Rectified Nearest Feature Line Segment for Pattern Classification

Hao Du and Yan Qiu Chen*

Department of Computer Science and Engineering,

School of Information Science and Engineering,

Fudan University,

Shanghai 200433, China

Emails: [duhao|chenyq]@fudan.edu.cn

Abstract

This paper points out and analyzes the advantages and drawbacks of the Nearest Feature Line (NFL) classifier. To overcome the shortcomings, a new feature subspace with two simple and effective improvements is built to represent each class. The proposed method, termed Rectified Nearest Feature Line Segment (RNFLS), is shown to possess a novel property of concentration as a result of the added line segments (features), which significantly enhances the classification ability. Another remarkable merit is that RNFLS is applicable to complex tasks such as the two-spiral distribution, which the original NFL cannot deal with properly. Finally, experimental comparisons with NFL, NN(Nearest Neighbor), k-NN and NNL (Nearest Neighbor Line) using both artificial and real-world datasets demonstrate that RNFLS offers the best performance.

Keywords

Pattern classification, nearest feature line, rectified nearest feature line segment, distribution concentration, interpolation and extrapolation accuracy

1 Introduction

Nearest Feature Line (NFL) [1], a newly proposed nonparametric classifier, has received considerable attention in the pattern classification field. It attempts to enhance the representational capacity of a sample set of limited size by using the lines passing through

*Corresponding author. Tel/Fax: +86-21-65643842

each pair of the samples belonging to the same class. Simple yet effective, NFL shows good performance in many applications, including face recognition [1] [2], audio retrieval [3], image classification [4], speaker identification [5] and object recognition [6]. The authors of NFL explain that a feature line provides information about the possible linear variants of two sample points not covered by themselves.

On the other hand, the feature lines may produce detrimental effects that lead to increased decision errors under certain conditions, limiting its further potential. The authors of [7] pointed out one of the drawbacks – extrapolation inaccuracy, and proposed a solution called Nearest Neighbor Line (NNL). The extrapolation inaccuracy, as we will analyze in this paper, may cause serious problems for classification tasks in a low dimensional feature space, although it makes limited harm when the dimensions become high. Due to this shortcoming, NFL is not widely applicable to classification tasks in a low dimensional feature space. Thus, a more general method taking advantage of the powerful feature lines of NFL while avoiding this type of drawback is desirable.

Another drawback of NFL is interpolation inaccuracy. It arises when one class ω_1 has multiple clusters and between two of them appears the area of another class ω_2. Distributions of a complex shape (two-spiral problem for example) often fall into this category, where, by the original NFL, the interpolating parts of the feature lines of class ω_1 break up the area of class ω_2, severely damaging the decision region. A similar problem happens in *feature midpoints* [8], in which the generalized feature midpoint of two sample points of the same class may fall into the territory of other classes, leading to inappropriate decisions.

In this paper, a new nonparametric classification method, *Rectified Nearest Feature Line Segment* (RNFLS), is proposed that overcomes both of the above-mentioned drawbacks and significantly improves the performance of NFL. The original NFL can conceptually be viewed as a two-stage algorithm – building representational subspaces for each class and then performing the nearest distance classification. We focus mainly on the first

stage. To overcome extrapolation inaccuracy, *Nearest Feature Line Segment subspace* (NFLS-subspace) is developed. Different from NNL, the proposed solution still keeps the attractive NFL characteristic that linearly expanding the representational capacity of the limited sample points. To overcome the interpolation inaccuracy, the "territory" of each sample point and each class is defined, and we obtain *Rectified Nearest Feature Line Segment subspace*(RNFLS-subspace) from NFLS-subspace by eliminating those feature line segments trespassing the territory of other classes. As a result, RNFLS works well for all shapes of sample distributions, which is a significant improvement.

Another remarkable advantage of the RNFLS method is that the added line segment features make the initial sample distribution more concentrated. It is demonstrated in this paper that the concentration property significantly enhances the classification performance in overlapping areas of two or more classes. An example is the classification problem with two classes taking the same Gaussian density but different distribution centers (means). We show that the decision boundary created by RNFLS gets closer to the one built by using the optimal Bayesian rule, bringing the correct classification rate higher. Comparisons with NN, k-NN, NFL, NNL using artificial and real-world datasets demonstrate that the proposed RNFLS method offers remarkably superior performance.

The main contributions of this paper include,

- Pointing out and analyzing the advantages and drawbacks of the original NFL method.
- Evaluating the detrimental effects of extrapolation inaccuracy in feature space of low and high dimension.
- Proposing the RNFLS classifier to improve NFL by overcoming the drawbacks.
- Analyzing the concentration property of NFLS-subspace.

