Pattern Recognition Letters 32 (2011) 485–493

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

# A novel classifier based on shortest feature line segment

De-Qiang Han *, Chong-Zhao Han, Yi Yang

*Institute of Integrated Automation, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, PR China*

ARTICLE INFO

*Article history:*
Received 11 January 2010
Available online 12 November 2010
Communicated by F. Roli

*Keywords:*
Nearest feature line (NFL)
Trespass inaccuracy
Feature line segment
Geometric relation
Neighborhood-based classifier

ABSTRACT

A new approach called shortest feature line segment (SFLS) is proposed to implement pattern classification in this paper, which can retain the ideas and advantages of nearest feature line (NFL) and at the same time can counteract the drawbacks of NFL. The proposed SFLS uses the length of the feature line segment satisfying given geometric relation with query point instead of the perpendicular distance defined in NFL. SFLS has clear geometric–theoretic foundation and is relatively simple. Experimental results on some artificial datasets and real-world datasets are provided, together with the comparisons between SFLS and other neighborhood-based classification methods, including nearest neighbor (NN), *k*-NN, NFL and some refined NFL methods, etc. It can be concluded that SFLS is a simple yet effective classification approach.



## 1. Introduction

With the development of pattern classification (Theodoridis and Koutroumbas, 2006), various classification methods have been proposed. The nearest neighbor (NN) rule is a non-parametric and neighborhood-based classification approach which is simple, yet effective. NN has the asymptotic error rate that is at most twice the Bayes error rate (Duda et al., 2001). NN also has its own drawbacks, i.e., the representational capacity of the dataset and the error rate of classification rely on how the training samples are chosen to account for possible variations and also how many training samples are available (Li and Lu, 1999). The performance of NN also relies on the definition of the distance measure used. Lots of modified approaches have been proposed aiming to suppress the drawbacks, such as *k*-NN, surrounding neighbor (SN) (Chaudhuri, 1996), graphic neighbor (GN) (Sanchez et al., 1997) and nearest feature line (NFL) (Li and Lu, 1999; Li et al., 2000).

In the neighborhood-based classifiers referred above, NFL is a non-parametric classifier which attempts to improve the representational capacity of a sample set with limited size by using the straight lines which pass through each pair of the samples from the same class. NFL can add extra information to the original sample set and it has shown wonderful performance in many applications, such as face recognition (Li and Lu, 1999), image classification (Li et al., 2000), audio retrieval (Li, 2000), speaker identification (Chen et al., 2002), etc. Though successful in improving the classification performance, NFL still has its obvious and even fatal drawbacks. Two types of trespassing can cause extrapolation inaccuracy and interpolation inaccuracy (Zheng et al., 2004; Du and Chen, 2007). In addition, NFL has relatively high computational complexity. As referred by the author of (Du and Chen, 2007): "extra-information is a double-edged sword."

To counteract drawbacks of NFL, there emerged several rectified or modified NFL methods (Li, 2008). In (Gao and Wang, 2007), center-based nearest neighbor (CNN) method was proposed to reduce the computational cost of the original NFL method by defining another kind of line called center-based line (CL). This CL connects a training sample point and the center of the sample's corresponding class, instead of two training sample points constituting the FL in NFL method. In (Zheng et al., 2004), the authors pointed out one of the inaccuracies caused by the trespass referred above-extrapolation inaccuracy, and they proposed a solution called nearest neighbor line (NNL). NNL can also significantly reduce the computational cost. In (Zhou et al., 2004b), a tunable nearest neighbor (TNN) method was proposed to improve the performance of NFL. In (Zhou and Kwoh, 2004), nearest feature midpoint (NFM) was proposed to refine NFL by defining the distance metric as the minimum Euclidean distance between the query point and the midpoints of the two sample point constituting FL. The computational complexity of NFM is significantly less than NFL. All the refined methods cannot counteract the interpolation inaccuracy. Rectified nearest feature line segment (RNFLS) was proposed in (Du and Chen, 2007) and the authors declared that RNFLS can counteract both the two inaccuracies mentioned above. Thus it can improve the classification performance further. RNFLS also has its deficiencies. The major are that the implementation of RNFLS is relatively complex and has the procedure of pre-selection of sample subspaces (Du and Chen, 2007). There are also other types of modified NFL methods. In (He, 2006), kernel method and subspace analysis are introduced to extend nearest feature line to nonlinear nearest feature line subspace. In (Pang et al., 2007),

* Corresponding author. Tel./fax: +86 029 82668775.
*E-mail address:* deqhan@gmail.com (D.-Q. Han).

0167-8655/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2010.11.009

subspaces are constructed by nearest feature line distance of intra-class to achieve a desirable discriminating ability. In this paper, we propose a novel modified NFL classification approach called shortest feature line segment (SFLS). Our approach has clear geometric-theory foundation and retains the ideas and advantages of original NFL, i.e., it uses a linear model of pairs of sample points within the same class. Instead of calculating the distance between the query point and the feature line, SFLS attempts to find the shortest feature line segment which satisfies the given geometric relation together with the query point. SFLS has wonderful classification ability and it can suppress the extrapolation inaccuracy and interpolation inaccuracy of original NFL. The SFLS's implementation is relatively simple and it has no pre-process procedure. Experiments are provided based on some artificial datasets and real-world datasets and corresponding experimental results show that the SFLS has wonderful classification performance when compared with other neighborhood-based classification methods such as NN, *k*-NN, NFL and some refined NFL, etc. This work is based on our previous work in (Yang et al., 2009).

## 2. Brief introduction of nearest feature line approach

Nearest feature line (NFL) approach is to use the information provided by each pair of points in the same class by constituting some feature line (FL) spaces. The NFL distance is defined as the Euclidean distance from the query point to the FL, i.e., the distance between a query point and its projection onto the FL as illustrated in Fig. 1. When the sample set's size is relatively small and the sample data's feature vector has relatively high dimension (Zhou et al., 2004b), NFL approach is consistently superior to the NN methods based on conventional distance definitions.

