCFL

Let $\boldsymbol{x}_i^{\theta}$ be a training sample from class $\theta$, $\theta = 1,2,...,M$. $\boldsymbol{o}^{\theta}$ denotes the center of all the training samples in class $\theta$. Let $N_{\theta}$ be the number of training samples from class $\theta$. Then, $\boldsymbol{o}^{\theta}$ can be calculated by

$$\boldsymbol{o}^{\theta} = \frac{\sum_{i=1}^{N_{\theta}} \boldsymbol{x}_i^{\theta}}{N_{\theta}}. \tag{5}$$

For $\boldsymbol{x}_{\mathrm{q}}$, find its nearest training sample denoted by $\boldsymbol{x}_{\mathrm{qn}}^{\theta}$ in class $\theta$, $\theta = 1,2,...,M$. Then, we can define the MFL $\overline{\boldsymbol{o}^{\theta}\boldsymbol{x}_{\mathrm{qn}}^{\theta}}$, which is a straight line passing through $\boldsymbol{x}_{\mathrm{qn}}^{\theta}$ and $\boldsymbol{o}^{\theta}$, as illustrated in Fig. 4.

Obviously, there are totally $M$ MFLs for a query point, where $M$ denotes the number of class.

MFL is used to capture information provided by the

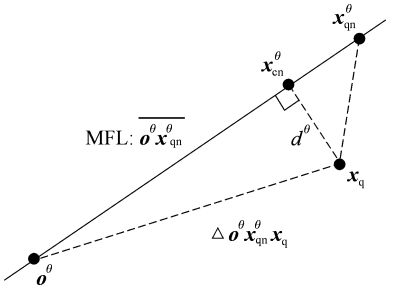


**Fig. 4** MFL and hyper-triangle in MCFL

interaction between the points $\boldsymbol{x}_{\mathrm{qn}}^{\theta}$ and $\boldsymbol{o}^{\theta}$ to achieve better classification performance. $\boldsymbol{x}_{\mathrm{qn}}^{\theta}$ implies the query point's neighborhood information in training samples, and $\boldsymbol{o}^{\theta}$ implies the global information of class $\theta$.

$\boldsymbol{x}_{\mathrm{cn}}^{\theta}$ is the projection point of $\boldsymbol{x}_{\mathrm{q}}$ on MFL $\overline{\boldsymbol{o}^{\theta}\boldsymbol{x}_{\mathrm{qn}}^{\theta}}$. $\boldsymbol{x}_{\mathrm{cn}}^{\theta}$ can be calculated based on

$$\boldsymbol{x}_{\mathrm{cn}}^{\theta} = (1-\mu)\boldsymbol{o}^{\theta} + \mu\boldsymbol{x}_{\mathrm{qn}}^{\theta}, \tag{6}$$

where the position parameter $\mu$ can be calculated as follows:

$$\mu = \frac{(\boldsymbol{x}_{\mathrm{q}} - \boldsymbol{o}^{\theta})^{\mathrm{T}}(\boldsymbol{x}_{\mathrm{qn}}^{\theta} - \boldsymbol{o}^{\theta})}{(\boldsymbol{x}_{\mathrm{qn}}^{\theta} - \boldsymbol{o}^{\theta})^{\mathrm{T}}(\boldsymbol{x}_{\mathrm{qn}}^{\theta} - \boldsymbol{o}^{\theta})}. \tag{7}$$

The distance from $\boldsymbol{x}_{\mathrm{q}}$ to MFL $\overline{\boldsymbol{o}^{\theta}\boldsymbol{x}_{\mathrm{qn}}^{\theta}}$ is calculated as follows:

$$d^{\theta}(\boldsymbol{x}_{\mathrm{q}}, \overline{\boldsymbol{o}^{\theta}\boldsymbol{x}_{\mathrm{qn}}^{\theta}}) = \left\|\boldsymbol{x}_{\mathrm{q}} - \boldsymbol{x}_{\mathrm{cn}}^{\theta}\right\|. \tag{8}$$

The area of the hyper-triangle $\triangle \boldsymbol{o}^{\theta}\boldsymbol{x}_{\mathrm{qn}}^{\theta}\boldsymbol{x}_{\mathrm{q}}$ is

$$\mathrm{Area}^{\theta} = \frac{1}{2} d^{\theta}(\boldsymbol{x}_{\mathrm{q}}, \overline{\boldsymbol{o}^{\theta}\boldsymbol{x}_{\mathrm{qn}}^{\theta}}) \left\|\boldsymbol{o}^{\theta} - \boldsymbol{x}_{\mathrm{qn}}^{\theta}\right\|. \tag{9}$$

$\boldsymbol{x}_{\mathrm{q}}$ is labeled class $\theta^*$, if

$$\theta^* = \arg\min_{1 \leqslant \theta \leqslant M} \mathrm{Area}^{\theta}.$$

### 3.2 Analysis of MCFL

Obviously, MCFL retains the ideas of original NFL, i.e., it uses a linear model of a pair of sample points within the same class instead of point feature. As introduced above, the NN used in MCFL represents query point's neighborhood (or local) information in each class in training samples. The class center represents the class-global information in training set. Both the local and global information are emphasized in MFCL. The smaller the area of hyper-triangle in Eq. (9) is, the lager the similarity degree between the query sample and the corresponding class is. However, in NNL, only query point's neighborhood in each class is considered (see details in Ref. [8]).