2 Background

2.1 The Nearest Feature Line Method

The Nearest Feature Line (NFL) [1] method constructs a feature subspace for each class, consisting of the straight lines passing through every pair of the samples belonging to that class. The straight line passing through samples x_i, x_j of the same class, denoted by $\overline{x_i x_j}$, is called a *feature line* of that class. Let $X^\omega = \{x_i^\omega | 1 \leq i \leq N_\omega\}$ be the set of N_ω samples belonging to class ω. A number $K_\omega = N_\omega(N_\omega - 1)/2$ of feature lines is then constructed and all these K_ω feature lines constitute the *NFL-subspace* to represent class ω, denoted by $\mathcal{S}_\omega = \{\overline{x_i^\omega x_j^\omega} | 1 \leq i, j \leq N_\omega, i \neq j\}$, which is a subset of the entire feature space. If there are m classes in the database, m NFL-subspaces will be constructed, containing a total number $N_{total} = \sum_{i=1}^{m} K_{\omega_i}$ of feature lines.

During classification, a query point q will be classified to class ω if q assumes the smallest distance to $\mathcal{S}_\omega$ than to any other $(m-1)$ NFL-subspaces. The distance from a feature point q to an NFL-subspace $\mathcal{S}_\omega$, $d(q, \mathcal{S}_\omega)$, is the shortest distance from q to any of the feature lines in $\mathcal{S}_\omega$

$$d(q, \mathcal{S}_\omega) = \min_{\overline{x_i x_j} \in \mathcal{S}_\omega} d(q, \overline{x_i x_j}). \tag{1}$$

The distance from a point q to an feature line $\overline{x_i x_j}$, $d(q, \overline{x_i x_j})$, is defined as

$$\begin{aligned} d(q, \overline{x_i x_j}) &= \min_{y \in \overline{x_i x_j}} \|q - y\| && (2) \\ &= \|q - p\| && (3) \end{aligned}$$

where y is a point in line $\overline{x_i x_j}$, $\|.\|$ is some norm and p is the projection point of q onto line $\overline{x_i x_j}$.

The projection point can be computed by

$$p = (1 - \mu)x_i + \mu x_j \tag{4}$$

where

$$\mu = \frac{(q - x_i).(x_j - x_i)}{(x_j - x_i).(x_j - x_i)}. \tag{5}$$

The parameter μ reflects the relative position of p to the two endpoints x_i and x_j. When $0 < \mu < 1$, p is an interpolation point between x_i and x_j. When $\mu > 1$, p is a "forward" extrapolation point on the x_j side. When $\mu < 0$, p is a "backward" extrapolation point on the x_i side.

Take Fig.1 as an example, there are two classes in the entire feature space. Samples x_1, x_2, x_3, marked with "circle", constitute class ω_1, and samples x_4, x_5, marked with "cross", constitute class ω_2. Solid straight lines show all the feature lines of the two classes. Point q is the query, and p_1, p_2, p_3, p_4 are the projection points of q on feature lines $\overline{x_1x_2}$, $\overline{x_2x_3}$, $\overline{x_1x_3}$, $\overline{x_4x_5}$ respectively. According to the NFL rule, the distance from q to $\mathcal{S}_{\omega_1}$ is $\min\{d(q,\overline{x_1x_2}), d(q,\overline{x_1x_3}), d(q,\overline{x_2x_3})\} = d(q,\overline{x_1x_3}) = \|q - p_3\|$, and the distance from q to $\mathcal{S}_{\omega_2}$ is $\min\{d(q,\overline{x_4x_5})\} = \|q - p_4\|$. Since $\|q - p_4\| \leq \|q - p_3\|$, q is classified to class ω_2. In addition, the projection points p_1, p_3 are on the extrapolating parts of the corresponding feature lines, while p_2, p_4 are on the interpolating parts.

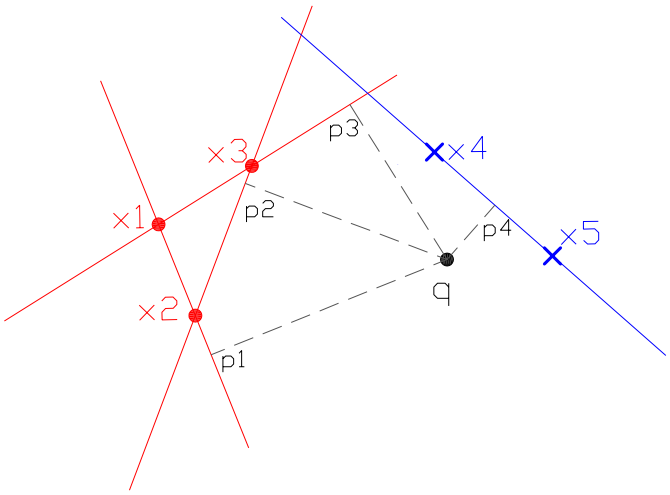


Figure 1: An example to illustrate the NFL classification method. The feature points p_1, p_2, p_3 and p_4 are the projections of query q on the feature lines $\overline{x_1x_2}$, $\overline{x_2x_3}$, $\overline{x_1x_3}$ and $\overline{x_4x_5}$ respectively.

2.2 Motivations and Advantages

The authors of [9] report that NFL gives better performance in pattern classification problems when the dimension of the feature space is high. This has been confirmed by several independent studies [1], [3], [4], [5] and [9].