### 2.1. Basics in NFL

Suppose that $x_i^\theta, x_j^\theta \quad (i \neq j, 1 \leq i,j \leq N_\theta)$ be two training samples belonging to the same class $\theta$, where $\theta = 1,\ldots,M$. Here $M$ represents the number of class and $N_\theta$ represents the number of samples belonging to class $\theta$. The dimensionality of the training sample's feature vector is denoted by $n$. Straight-line $\overline{x_i^\theta x_j^\theta}$ passing through $x_i^\theta$ and $x_j^\theta$ is named a feature line (FL) of the class $\theta$. Let $x_q$ be a query sample point and $x_{ij}^\theta$ be the projection point of the $x_q$ on the FL $\overline{x_i^\theta x_j^\theta}$. $\mathbf{x}_{ij}^\theta$ can be calculated based on (1) and (2). Here suppose that $x_q$, $x_i^\theta$ and $x_j^\theta$ are all column vectors.

$$x_{ij}^\theta = (1-\mu)x_i^\theta + \mu x_j^\theta, \tag{1}$$

where

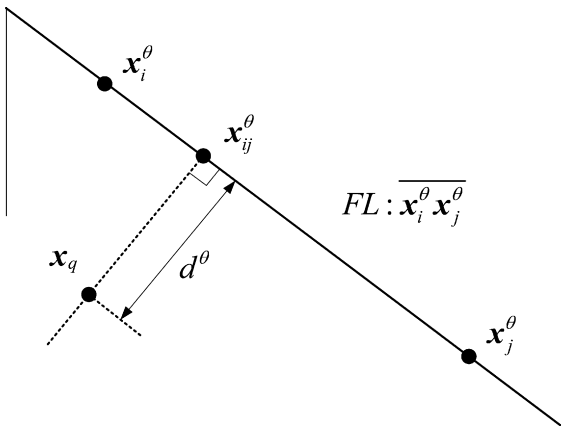


**Fig. 1.** Query point, feature line and the distance between them.

$$\mu = \frac{(x_q - x_i^\theta)^T(x_j^\theta - x_i^\theta)}{(x_j^\theta - x_i^\theta)^T(x_j^\theta - x_i^\theta)}. \tag{2}$$

The distance from the query point $\mathbf{x}_q$ to the FL $\overline{x_i^\theta x_j^\theta}$ can be calculated as follows:

$$d^\theta(x_q, \overline{x_i^\theta x_j^\theta}) = \|x_q - x_{ij}^\theta\|, \tag{3}$$

where $\|\cdot\|$ denotes some norm, e.g. Euclidean norm.

The classification decision can be made according to (4):

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

where $\theta^*$ is the class label assigned to query point $\mathbf{x}_q$. The NFL distance is defined in (5):

$$d_{\mathrm{NFL}} = d^{\theta^*}. \tag{5}$$

One of the advantages of the NFL is that the representational capacity of training samples can be generalized by NFL, i.e., NFL can add extra information to training samples. For a given training sample set, quantity of line features (corresponding to the FLs) based on NFL is always larger than that of point features (corresponding to sample points). Because if there are $N_\theta$ training samples belonging to class $\theta$, there will be $N_\theta(N_\theta-1)/2$ feature lines. $N_\theta(N_\theta-1)/2$ feature lines always carry more information than the $N_\theta$ sample points.

However the extra information might be a double-edged sword. Although successful in improving the classification performance, there are still some drawbacks in NFL that will limit its further application in practice. Two main categories of drawbacks can be summarized as follows:

(a) It will encounter large computational complexity when there are too many samples in each class. As referred above, if there are $N_\theta$ training samples belonging to class $\theta$, there will be $N_\theta(N_\theta-1)/2$ feature lines. Suppose that each sample's feature vector is $n$ dimension, for the traditional NFL, there are $(3\times n+1)\times(N_\theta\times(N_\theta-1)/2)$ multiplication operations. Some research attempted to resolve such a problem, e.g. in (Chen et al., 2002), a fast algorithm to calculate $d^\theta(x_q, x_i^\theta x_j^\theta)$ is proposed, the computation time of which is only 1/3 of the traditional NFL.

(b) It will encounter the problem of extrapolation inaccuracy and interpolation inaccuracy, which are discussed further as follows.

### 2.2. Inaccuracy of NFL caused by trespass

In NFL, the perpendicular distance between the query sample and the FL is used as the decision criterion. When a straight line of one class trespasses into one of other classes area, according to NFL, it may lead to classification error. Two types of trespassing are discussed in this section, which can cause two types of inaccuracies: extrapolation inaccuracy and interpolation inaccuracy (Du and Chen, 2007).

#### 2.2.1. Extrapolation inaccuracy

In Fig. 2, the query sample $x_q$ is surrounded by the samples belonging to class "Star" ($x_1^s,\ldots,x_4^s$ in Fig. 2), but it is classified to class "Circle" with the decision criterion of NFL illustrated in (4). This classification error is due to the extrapolating part of FL $\overline{x_1^c x_2^c}$.

It can be proved that extrapolation inaccuracy can be ignored if the dimension of feature space is large enough (Du and Chen, 2007). But in a feature space with low dimension, it actually harms. To counteract the extrapolation inaccuracy, several researchers made their helpful attempts. The author in (Zheng et al., 2004) proposed nearest neighbor line (NNL). In NNL, only one feature line is

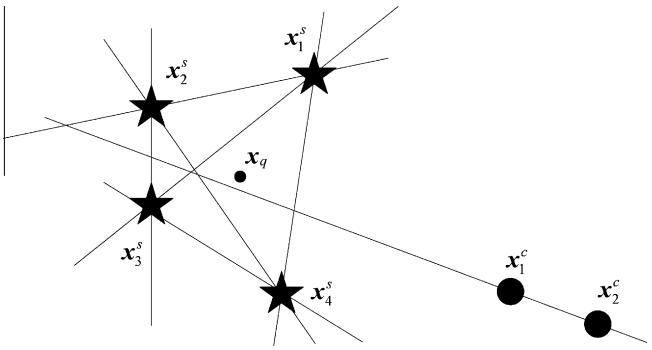


**Fig. 2.** Extrapolation inaccuracy.

used to represent one class, which links the query sample's nearest two sample points. NNL can avoid the extrapolation problem, but it may also cause the loss of classification ability (Du and Chen, 2007). In (Zhou and Kwoh, 2004), nearest feature midpoint (NFM) was proposed which attempt to overcome the extrapolation inaccuracy by defining the distance metric as the minimum Euclidean distance between the query point and the feature line's midpoint instead of the perpendicular distance defined in NFL. In other researches such as tunable nearest neighbor (TNN) (Zhou et al., 2004b), rectified nearest feature line segment (RNFLS) (Du and Chen, 2007) and extended nearest feature line(ENFL) (Zhou et al., 2004a), they modified the original NFL distance metric definitions according to the relative position relationship between the query point and the feature line (According to the definitions and rules in RNFLS, the two points constructing the line segment can be the same point. The distance definition in RNFLS also includes the point to point distance). Thus the query points with the same original NFL distance can be dealt with discrimination and extrapolation inaccuracy can be counteracted.

*2.2.2. Interpolation inaccuracy*

In Fig. 3, an example of interpolation inaccuracy in NFL is illustrated. The query sample $x_q$ is surrounded by the samples belonging to class "Star", but $x_q$ is classified to class "Circle". This is due to the territory of class "Star" is trespassed by interpolating part of FL $\overline{x_1^c x_2^c}$.

To counteract the interpolation inaccuracy, authors in (Du and Chen, 2007) proposed a method for sample selection. They construct rectified nearest feature line segment (RNFLS) subspaces to represent each class by having removed those line segments trespassing the class-areas of other classes.