NNL does not consider the class-global information. Compared with CNN [7], MCFL and CNN both use the class center, i.e., they utilize the class-global information, but unlike CNN, MCFL also uses NN in each class. Then, the extrapolation inaccuracies will be not as significant as those of CNN and NFL. Based on the analysis above, MCFL proposed is relatively more rational thus competitive classification performance can be maintained.

The most important thing is that the computational cost of MCFL is relatively low. If there are $N_\theta$ samples belonging to class $\theta$ in training set, for each class, there are $N_\theta(N_\theta-1)/2$ FLs in NFL, $N_\theta$ CLs in CNN, but only one MFL in MCFL. That is, the computational cost can be reduced by redefining the feature line in MCFL, which is called MFL. In addition, the calculation of class centers can be implemented offline. For the classification task of large-scale dataset, the computational cost of the classifier used should be more crucial.

## 4 Experiments

In the experiments, we use six datasets from UCI dataset (http://archive.ics.uci.edu/ml/) and the standard face recognition dataset of ORL. The experiments are executed on a PC with CPU of AMD Athlon™ 4000 + dual core processor and RAM of 2 G DDR II.

### 4.1 Experiments based on datasets from UCI

The datasets adopted are listed in Table 1. Because, in the experiments, we do not deal with the issue of missing values, the instances with missing values are removed.

In the experiments, the leave-one-out cross-validation approach is used, because it is considered to be the most effective way, which is more rigorous and reliable. Moreover, for the fairness of the classification procedure, attributes of the instances are standardized based on Eq. (10) before being submitted to the classifiers.

$$a_i = \frac{v_i - \min\ v_i}{\max\ v_i - \min\ v_i}, \tag{10}$$

where $v_i$ denotes the actual value of attribute $i$. Maximum and minimum operations are over all samples in training dataset. We use NNL, NN, NFL, CNN, and MCFL proposed to execute the same classification tasks. The classification accuracies are illustrated in Fig. 5.

The classification time records (in seconds) for the five classifiers respectively based on six datasets are listed in Table 2.

It can be concluded from Fig. 5 and Table 2 that the classification accuracies of MCFL proposed are still competitive, which are very approximate to those of NFL and NFL-refined methods, such as NNL and CNN, and MCFL has significantly lower computational cost,

**Table 1** Datasets from UCI

| dataset | number of classes | number of instances | number of attributes |
|---|---|---|---|
| Iris | 3 | 150 | 4 |
| Wine | 3 | 178 | 13 |
| Glass | 6 | 214 | 9 |
| Ionosphere | 2 | 351 | 34 |
| SPECTF Heart | 2 | 267 | 44 |
| Hepatitis | 2 | 80 | 19 |

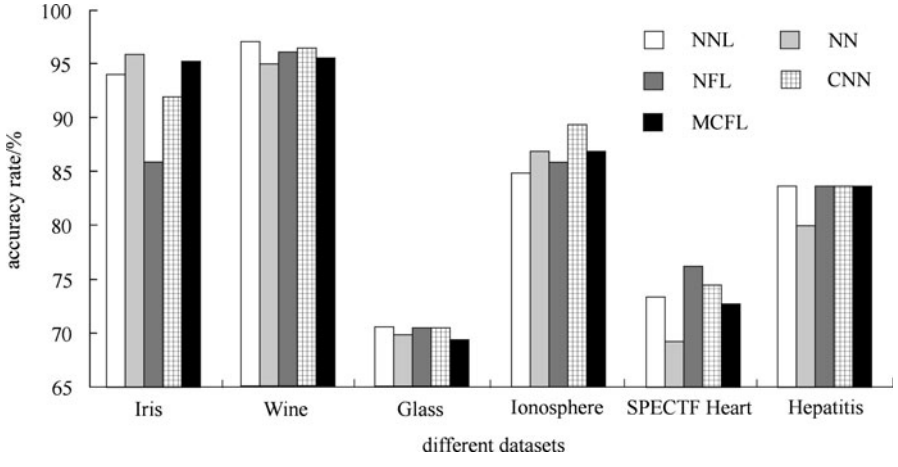


**Fig. 5** Classification accuracies in UCI

**Table 2** Computational costs comparison

| dataset | classification time/s | | | | |
|---|---|---|---|---|---|
| | NNL | NN | NFL | CNN | MCFL |
| Iris | 1.00 | 0.69 | 79.28 | 2.77 | 1.00 |
| Wine | 1.25 | 0.86 | 142.28 | 3.59 | 1.25 |
| Glass | 1.98 | 1.02 | 176.06 | 4.91 | 1.98 |
| Ionosphere | 3.50 | 2.20 | 5382.90 | 10.70 | 3.50 |
| SPECTF Heart | 2.20 | 0.40 | 3138.60 | 5.90 | 2.20 |
| Hepatitis | 0.50 | 0.30 | 31.69 | 0.72 | 0.50 |

which can be seen in Table 2. MCFL's capability to reduce computational cost is the same as that of NNL and better than that of CNN. For the applications with relatively large-scale dataset, e.g., Ionosphere and SPECTF Heart, the advantages of MCFL in computational costs are more significant when compared to NFL and other NFL-refined approaches listed in Table 2.