One explanation provided by [1] is that a feature line virtually provides an infinite number of additional samples along the linear variance of a pair of sample points. When there is linear relationship among the sample points, NFL shows an impressive improvement.

Suppose that we have two classes of samples in a two-dimensional feature space as shown in Fig.2(a). One such situation is to separate a bag of mixed fruits when there are few samples(4 bananas and 4 apples). We have two features – the redness x, and the number of pockmarks on the surface y. According to general knowledge, the redness is very important for separating red apples from yellow bananas, while the number of pockmarks does not contribute useful information. However, we may assume that the two features are treated equally since the designer of the classifier has no prior knowledge about the different nature of the two attributes.

Naturally, we expect a classifier to produce a hyperplane (a line perpendicular to axis-x in this problem) separating the two classes. The NN rule, however, leads to decision boundary that is not favorable, as shown in Fig.2(b). Another choice would be NFL, in which the feature lines support the original samples as a complement, Fig.2(c), and the decision boundary constructed by NFL, as shown in Fig.2(d), precisely match what we expect, bringing superior classification results.

2.3 Shortcomings

NFL extends the samples of one class by adding straight lines linking each pair. A good argument for doing this is that it adds extra information to the sample set. The extra information, however, is a double-edged sword. When a straight line of one class trespasses

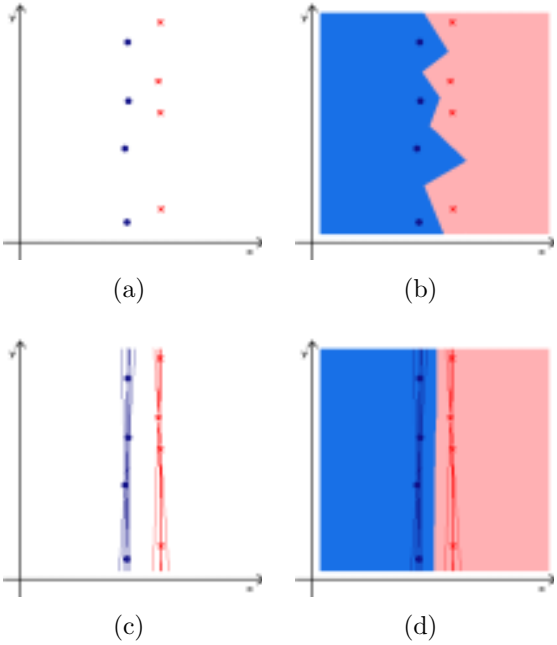


Figure 2: (a)The sample distribution in two-dimensional feature space. (b)Decision regions by NN rule. (c)The feature lines of each class constitute the NFL-subspace. (d)Decision regions by NFL-classifier.

into the territory of another class, it will lead to increased error probability. There are two types of trespassing, causing two types of inaccuracies: extrapolation inaccuracy and interpolation inaccuracy.

2.3.1 Extrapolation Inaccuracy

In Fig.3, the query point q surrounded by four "cross" sample points lies in the territory of the "cross" class, hoping to be classified into the "cross" class. But the extrapolating part of feature line $\overline{x_1x_2}$ makes the distance from q to $\overline{x_1x_2}$ smaller. Thus, $d(q, \mathcal{S}_{circle}) < d(q, \mathcal{S}_{cross})$, and NFL will assign q the label "circle", not "cross". This is very likely to be a decision error.

Further analysis of the extrapolation inaccuracy is presented in Section 4, where we will show that the probability of a feature line of one class trespassing the area of another class tends to zero when the feature space dimension is high. This means that extrapolation inaccuracy can be ignored if the feature space dimension is large enough. In a feature

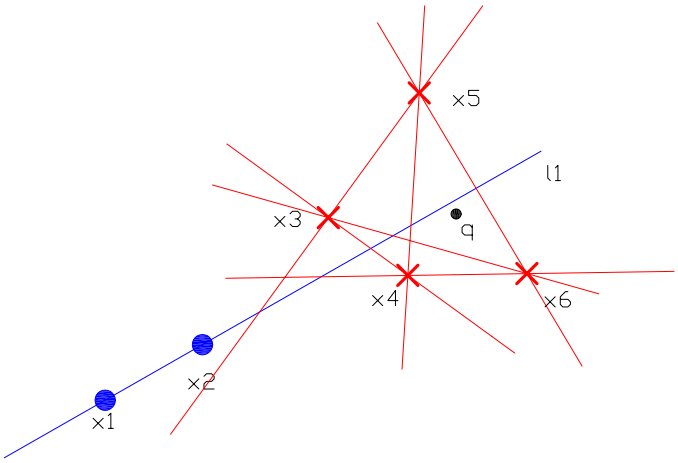


Figure 3: An example to show extrapolation inaccuracy. The query q surrounded by "cross"-samples is classified to "circle"-class because of the extrapolating part of the nearest feature line $\overline{x_1x_2}$. This is likely to be a decision error.

space of low dimension, however, it actually harms.