Fig. 4 illustrates a case which can produce both the extrapolation inaccuracy and interpolation inaccuracy. The NFL-refined methods referred above, including NNL, NFM, TNN, can only suppress the extrapolation inaccuracy. The RNFLS method can counteract both extrapolation and interpolation inaccuracy and it can implement classification well in the case as illustrated in Fig. 4. In the same environment, the classification error will increase significantly for NFL. However, RNFLS also has its drawbacks. Its implementation is relatively complex. To overcome the interpolation inaccuracy, RNFLS has the procedure of sample subspace selection. To overcome the extrapolation inaccuracy, its calculation methods (or definitions) of the distance metric are various in different cases. Its computational complexity is relatively high. In next section, a novel feature line segment approach is proposed, which is simple yet effective and can suppress both the interpolation inaccuracy and extrapolation inaccuracy.

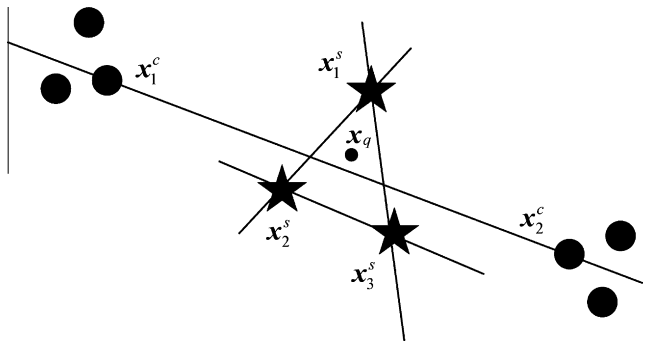


**Fig. 3.** Interpolation inaccuracy.

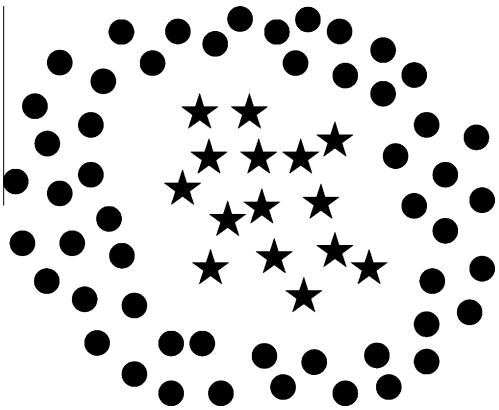

**Fig. 4.** Sample distribution causing significant interpolation and extrapolation inaccuracy problem.

## 3. Shortest feature line segment classification approach

To utilize the advantages brought by the line features and to counteract the drawbacks of the traditional NFL and the modified NFL methods, SFLS is proposed in our research. The basic ideas are as follows. SFLS does not calculate the distance between the query point and the feature line. Instead, it attempts to find the shortest feature line segment which satisfies given geometric relation constraints together with the query sample. Two samples in the same class constitute a feature line segment. If the query point is inside or on the hypersphere centered at the midpoint of the feature line segment (obviously, the diameter of the hypersphere is the length of the feature line segment), the corresponding feature line segment will be tagged. Then find out the shortest feature line segment in all the tagged feature line segments and assign the shortest feature line's class label to the query sample. The specific classification procedure is as follows:

Let $x_i^\theta, x_j^\theta \quad (i \neq qj, 1 \leq i, j \leq N_\theta)$ be two training samples belonging to the same class $\theta$, $\theta = 1, \ldots, M$, where $M$ represents the number of class and $N_\theta$ represents the number of samples belonging to the class $\theta$. When there are $N_\theta$ samples belong to class $\theta$ in training set, there will be $N_\theta(N_\theta - 1)/2$ feature lines segments. For a query sample point $x_q$, execute the steps as follows:

1) Calculate the angle (denoted by $\alpha$) between vector $x_q - x_i^\theta$ and vector $x_q - x_j^\theta$ based on (6):

$$\alpha = \frac{180}{\pi} \cdot \arccos \frac{(x_q - x_i^\theta)^T (x_q - x_j^\theta)}{\|\overline{x_q x_i^\theta}\| \cdot \|\overline{x_q x_j^\theta}\|}, \tag{6}$$

where $x_q$, $x_i^\theta$ and $x_j^\theta$ are all column vectors and $\|\cdot\|$ denotes the Euclidean norm. The unit of $\alpha$ is deg.
If $\alpha$ is an acute angle (i.e., $0 \leqslant \alpha < 90$), leave feature line segment $\overline{x_i^\theta x_j^\theta}$ untagged.

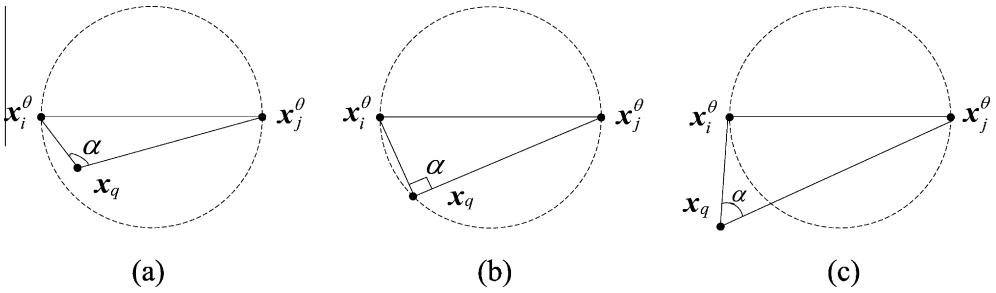


**Fig. 5.** Relative position relation between query point and feature line segment.

If $\alpha$ is a right angle or an obtuse angle (i.e., $90 \leqslant \alpha \leqslant 180$), tag feature line segment $\overline{x_i^\theta x_j^\theta}$.

2) Find out the shortest tagged feature line segment:

$\overline{x_i^{\theta^*} x_j^{\theta^*}}$, label class $\theta^*$ to query point $x_q$. If there is no tagged feature line, then the corresponding query point is rejected to be classified.

Based on the basic geometric theorem, as illustrated in Fig. 5, if $\alpha$ is an acute angle, then $x_q$ is outside the hypersphere; if $\alpha$ is a right angle, then $x_q$ is on the hypersphere; if $\alpha$ is an obtuse angle, then $x_q$ is inside the hypersphere. The hypersphere's center is the midpoint of the feature line segment $\overline{x_i^\theta x_j^\theta}$ and its diameter is the length of the feature line segment $\overline{x_i^\theta x_j^\theta}$.

The classification procedure can be briefly illustrated in Fig. 6.

In Fig. 6, query point $x_q$ is inside hypersphere centered at midpoint of $\overline{x_1^s x_2^s}$, whose diameter is $\|x_1^s - x_2^s\|$. $x_q$ is also inside hypersphere $C_c$ centered at midpoint of $\overline{x_1^c x_2^c}$, whose diameter is $\|x_1^c - x_2^c\|$. Feature line segment $\overline{x_1^c x_2^c}$ and $\overline{x_1^s x_2^s}$ are both tagged. $\overline{x_1^c x_2^c}$ is shorter than $\overline{x_1^s x_2^s}$, so query point $x_q$ is labeled class "Circle".