### 4.2 Experiments based on ORL

This experiment compares NNL, NN, NFL, and CNN with the MCFL proposed use the ORL database.

Some face images are illustrated in Fig. 6.

**Fig. 6** Face images in ORL

The ten images of each of the 40 persons are partitioned into training set and testing set. For each person, three images are selected randomly for training, while the remainders (seven images) are reserved for testing. The classification task is executed for ten times, and in each time, the training images are reselected randomly. There, a face space is constructed or spanned by a number of eigenfaces derived from a set of training face images by using principal component analysis (PCA). The average classification performances for the five classifiers over ten times' execution are listed in Table 3.

**Table 3** Average classification accuracy and average classification time

| classifiers | classification accuracy/% | classification time/s |
|---|---|---|
| NNL | 89.50 | 2.04 |
| NN | 88.21 | 0.15 |
| NFL | 90.11 | 11.41 |
| CNN | 89.93 | 2.73 |
| MCFL | 89.75 | 2.04 |

It can be seen in Table 3 that MCFL can achieve competitive classification accuracies being approximate to those of NFL and refined NFL methods used here; however, the computational time of MCFL and NNL are least when compared with NFL and NFL-refined approaches, as illustrated in Table 3, with item "classification time".

It can be concluded from the experimental results in Sects. 4.1 and 4.2 that MCFL proposed can reduce the high computational cost in NFL and at the same time maintain the classification accuracies.

## 5 Conclusions

In this paper, a novel classification approach called MCFL is proposed. It can be considered as an alternative classifier of other two kinds of NFL-refined methods (NNL and CNN) aiming to modify the original NFL. MCFL utilizes their advantages and may counteract their drawbacks at the same time. MCFL's classification decision is based on the area of the hyper-triangle, which can describe similarity between the query sample and the corresponding class by emphasizing both the local and global information. Thus, competitive classification performance can be achieved, and lower computational cost is generated. Experimental results show that MCFL is a simple yet effective classification approach. Although the classification accuracy cannot be significantly improved by MCFL, it can dramatically reduce the computational costs with less accuracy decrease. To select a proper classifier, the classification accuracy should not be the unique criterion. Trade-off should be done between classification accuracy and computational cost. In the classification task of large training sample set, the MCFL might be a better choice.

To further improve the classification performance of MCFL, we can use clustering analysis to generate subclasses. Thus, the MFLS can be constituted based on subclass centers and nearest neighbor, and better performance can be achieved without generating much more computational cost. Further research seems warranted.

**Acknowledgements** This work was supported by the State Key Development Program for Basic Research of China (No. 2007CB311006).

## References


1. Theodoridis S, Koutroumbas K. Pattern Recognition. 2nd ed. San Diego, CA: Academic Press, 2003
2. Duda R O, Hart P E, Stork D G. Pattern Classification. 2nd ed. New York: Wiley-Interscience Publication, 2000
3. Li S Z, Lu J W. Face recognition using the nearest feature line method. IEEE Transactions on Neural Networks, 1999, 10(2): 439–443
4. Li S Z. Content-based audio classification and retrieval using the nearest feature line method. IEEE Transactions on Speech and Audio Processing, 2000, 8(5): 619–625
5. Chen K, Wu T Y, Zhang H J. On the use of nearest feature line for speaker identification. Pattern Recognition Letters, 2002, 23(14): 1735–1746
6. Li S Z, Chan K L, Wang C L. Performance evaluation of the nearest feature line method in image classification and retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(11): 1335–1339
7. Gao Q B, Wang Z Z. Rapid and brief communication: center-based nearest neighbor classifier. Pattern Recognition, 2007, 40(1): 346–349
8. Zheng W M, Zhao L, Zou C R. Locally nearest neighbor classifiers for pattern classification. Pattern Recognition, 2004, 37(6): 1307–1309
9. Zhou Y L, Zhang C S, Wang J C. Tunable nearest neighbor classifier. Lecture Notes in Computer Science, 2004, 3175: 204–211
10. Zhou Z L, Kwoh C K. The pattern classification based on the nearest feature midpoints. In: Proceedings of the 17th International Conference on Pattern Recognition. 2004, 3: 446–449
11. Du H, Chen Y Q. Rectified nearest feature line segment for pattern classification. Pattern Recognition, 2007, 40(5): 1486–1497