2.3.2 Interpolation Inaccuracy

Fig.4 shows an example to illustrate interpolation inaccuracy. We note that the samples from class "circle" distribute as two clusters with the samples from class "cross" lying in between, and feature lines such as $\overline{x_1x_2}$ linking the two clusters of class "circle" trespass the territory of class "cross". The query q in the territory of class "cross" which should have been labelled as "cross" is inappropriately classified to class "circle" by the NFL rule because $d(q, \overline{x_1x_2})$ is smaller than the distance from q to any feature lines of class "cross". This type of error occurrs because the territory of class "cross" is trespassed by the interpolating part of feature lines of class "circle".

The extrapolation and interpolation inaccuracies are serious drawbacks that limit the applicability of NFL. The authors of [7] tried to reduce the extrapolation inaccuracy by using Nearest Neighbor Line (NNL), where only one feature line linking the nearest two sample points to the query is used to represent the class. This approach, while mitigating the extrapolation inaccuracy, also reduces the classification ability of the original NFL.

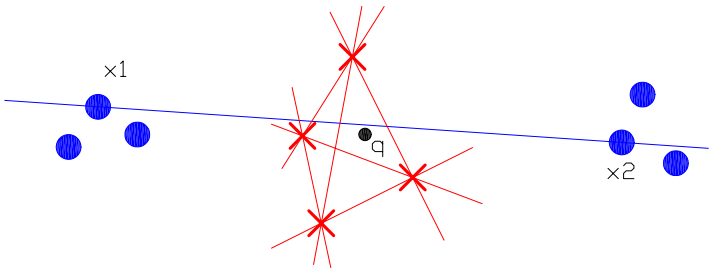


Figure 4: An example to show the interpolation inaccuracy. The query q surrounded by "cross"-samples is classified to "circle"-class because of the interpolating part of feature line $\overline{x_1x_2}$, which probably causes a decision error.

In the following section we pursue a more systematic approach in which a new feature subspace for each class is constructed to avoid both drawbacks. The original advantage of NFL that linearly extending the representational capacity of the original samples is retained in our method.

3 Rectified Nearest Feature Line Segment

In this section, we introduce the new method - *Rectified Nearest Feature Line Segment* (RNFLS) - to improve NFL. The main difference between the two lies in the way of constructing the feature subspace for each class. In the original NFL [1], the NFL-subspace for class ω_k is

$$\mathcal{S}_{\omega_k} = \{\overline{x_i x_j} | x_i, x_j \in \omega_k, x_i \neq x_j\}. \tag{6}$$

To overcome the drawbacks described in the previous section, a two-step improvement is developed, as shown in Fig. 5. Not only has it avoided the shortcomings, it also possesses a remarkable property – concentration property – which generates new RNFLS features (line segments), and consequently makes the resultant distribution more concentrated

than the distribution of the original samples. This property significantly enhances the classification ability for a wide range of applications.

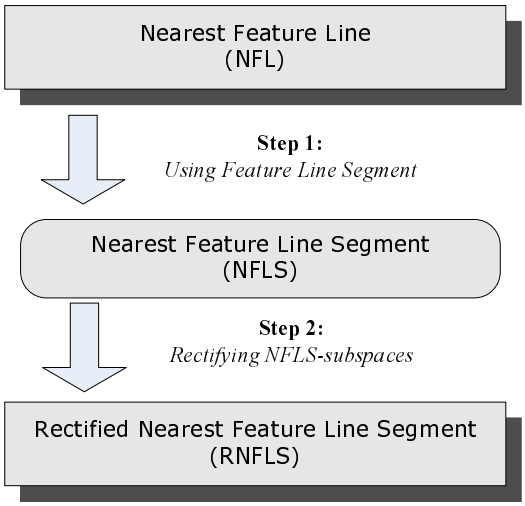


Figure 5: RNFLS can be considered as augmented NFL with two improvement steps.

3.1 Using Feature Line Segment

To avoid extrapolation inaccuracy, we use line segments between pairs of the sample points to construct a *Nearest Feature Line Segment subspace* (NFLS-subspace) instead of the original NFL-subspace to represent each class. Let $X^\omega = \{x_i^\omega | 1 \le i \le N_\omega\}$ be the set of N_ω samples belonging to class ω. The NFLS-subspace ($\widetilde{\mathcal{S}}_\omega$) representing class ω is

$$\widetilde{\mathcal{S}}_\omega = \{\widetilde{x_i^\omega x_j^\omega} | 1 \le i, j \le N_\omega\}, \tag{7}$$

where $\widetilde{x_i^\omega x_j^\omega}$ denotes the line segment connecting the point x_i^ω and x_j^ω. Note that a degenerative line segment $\widetilde{x_i^\omega x_i^\omega}$ $(1 \le i \le N_\omega)$, representing the original sample point x_i^ω, is also a member of $\widetilde{\mathcal{S}}_\omega$.