In the worst case, i.e., there is no tagged feature line for a test sample $x_q$, if the rejection decision is not permitted, just use the rule of nearest neighbor (NN) instead to make the classification decision for the samples which are reject to be classified by SFLS.

## 4. Analysis of SFLS

SFLS proposed in this paper has powerful classification ability and it can suppress some significant drawbacks of NFL. SFLS is also relatively easy to be implemented and it has relatively low computational complexity.

### 4.1. Classification ability analysis

#### 4.1.1. SFLS maintains the advantages of the NFL

Obviously, one feature line in NFL corresponds to one feature line segment in SFLS. SFLS retains the ideas and advantages of NFL, i.e., it uses a linear model of each pair of sample points within the same class to generalize the representational capacity of sample set.

Although perpendicular distance used in NFL has its deficiencies, it still has its rationality to represent the similarity between the query point and some class. SFLS does not directly use the perpendicular distance from the query sample $x_q$ to the feature line segment $\overline{x_i^\theta x_j^\theta}$ in classification, but it can be considered to use the perpendicular distance indirectly, which can be justified as follows:

When the shortest feature line segment of each class which satisfies the geometric condition defined is found, the maximum value of the distance from the query sample $x_q$ to the shortest feature line segment $\overline{x_i^\theta x_j^\theta}$ of each class is determined, as illustrated in Fig. 7.

In Fig. 7, the perpendicular distance from $\mathbf{x}_q$ to the shortest feature line segment $\overline{x_i^\theta x_j^\theta}$ of class $\theta$, denoted by $d_q^\theta$, satisfies $d_q^\theta \leq d_{\max}^\theta = \|x_i^\theta - x_j^\theta\|/2$. The classification decision of SFLS in fact can be considered as being based on the perpendicular distance, but the distance used for decision is the maximum value (or upper bound) of the corresponding distance from the query point $x_q$ and the shortest feature line segment $\overline{x_i^\theta x_j^\theta}$ in each class. To compare the upper bound of perpendicular distance is more conservative than to compare the distance directly. SFLS is equivalent to NFL to some extent, but it is relatively conservative. Sometimes, more conservative means more reliable.

#### 4.1.2. SFLS has the concentration property of feature line segment

SFLS uses the feature line segment instead of feature line. The feature space constituted by feature line segments has a good property, i.e., the concentration (Du and Chen, 2007). The distribution of line segments is denser at the center than at the boundary if the distribution of original sample points is under a uniform density. A Gaussian distribution can be viewed as a pile-up of several uniform distribution disks with the same center but different radius. It is conjectured that this concentration property also applies to the Gaussian case. For a two-class classification problem with

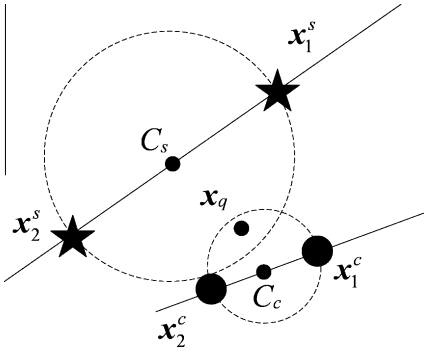


**Fig. 6.** Shortest feature line segment classification.

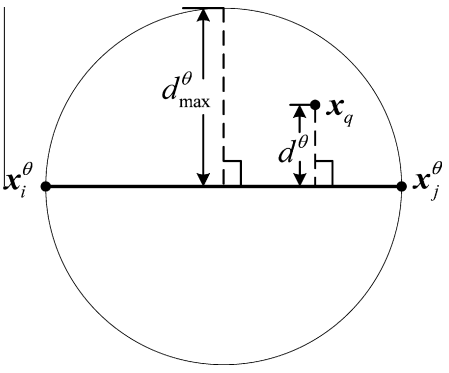


**Fig. 7.** Perpendicular distance and length of the feature line segment.

Gaussian distribution, it can be concluded that stronger concentration property can bring the improvement of classification accuracy (Du and Chen, 2007). The author in (Du and Chen, 2007) referred that the concentration property can be extended to classification problems in which the overlapping is caused by noise scattering of two or more classes under similar distribution but different centers. It reverses the scattering and achieves a substantial improvement. Detailed descriptions on the concentration property and related proof can be found in (Du and Chen, 2007).

*4.1.3. SFLS has more reasonable similarity definition*

NFL only uses the perpendicular distance from the query point to the feature line as the criterion for classification decision. The perpendicular distance is in fact the similarity between point feature and line feature. What we concern is the similarity between the query sample and the given class. Only such a distance is not sufficient to represent the similarity between query sample and the feature line's corresponding class. Less perpendicular distance does not always represent more similarity. Such a proposition can be supported by the two kinds of inaccuracies aforementioned. In SFLS, as referred above, both the distance and the volume of the hypersphere constituted by feature line segment and query point are emphasized to define the similarity. For a tagged feature line segment, the shorter the feature line segment is, the less the volume of the hypersphere covering the query point and the feature line segment is. Less volume represents more similarity between query point and feature line segment.

According to the similarity definition in SFLS, there exists a good property: the query point is always near to both the two points constituting the shortest feature line segment. It can better reflect the similarity between the query point and the corresponding class. Such a property can be proved as follows:

As illustrated in Fig. 5, according to the basic geometric theorem:

If a query sample $x_q$ is outside the hypersphere generated based on a feature line segment $\overline{x_i^\theta x_j^\theta}$, then

$$\|x_q x_i^\theta\|^2 + \|x_q x_j^\theta\|^2 > \|x_i^\theta x_j^\theta\|^2 \tag{7}$$

comes into existence;

If a query sample $x_q$ is inside or on the hypersphere generated based on a feature line segment $\overline{x_i^\theta x_j^\theta}$, then

$$\|x_q x_i^\theta\|^2 + \|x_q x_j^\theta\|^2 \leq \|x_i^\theta x_j^\theta\|^2 \tag{8}$$

comes into existence.

Suppose that $\overline{x_i^{\theta^*} x_j^{\theta^*}}$ is the shortest feature line segment of $x_q$ (i.e., SFLS label class $\theta^*$ to query point $x_q$) and $\|x_i^{\theta^*} x_j^{\theta^*}\| = r$, then

$$\|x_q x_i^{\theta^*}\|^2 + \|x_q x_j^{\theta^*}\|^2 \leq r^2. \tag{9}$$

The quadratic sum has an upper bound.

For the shortest feature line segment, the length $r$ is always relatively small thus the quadratic sum has a relatively small upper bound. So, in SFLS, the query sample $x_q$ is always near to both the two points $(x_i^{\theta^*}, x_j^{\theta^*})$ constituting the shortest feature line segment.

Based on the analysis above, it can be concluded that the similarity definition in SFLS should be more reasonable and more comprehensive to describe the query point and the class represented by the feature line segment.