If we consider each line segment as a set of feature points, $\widetilde{\mathcal{S}}_\omega$ can also be viewed as a point set. Thus the distance from a query point q to an NFLS-subspace $\widetilde{\mathcal{S}}_\omega$ is defined as

$$d(q, \widetilde{\mathcal{S}}_\omega) = \min_{y \in \widetilde{\mathcal{S}}_\omega} \|q - y\|, \tag{8}$$

where $\|\cdot\|$ is some norm and $y \in \widetilde{\mathcal{S}}_\omega \subset \Re^n$ is a point. The query q is classified to class ω_k when $d(q, \widetilde{\mathcal{S}}_{\omega_k})$ is smaller than the distance from q to any other $\widetilde{\mathcal{S}}_{\omega_i}$ $(i \neq k, 1 \leq i \leq m)$.

Since $\widetilde{\mathcal{S}}_{\omega_k}$ is made up of line segments, Equ.(8) is equivalent to

$$d(q, \widetilde{\mathcal{S}}_{\omega_k}) = \min_{\widetilde{x_i x_j} \in \widetilde{\mathcal{S}}_{\omega_k}} d(q, \widetilde{x_i x_j}) \tag{9}$$

where

$$d(q, \widetilde{x_i x_j}) = \min_{y \in \widetilde{x_i x_j}} \|q - y\|. \tag{10}$$

And to calculate $d(q, \widetilde{x_i x_j})$, there are two cases. If $x_i = x_j$, the answer is simply the point to point distance,

$$d(q, \widetilde{x_i x_i}) = \|q - x_i\|. \tag{11}$$

Otherwise, the projection point p of q onto $\overline{x_i x_j}$ is located first by using Equ.(4) and Equ.(5). Then, different reference points are chosen to calculate $d(q, \widetilde{x_i x_j})$ according to the position parameter μ. When $0 < \mu < 1$, p is an interpolation point between x_i and x_j, so $d(q, \widetilde{x_i x_j}) = \|q - p\|$. When $\mu < 0$, p is a "backward" extrapolation point on the x_i side, so $d(q, \widetilde{x_i x_j}) = \|q - x_i\|$. When $\mu > 1$, p is a "forward" extrapolation point on the x_j side, so $d(q, \widetilde{x_i x_j}) = \|q - x_j\|$. Fig.6 shows an example.

3.2 Rectifying Nearest Feature Line Segment Subspaces

The next step is to rectify the NFLS-subspace to eliminate interpolation inaccuracy. Our motivation is to have the inappropriate line segments removed from the NFLS-subspace $\widetilde{\mathcal{S}}_{\omega_k}$ for each class ω_k,. The resulting subspace denoted by $\widetilde{\mathcal{S}}^*_{\omega_k}$ is a subset of $\widetilde{\mathcal{S}}_{\omega_k}$ termed *Rectified Nearest Feature Line Segment subspace* (RNFLS-subspace).

3.2.1 Territory

We begin with the definitions of two types of territories. One is *sample-territory*, $T_x \subseteq \Re^n$, that is the territory of a sample point x; the other is *class-territory*, $\mathcal{T}_\omega \subseteq \Re^n$, that is the territory of class ω. Suppose the sample set X is $\{(x_1, \theta_1), (x_2, \theta_2), ..., (x_m, \theta_m)\}$, which

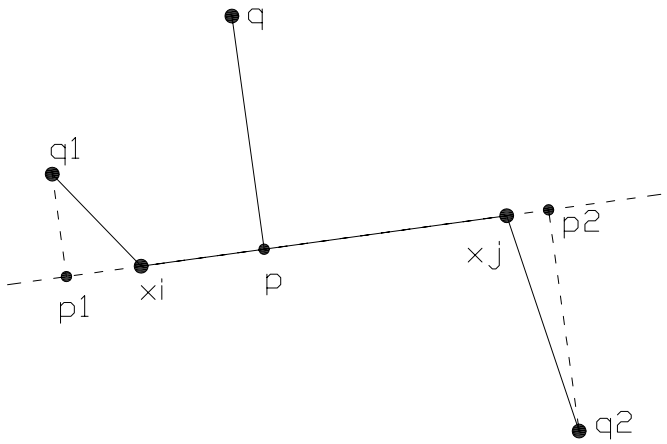


Figure 6: Distance from a feature point to a feature line segment. The projection of the query point q_1 on $\widetilde{x_i x_j}$ is in the extrapolating part, so the nearest endpoint x_i is chosen to be the reference point for calculating the distance. The process is similar for q_2. For query q, $d(q, \widetilde{x_i x_j})$ is the same as the distance in NFL since its projection is in the interpolating part of $\widetilde{x_i x_j}$.

means x_i belongs to class θ_i. The *radius* r_{x_k} of the sample-territory T_{x_k} is defined as

$$r_{x_k} = \min_{\forall x_i, \theta_i \neq \theta_k} \|x_i - x_k\|. \tag{12}$$

Thus,

$$T_{x_k} = \{y \in \Re^n | \|y - x_k\| < r_{x_k}\}. \tag{13}$$

It is not required that $T_{x_i} (1 \leq i \leq m)$ is a partition of $\Re^n$. That is to say $T_{x_i} \bigcap T_{x_j} (x_i \neq x_j)$ may not be empty, and $\bigcup_{1 \leq i \leq m} T_{x_i}$ may not be the whole feature space $\Re^n$.

The class-territory $\mathcal{T}_{\omega_k}$ is defined to be

$$\mathcal{T}_{\omega_k} = \bigcup_{\theta_i = \omega_k} T_{x_i}, \quad (x_i, \theta_i) \in X. \tag{14}$$

In Fig.7, the points denoted by "circle" and "cross" represent the samples from two classes. Each of the "cross"-points (y_1, y_2, y_3) has its own sample-territory as shown by the dashed circle. The union of these sample-territories is $\mathcal{T}_{cross}$. $\mathcal{T}_{circle}$ is obtained in a similar way.

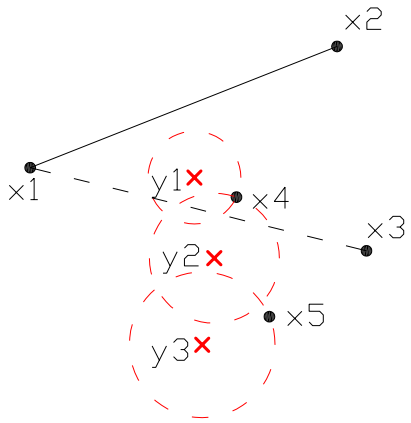


Figure 7: The territory of "cross"-samples is shown in dashed circle. Their union constitutes the territory of "cross"-class. The feature line segment $\widetilde{x_1x_3}$ of "circle"-class trespassing the territory of "cross"-class will be removed, while $\widetilde{x_1x_2}$ will not.

3.2.2 Building RNFLS-subspace

For class ω_k, its RNFLS-subspace $\widetilde{\mathcal{S}}^*_{\omega_k}$ is built from the NFLS-subspace $\widetilde{\mathcal{S}}_{\omega_k}$ by having those line segments trespassing the class-territories of other classes removed. That is,

$$\widetilde{\mathcal{S}}^*_{\omega_k} = \widetilde{\mathcal{S}}_{\omega_k} \setminus \widetilde{U}_{\omega_k}, \tag{15}$$

where '\' is the set minus operator, and

$$\begin{aligned} \widetilde{U}_{\omega_k} &= \{\widetilde{x_ix_j} | \exists \omega_y, \omega_k \neq \omega_y \wedge \\ &\quad \wedge \widetilde{x_ix_j} \in \widetilde{\mathcal{S}}_{\omega_k} \wedge \widetilde{x_ix_j} \cap \mathcal{T}_{\omega_y} \neq \phi\} \\ &= \{\widetilde{x_ix_j} | \exists (x_y, \theta_y) \in X, \widetilde{x_ix_j} \in \widetilde{\mathcal{S}}_{\omega_k} \wedge \\ &\quad \wedge \omega_k \neq \theta_y \wedge d(x_y, \widetilde{x_ix_j}) < r_{x_y}\}. \end{aligned} \tag{16}$$

For example, in Fig.7, $\widetilde{x_1x_3} \notin \widetilde{\mathcal{S}}^*_{circle}$ because $\widetilde{x_1x_3} \cap \mathcal{T}_{cross} \neq \phi$. $\widetilde{x_1x_2} \in \widetilde{\mathcal{S}}^*_{circle}$ because $\widetilde{x_1x_3} \cap \mathcal{T}_{cross} = \phi$ and class "cross" is the only class different from class "circle".

3.2.3 Classifying using RNFLS-subspaces

To perform classification using RNFLS-subspaces is similar to using NFLS-subspaces, since the only difference between an RNFLS-subspace and an NFLS-subspace is $\widetilde{\mathcal{S}}^*_{\omega_k} =$

$\widetilde{\mathcal{S}}_{\omega_k} \setminus \widetilde{U}_{\omega_k}$, where, except for some removed line segments, $\widetilde{\mathcal{S}}^*_{\omega_k}$ is still a set consisting of line segments. The distance measure from a query point to the RNFLS-subspace remains the same.

4 Analyzing the Method

In this section, we analyze the proposed RNFLS method from two important viewpoints - *concentration property* and *the correlation between trespassing probability and the feature space dimension*. The concentration property is a remarkable advantage of RNFLS, which significantly enhances the classification performance by generating new RNFLS features that are more concentrated in distribution than the original samples of the corresponding class. The second property shows the necessity for classification problems in low dimensional feature space to eliminate the extrapolation and interpolation inaccuracies of NFL by the two-step improvement of RNFLS, leading to better classification performance.

4.1 Analyzing the Concentration Property

In many real-world pattern recognition problems, samples from one class tend to scatter around its central manifold due to systematic deviation and random noise. Two scattered classes may overlap each other, causing decision errors. A method which is able to generate features with a more concentrated distribution than the distribution of original samples may improve the classification performance in the overlapping areas, leading to a higher correct classification rate.