*4.1.4. SFLS can suppress the inaccuracies caused by the trespass*

**First**, SFLS uses line segment instead of straight line. Intuitively, a line segment has finite length while a straight line extends to infinity. Thus the probability to produce trespass problems can be reduced when using feature line segment.

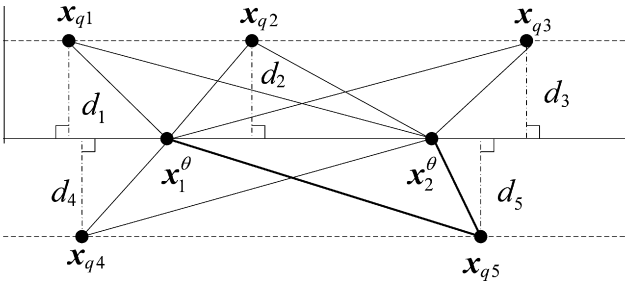


**Fig. 8.** Sample points with the same NFL distance.

**Second**, in Fig. 8 we can see that, all the points at two parallel dash lines have the same NFL distance according to FL $\overline{x_1^\theta x_2^\theta}$. All the query points with the same NFL distance will be treated indiscriminatingly. This is an important reason for generation of extrapolation and interpolation inaccuracies. While in SFLS, all the query samples with the same NFL distance are treated discriminatingly. Because, not all the query samples satisfy the geometric relation defined in SFLS with $\overline{x_1^\theta x_2^\theta}$ in Fig. 8.

**Third**, SFLS has no problem of extrapolation inaccuracy. This can be proved geometrically as follows:

In NFL, $\mu$ denotes the position parameter. Rewrite (2) as follows:

$$\mu = \frac{(x_q - x_i^\theta)^T (x_j^\theta - x_i^\theta)}{(x_j^\theta - x_i^\theta)^T (x_j^\theta - x_i^\theta)}.$$

In Fig. 1, the projection of $x_q$ onto feature line $\overline{x_i^\theta x_j^\theta}$ is denoted as $x_{ij}^\theta$. Define the direction from $x_i^\theta$ to $x_j^\theta$ as the forward direction.

When $\mu < 0$, $x_{ij}^\theta$ is a backward extrapolating point on the $x_i^\theta$ side.
When $\mu > 1$, $x_{ij}^\theta$ is a forward extrapolating point on the $x_j^\theta$ side.
When $0 < \mu < 1$, $x_{ij}^\theta$ is an interpolating point between $x_i^\theta$ and $x_j^\theta$.

In SFLS, for query point $x_q$ and the tagged feature line segment $\overline{x_i^\theta x_j^\theta}$, $0 \leqslant \mu \leqslant 1$ definitely comes into existence. Thus, there is no problem of extrapolation inaccuracy in SFLS.

**Fourth**, SFLS can suppress the interpolation inaccuracy in some cases as illustrated in Fig. 9.

In Fig. 9, for query point $x_q$, the shortest feature line segment in class "circle" is $\|\overline{x_i^c x_j^c}\|$ and in class "Star" is $\|\overline{x_i^s x_j^s}\|$. The query point $x_q$ surrounded by samples of class "Star" is classified to class "Circle" by NFL due to the interpolation inaccuracy ($d^c < d^s$). But it is correctly classified to class "Star" by SFLS because $\|\overline{x_i^c x_j^c}\| > \|\overline{x_i^s x_j^s}\|$.

As referred above, the similarity definition in SFLS is more reasonable. In SFLS, not all the feature line segments are proper to be used as representatives for a class. SFLS can be considered as performing a selection of all feature line segments according to the

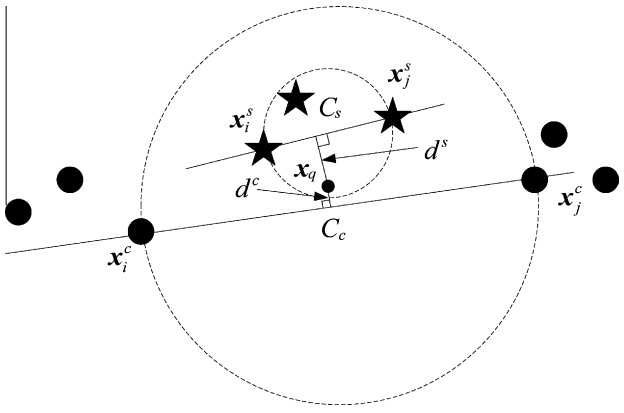


**Fig. 9.** NFL and SFLS in interpolation inaccuracy problem.

geometric relation defined before performing the classification. The sample pair constituting the shortest feature line segment is always in the neighborhood of the query point in SFLS. The occurrence possibility of trespassing problems (including extrapolation and interpolation) is relatively low. Consequently, interpolation inaccuracy problem can be suppressed to some extent in SFLS.

It should be noted that SFLS can only reduce the probability to produce interpolation inaccuracy, i.e., SFLS can not thoroughly resolve interpolation inaccuracy. No classification approach is panacea.

It can be concluded that SFLS proposed in this paper retains the advantages of NFL and it has the good property of concentration. SFLS has relatively more reasonable similarity definition and it can counteract the inaccuracies caused by the trespass. Based on analysis above, for SFLS, better classification performance can be expected.

### 4.2. Computational complexity analysis

If there are $N_\theta$ samples belong to class $\theta$ in training set, there exist $N_\theta(N_\theta-1)/2$ feature line segments. Suppose that feature vector of each sample has $n$ dimension, the computation in SFLS for each class includes $(3\times n)\times(N_\theta\times(N_\theta-1)/2)$ multiplication operations which is less than that of original NFL: $(3\times n+1)\times(N_\theta\times(N_\theta-1)/2)$. And it should be noted that there is no offline samples subspace selection or preparation or the classification based on SFLS. In RNFLS, offline preparation is needed, which has computation time complexity of $O(N^3)$, where $N$ denotes the number of training samples.

In general, SFLS proposed in this paper is a simple yet effective classification method. It has clear geometric–theoretic foundation. It retains the main ideas and advantages of NFL and at the same time it can counteract extrapolation inaccuracy and interpolation inaccuracy. SFLS attempts to find the shortest feature line segment which is "near" to the query point. However, the "near" is not defined based on the perpendicular distance from the query point to the feature line. It uses the geometric relation to define a more reasonable "near", i.e., the similarity.

## 5. Experiments

The classification performance of SFLS is compared with classification approaches including NN, $k$-NN, traditional NFL, NNL, TNN, RNFLS. The experiments are executed on four artificial datasets and some real-world datasets including UCI (BLAKE and Merz, 1998).