In the following analysis, we show the impressive concentration property of the proposed RNFLS method using a simple case - a uniform distribution in two-dimensional feature space. A further simplification is to show it in NFLS-subspace.

Proposition 1. *Consider the NFLS-subspace of the class ω as shown in Fig.8, where the sample points of class ω are uniformly distributed in disk D with radius R and center O. Let $M(a, r)$ $(a \leq R)$ be a round area with an arbitrarily small radius r and distance a*

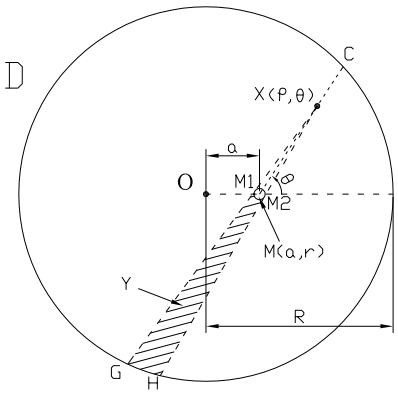


Figure 8: Calculating N_a^ω for a uniform sample point density on a disk.

from O, and N_a^ω be the probability of a randomly chosen feature line segment of class ω passing through $M(a,r)$. Given an arbitrarily small r, N_a^ω is decreasing with increasing a.

Proof. We calculate N_a^ω in a polar coordinate system by choosing the center of $M(a,r)$ as pole and $\overrightarrow{OM}$ as polar axis. For a line segment $\widetilde{XY}$ passing through $M(a,r)$, given one endpoint $X(\rho,\theta)$ in D, the other endpoint Y has to appear in the corresponding $\square_{M_1M_2HG}$, as shown in Fig.8. Thus we obtain

$$\begin{aligned} N_a^\omega &= \iint_D \frac{1}{\pi R^2} A(\rho,\theta)\rho d\rho d\theta \\ &= \int_0^{2\pi}\int_0^{|MC|} \frac{1}{\pi R^2} A(\rho,\theta)\rho d\rho d\theta \end{aligned} \tag{17}$$

where $A(\rho,\theta)$ is the probability that the randomly generated endpoint Y appears in $\square_{M_1M_2HG}$.

We can calculate $A(\rho,\theta)$, $|MC|$, $|MG|$ and $|GH|$ by

$$A(\rho,\theta) = \frac{1}{\pi R^2}\Big[\frac{1}{2}(2r+|GH|)\cdot|MG| + o(r)\Big] \tag{18}$$

$$|MC| = \sqrt{R^2 - a^2\sin^2\theta} - a\cos\theta \tag{19}$$

$$|MG| = \sqrt{R^2 - a^2\sin^2\theta} + a\cos\theta \tag{20}$$

$$|GH| = \frac{2r(|MG|+\rho)}{\rho} \tag{21}$$

According to Equ.(17), (18), (19), (20) and (21),

$$N_a^\omega = \frac{2r(R^2-a^2)}{(\pi R^2)^2}\int_0^{2\pi}\sqrt{R^2 - a^2\sin^2\theta}\cdot d\theta + o(r). \tag{22}$$

Since $r \to 0$, $o(r)$ is ignored. Thus, for a fixed r, N_a^ω gets smaller when a gets larger. □

Proposition 1 indicates that the distribution of line segments in the NFLS-subspace is denser at the center than at the boundary if the original sample points distribution 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.

Consider a given location whose distance to the center is a. The relationship between $\rho_\omega(a)$ and N_a^ω can be expressed as

$$N_a^\omega = k(a)\cdot\rho_\omega(a), \tag{23}$$

$k(a)$ is a positive valued function decreasing on a.

Proposition 2. *Following the above analysis, we duplicate m copies of class ω with different class centers μ_i, denoted by $\omega_1, \omega_2, ..., \omega_m$. In the m-class classification problem, given a small round area M in the overlapping area of these m classes, the probability of a randomly chosen query x appeared in M being classified to class ω_k, $P_{NFLS}(\omega_k|x)$, is*

$$P_{NFLS}(\omega_k|x) = \frac{N_{a_k}^{\omega_k}}{\sum_{i=1}^m N_{a_i}^{\omega_i}} \tag{24}$$

where $a_i = \|x-\mu_i\|$ is the distance from x to the distribution center of ω_i.