### 5.1. Artificial dataset: two-spiral

The two spiral curves can be generated based on (10) and (11) as follows (Denoeux and Lengelle, 1993; Sin and DeFigueiredo, 1993; Chen et al., 1994; Du and Chen, 2007):

$$spiral1: \begin{cases} x = k\theta\cos(\theta) \\ y = k\theta\sin(\theta) \end{cases} \tag{10}$$

$$spiral2: \begin{cases} x = k\theta\cos(\theta+\pi) \\ y = k\theta\sin(\theta+\pi) \end{cases} \tag{11}$$

Set parameter $\pi/2 \leqslant \theta \leqslant 3\pi$ and suppose that each class's probability density is uniform along the corresponding curve. The two-spiral dataset for experiment is polluted by Gaussian noise whose mean is zero and variance is $\sigma = 1.5$. The two spiral curves and the polluted two spiral curves for experiment are illustrated in Fig. 10.

Totally 500 samples are generated. Class 1 and class 2 each has 250 samples. Randomly select 125 training samples from each class, the remainder are reserved for test. Then we use NN, $k$-NN, NFL, NNL, RNFLS and the SFLS proposed to perform the classification. The $k$ parameter in $k$-NN is set to 3. The classification procedure referred above is executed for 10 times. In each time, samples for training and test are re-selected randomly. The experimental results are listed in Table 1.

Based on the results listed in Table 1, it can be seen that NFL is not proper to be used in two-spiral classification problem. In this experiment, NN and $k$-NN perform better than NFL, because the sample point's corresponding feature vector is 2-dimension, which is low. As referred in section 2.2.1, the trespass will harm when the feature vector of the sample point has relatively low dimension. The scale of sample set in this experiment is not too limited. In such a case, generalizing representational capacity by using line feature instead of point feature to improve classification performance may not be as effective as the one in small-scale sample sets. RNFLS gains better performance due to the building of the RNFLS subspace and various definition of distance. SFLS proposed in this paper is relatively simple but it also achieves the same classification accuracy as RNFLS due to its classification ability analyzed in section 4.1.

### 5.2. Artificial dataset: three-spiral

The three spiral curves can be generated based on (12)–(14) as follows:

**Table 1**
Two-spiral classification performance.

| Classification approach | Correct rate (min value) (%) | Correct rate (max value) (%) | Correct rate (mean value) (%) |
|---|---|---|---|
| NN | 93.6 | 96.8 | 95.5 |
| $k$-NN | 94.8 | 98.4 | 96.6 |
| NFL | 52.0 | 63.2 | 56.0 |
| NNL | 85.2 | 91.6 | 89.2 |
| RNFLS | 94.0 | 97.2 | 96.0 |
| SFLS | 94.0 | 97.2 | 96.0 |

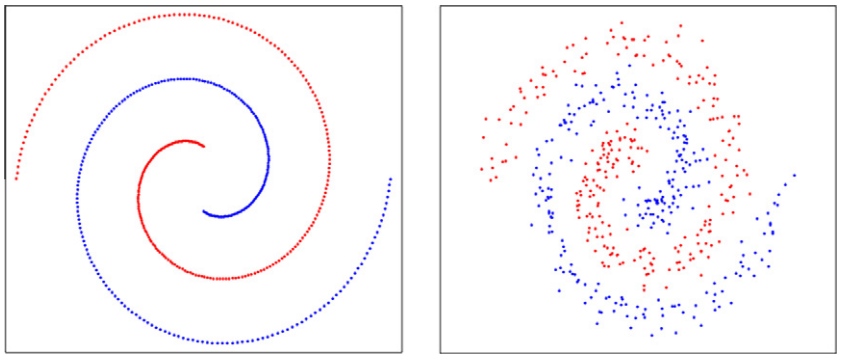

**Fig. 10.** Two-spiral curves (a) Two-spiral; (b) Two-spiral polluted with Gaussian noise.

$$spiral1: \begin{cases} x = (k_1\theta + a_1)\cos(\theta) \\ y = (k_1\theta + a_1)\sin(\theta) \end{cases} \quad (12)$$

$$spiral2: \begin{cases} x = (k_2\theta + a_2)\cos(\theta) \\ y = (k_2\theta + a_2)\sin(\theta) \end{cases} \quad (13)$$

$$spiral3: \begin{cases} x = (k_3\theta + a_3)\cos(\theta) \\ y = (k_3\theta + a_3)\sin(\theta) \end{cases} \quad (14)$$

Set parameter $a_1 = 0$, $a_2 = 6$, $a_3 = 12$, $k_1 = k_2 = k_3 = 3$ and suppose that each class's probability density is uniform along the corresponding curve. The three-spiral dataset for experiment is polluted by Gaussian noise whose mean is zero and variance is $\sigma = 1.5$. The three-spiral curves and the polluted three-spiral curves for experiment are illustrated in Fig. 11.

Totally 750 samples are generated. Class 1, class 2 and class 3 each has 250 samples. Randomly select 125 training samples from each class, the remainder are reserved for test. Then we use NN, $k$-NN, NFL, NNL, RNFLS and the SFLS proposed to perform the classification. The $k$ parameter in $k$-NN is set to 3. The classification procedure referred above is executed for 10 times. In each time, samples for training and test are re-selected randomly. The experimental results are listed in Table 2.

For the data of three classes of spirals, the chance to encounter trespass has been significantly increased. Based on the experimental results listed in Table 2, it can be concluded that NFL is not proper to be used in such an environment. Other NFL modified approaches also cannot achieve ideal classification performance. SFLS proposed in this paper achieve the best performance among all the NFL modified or refined approaches.

### 5.3. Artificial dataset: two classes with Gaussian distribution

To compare the classification performance of the SFLS and other classification methods with that of Bayesian classifier, we generate the artificial two-class dataset with known probability density. Any sample in our dataset has two dimensions $(x, y)$. $x$ is a Gaussian distributed variable and $y$ is a uniform distributed variable. Their joint-probability density function is as follows (Du and Chen, 2007):

$$p_{class1}(x,y) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma(b-a)}\exp[-\frac{1}{2}(\frac{x}{\sigma})^2], & a \le y \le b, \\ 0, & \text{otherwise}; \end{cases}$$
$$p_{class2}(x,y) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma(b-a)}\exp[-\frac{1}{2}(\frac{x-\mu}{\sigma})^2], & a \le y \le b, \\ 0, & \text{otherwise}; \end{cases} \quad (15)$$

where $\mu$ denotes the distance between the two Gaussian centers. Totally 500 samples are generated. Class 1 and class 2 each have 250 samples. Randomly select 125 training samples from each class, the remainder are reserved as test samples. About 10 different $\mu$ values are used in our experiment ($\mu = 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5$). The classification procedure referred above is executed for 10 times. In each time, samples for training and test are re-selected randomly. We use NN, $k$-NN, NFL, NNL, RNFLS and the SFLS proposed to perform the classification. The $k$ parameter in $k$-NN is set to 3. The experimental results are listed in Fig. 12 and Table 3.

Based on the results, it can be concluded that the NFL is not fit for this type of dataset. Because the sample point's corresponding feature vector is 2-dimension, which is low, the extrapolation inaccuracies for NFL actually harm. The scale of sample set in this experiment is not too limited. The NN, $k$-NN and the NFL-refined approaches can achieve relatively good classification performance.

**Table 2**
Three-spiral classification performance.

| Classification approach | Correct rate (min value) (%) | Correct rate (max value) (%) | Correct rate (mean value) (%) |
|---|---|---|---|
| NN | 68.8 | 75.7 | 73.3 |
| $k$-NN | 74.9 | 79.5 | 77.4 |
| NFL | 34.9 | 42.4 | 38.1 |
| NNL | 63.2 | 73.3 | 68.6 |
| RNFLS | 73.1 | 79.2 | 75.7 |
| SFLS | 74.7 | 80.3 | 77.4 |

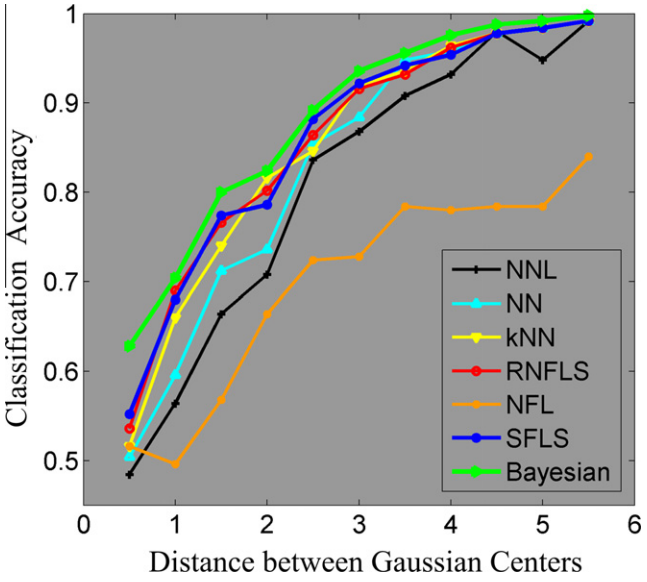


**Fig. 12.** Classification performances of two-Gaussian distribution samples.

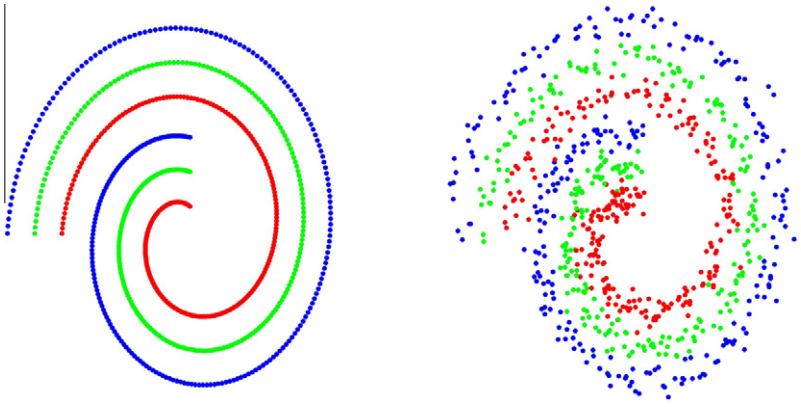

**Fig. 11.** Three-spiral curves: (a) three-spiral; (b) three-spiral polluted with Gaussian noise.

**Table 3**
Classification performances of two-Gaussian distribution samples (mean values).

| Classification approach | Mean classification correct rate of experiments with various distance of Gaussians centers (%) |
|---|---|
| Bayesian | 88.1 |
| NN | 83.1 |
| 3-NN | 85.0 |
| NFL | 69.7 |
| NNL | 80.8 |
| RNFLS | 86.2 |
| SFLS | 86.6 |

SFLS and RNFLS both achieve two top performances except for Bayesian approach due to their wonderful classification abilities, however, SFLS is simpler.

### 5.4. Artificial dataset: Ripley

Ripley dataset (Ripley, 1996) is a two-class dataset, which includes 125 training samples per class and 500 test samples per class. The data in each class are generated with the mixed-Gaussian distribution. The classification procedure referred above is executed for 10 times. In each time, samples for training and test are re-selected randomly. NN, $k$-NN, NFL, NNL, RNFLS and the SFLS proposed are used to perform the classification. The $k$ parameter in $k$-NN is set to 3. The experimental results are listed in Table 4.

Based on the results, it can be concluded that the NFL is not fit for this type of dataset. Because the sample point's corresponding feature vector is 2-dimension, which is low, thus the extrapolation inaccuracies actually harm. The scale of sample set in this experiment is not too limited. The NN, $k$-NN and the NFL-refined approaches can achieve relatively good classification performance. SFLS proposed in this paper derive the best performance.

**Table 4**
Classification performances of Ripley dataset.

| Classification approach | Mean classification correct rate of experiments with various distance of Gaussians centers (%) |
|---|---|
| NN | 85.0 |
| 3-NN | 86.6 |
| NFL | 69.0 |
| NNL | 80.7 |
| RNFLS | 88.8 |
| SFLS | 89.7 |

**Table 5**
UCI datasets used in the experiments.

| Datasets | Number of class | Dimension of features | Number of samples |
|---|---|---|---|
| Iris | 3 | 4 | 150 |
| Housing | 6 | 13 | 506 |
| Pima | 2 | 8 | 768 |
| Wine | 3 | 13 | 178 |
| Bupa | 2 | 6 | 345 |
| Ionosphere | 2 | 34 | 351 |
| WDBC | 2 | 30 | 569 |
| Glass | 6 | 9 | 214 |

### 5.5. Classification problem based on UCI dataset

To further verify the SFLS's performance, we use some real-world datasets from UCI(Blake and Merz, 1998) listed in Table 5, which include some multi-class (class number $\geqslant 3$) datasets:

In our experiment, we do not deal with the missing data problem, all the samples with missing values are eliminated. Features of the samples are normalized by their means and standard deviations before classification. Leave-one-out cross-validation approach is used in our experiment. The classification performance derived based on different classification approaches are listed in Table 6.

Based on the experimental results listed in Table 6, it can be concluded that for the datasets with relatively low feature dimension and high number of class, the performance of NFL would be even worse, which can be verified by the classification performance on the datasets of Iris, Pima and Glass, etc. For the datasets with less class number and higher feature dimension, e.g. WDBC and Ionosphere, the performances of NFL are relatively ideal. These agree with the analysis referred in section2, i.e., when there are lower feature dimension and more classes in some dataset, the trespass inaccuracies of NFL are more significant.

Both the RNFLS and the SFLS approach proposed in this paper always achieve better performance among all the methods listed in Table 6. Especially for the datasets with relatively low feature dimension, the proposed SFLS performs the best. This can be verified by the classification performance on datasets of Iris and Pima listed in Table 6.

### 5.6. Discussion

Several NFL-refined classification approaches are used in the experiments. SFLS and RNFLS always perform better than original NFL and other NFL-refined methods. When compared to other methods based on the line feature, SFLS has more rational and more comprehensive definition of similarity, which is crucial for the classification problems. Compared to RNFLS, both SFLS and RNFLS uses feature line segment which has the good property of concentration and they can suppress the inaccuracies caused by trespass, but SFLS does not need the preparation procedure for samples subspace. SFLS uses simply the length of feature line segment instead of various distance definitions in RNFLS and has no pre-selection procedure. In general, SFLS is relatively simple yet effective. These can be supported by the provided experimental results.

When the scale of sample set is not too limited and when the feature dimension is relatively low, the advantages of line features (or line segment features) can not be distinctly shown when compared with the traditional classification approach such as NN and $k$-NN. It can be found in the experimental results that under such a circumstance (e.g. the experiments based on artificial datasets),

**Table 6**
Classification performances on UCI datasets.

| Classification approach | Iris (%) | Housing (%) | Pima (%) | Wine (%) | Bupa (%) | Iono-sphere (%) | WDBC (%) | Glass (%) |
|---|---|---|---|---|---|---|---|---|
| NN | 94.7 | 70.8 | 70.6 | 95.5 | 63.2 | 86.3 | 95.1 | 70.1 |
| 3-NN | 94.7 | **73.0** | **73.6** | 95.5 | 65.2 | 84.6 | 96.5 | **72.0** |
| NFL | 88.7 | 71.1 | 67.1 | 92.7 | 63.5 | 85.2 | 95.3 | 66.8 |
| NNL | 94.7 | 67.6 | 62.8 | 78.7 | 57.4 | 87.2 | 64.0 | 65.4 |
| RNFLS | **95.3** | **73.5** | 73.0 | **97.2** | **66.4** | **94.3** | **97.2** | **72.0** |
| SFLS | **96.0** | 72.7 | **73.6** | **96.1** | **65.5** | **92.4** | **96.8** | 70.1 |

NFL and other NFL –refined approaches lose their classification ability while the proposed SFLS can also achieve the relatively ideal classification performance. Furthermore, the proposed SFLS is a nonparametric classifier, i.e., it has no problem of parameter selection which is necessary in the classifier such as $k$-NN (the choice of $k$).

## 6. Conclusion

In this paper, a novel nonparametric classification approach called SFLS is proposed. SFLS can be considered as a refined NFL method which can substantially improve the classification performance of NFL. Theoretical justification and analysis of SFLS' rationality are provided. Experimental results based on artificial dataset and real-world dataset also show that SFLS is an effective classification approach.

In the future, we will do further research in SFLS. The computational cost of SFLS is still relatively high. Faster and more effective algorithms for SFLS are worth researching. Some modifications to SFLS can be made according to other effective classification approaches. For example, based on SFLS, $k$-SFLS can be developed, just like the idea of $k$-NN, better performance can be expected. In this paper, we only discuss about the classification problem with Euclidean distance. Other distance definition can also be used to implement the SFLS approach.

It should be noted that when the sample set is too small, the rejection rate of the proposed SFLS will increase, because the geometric relation requirement in SFLS might be relatively strict. As referred above when the decision of rejection is not permitted, NN and other approach can be used to temporarily take place of SFLS for the sample point encountered rejection. Such a method is just an expedient. Further research works are required to reduce the rejection rate of SFLS. This is also an important work in future.

## Acknowledgements

This work is supported by the Grant for State Key Program for Basic Research of China (973): No. 2007CB311006. We are also grateful to the anonymous reviewers for their helpful comments and suggestions that greatly improve the paper.

## References

Blake, C.L., Merz, C.L., 1998. UCI repository of machine learning databases. <http://www.ics.uci.edu/~mlearn/MLRepository.html>.
Chaudhuri, B.B., 1996. A new definition of neighborhood of a point in multi-dimensional space. Pattern Recognit. Lett. 17, 11–17.
Chen, K., Wu, T.Y., Zhang, H.J., 2002. On the use of nearest feature line for speaker identification. Pattern Recognit. Lett. 23, 1735–1746.
Chen, Y.Q., Thomas, D.W., Nixon, M.S., 1994. Generating-shrinking algorithm for learning arbitrary classification. Neural Networks 7, 1477–1489.
Denoeux, T., Lengelle, R., 1993. Initializing back propagation networks with prototypes. Neural Networks 6, 351–363.
Du, H., Chen, Y.Q., 2007. Rectified nearest feature line segment for pattern classification. Pattern Recognit. 40, 1486–1497.
Duda, O.R., Hart, E.P., Stork, D.G., 2001. Pattern Classification. Wiley Inter-Science Publication, New York.
Gao, Q.B., Wang, Z.Z., 2007. Center-based nearest neighbor classifier. Pattern Recognit. 40, 346–349.
He, Y.H., Year. of Conference Face recognition using kernel nearest feature classifiers. 2006 International Conference on Computational Intelligence and Security, 1, pp. 678–683.
Li, S.Z., 2000. Content-based audio classification and retrieval using the nearest feature line method. IEEE Trans. Speech Audio Process. 8, 619–625.
Li, S.Z., 2008. Nearest feature line. Scholarpedia 3, 4357.
Li, S.Z., Chan, K.L., Wang, C.L., 2000. Performance evaluation of the nearest feature line method in image classification and retrieval. IEEE Trans. Pattern Anal. Mach. Intell. 22, 1335–1339.
Li, S.Z., Lu, J.W., 1999. Face recognition using the nearest feature line method. IEEE Trans. Neural Networks 10, 439–443.
Pang, Y., Yuan, Y., Li, X., 2007. Generalized nearest feature line for subspace learning. IEEE Electron. Lett. 43, 1079–1080.
Ripley, B.D., 1996. Pattern recognition and neural networks. Cambridge University Press, Cambridge.
Sanchez, J.S., Pla, F., Ferri, F.J., 1997. On the use of neighbourhood-based non-parametric classifiers. Pattern Recognit. Lett. 18, 1179–1186.
Sin, S.-K., DeFigueiredo, R.J.P., 1993. Efficient learning procedures for optimal interpolative nets. Neural Networks 6, 99–113.
Theodoridis, S., Koutroumbas, K., 2006. Pattern Recognit.. Elsevier, Singapore.
Yang, Y., Han, C.Z., Han, D.Q., 2009. A novel feature line segment approach for pattern classification. 12th International Conference on Information Fusion, pp. 490–497.
Zheng, W.M., Zhao, L., Zou, C.R., 2004. Locally nearest neighbor classifiers for pattern classification. Pattern Recognit. 37, 1307–1309.
Zhou, Y.L., Zhang, C.S., Wang, J.C., 2004a. Extended nearest feature line classifier. In: PRICAI 2004 Trends in Artificial Intelligence. Lecture Notes in Computer Science. pp. 183190.
Zhou, Y.L., Zhang, C.S., Wang, J.C., 2004b. Tunable nearest neighbor classifier. Pattern Recognition. Lecture Notes in Computer Science. Springer, Berlin / Heidelberg. pp. 204-211.
Zhou, Z.L., Kwoh, C.K., 2004. The pattern classification based on the nearest feature midpoints. In: Proceedings of the 17th Internat. Conf. on Pattern Recognition. pp. 446–449.