Proof. The location of each feature line segment of these m classes passing through M is random, therefore, $P_{NFLS}(\omega_k|x)$ is determined by the portions of feature line segments of class $\omega_i (i = 1, 2, ..., m)$ that pass through M. On the other hand, since the m classes possess an equal prior probability and $N_{a_k}^{\omega_k}$ represents the probability of a randomly chosen feature line segment of class ω_k passing through M, the proposition is obvious. □

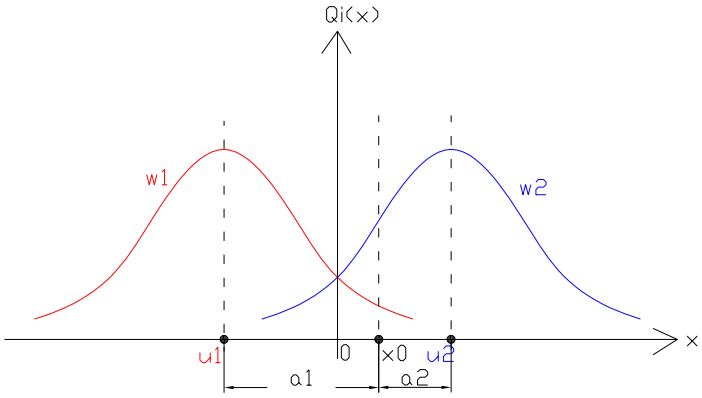


Figure 9: Two-category problem of the same Gaussian distribution. After the concentration of the probability density, the conditional risk $\mathcal{R}_{NFL}$ will be smaller than $\mathcal{R}_{NN}$.

To show the benefits of density concentration, suppose there are two Gaussian distributions of the same variance but different centers (means) in two-dimensional feature space. Projected on the straight line linking the two Gaussian centers, the probability density is shown in Fig.9, where σ is the variance of the projected one dimensional Gaussian density, and μ_1, μ_2 are the projected centers, satisfying

$$\mu_1 + \mu_2 = 0 \quad (\mu_1 < 0). \tag{25}$$

Consider a query point x_0 ($x_0 > 0$). For i=1,2, let

$$Q_i(x) = P\{x \in \omega_i|x\} \tag{26}$$

$$a_i = \|x_0 - \mu_i\| \tag{27}$$

$$\rho_{\omega_i}(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\{-\frac{1}{2}(\frac{x}{\sigma})^2\}. \tag{28}$$

We obtain

$$Q_i(x_0) = \frac{\rho_{\omega_i}(a_i)}{\rho_{\omega_1}(a_1) + \rho_{\omega_2}(a_2)} \qquad (i = 1, 2) \tag{29}$$

$$a_1 > a_2. \tag{30}$$

On the other hand, consider a small round area M with radius r centered at x_0. Let $N_{a_i}^{\omega_i}$ be the probability of a randomly chosen feature line segment of class ω_i passing through M $(i = 1, 2)$. If we denote them as Equ.(23),

$$N_{a_1}^{\omega_1} = k_1 \cdot \rho_{\omega_1}(a_1) \tag{31}$$

$$N_{a_2}^{\omega_2} = k_2 \cdot \rho_{\omega_2}(a_2), \tag{32}$$

from Equ.(23), (28), (30), (31) and (32),

$$k_1 < k_2. \tag{33}$$

According to the NN rule, a query x_0 is classified to class ω if the nearest sample point x_0' belongs to ω. An error occurs when $x_0 \in \omega_1$ but $x_0' \in \omega_2$ or $x_0 \in \omega_2$ but $x_0' \in \omega_1$. If the sample distribution is not sparse, $P\{x_0' \in \omega | x_0'\}$ can be approximately replaced by $P\{x_0 \in \omega | x_0\}$. Therefore, the conditional risk $\mathcal{R}_{NN}$ given x_0 is

$$\begin{aligned} \mathcal{R}_{NN}(x_0) &= P\{x_0 \in \omega_1 \wedge x_0' \in \omega_2\} \vee \\ &\quad P\{x_0 \in \omega_2 \wedge x_0' \in \omega_1\} \\ &= 2Q_1(x_0)Q_2(x_0). \end{aligned} \tag{34}$$

Let $C(x_0)$ be the classification result of query x_0 by NFLS,

$$\begin{aligned} \mathcal{R}_{NFLS}(x_0) &= P\{x_0 \in \omega_1 \wedge C(x_0) = \omega_2\} \vee \\ &\quad P\{x_0 \in \omega_2 \wedge C(x_0) = \omega_1\} \\ &= Q_1(x_0)P_{NFLS}(\omega_2|x_0) + \\ &\quad Q_2(x_0)P_{NFLS}(\omega_1|x_0). \end{aligned} \tag{35}$$

According to Equ.(24), (28), (29), (31), (32), (33), (34) and (35), we conclude

$$\mathcal{R}_{NFLS}(x_0) < \mathcal{R}_{NN}(x_0). \tag{36}$$

This 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.

4.2 Trespassing Probability and Feature Space Dimension

The following example gives the result that the extrapolation inaccuracy of NFL makes limited harm in high dimensional feature space but causes severe problems in a feature space of low dimension. Thus, it is necessary to design solutions such as RNFLS to eliminate these inaccuracies for its applicability on classification problems in low dimension.

Consider, in an n-dimensional feature space, a two-category problem in which $p(x|\omega_1)$ and $p(x|\omega_2)$ are the prior densities of the two classes in region D_1 and D_2 respectively $(D_1, D_2 \subseteq \Re^n)$. To randomly generate a feature line l from the NFL-subspace of class ω_1 is equivalent to firstly ordering all the sample points of class ω_1, so that they form a chain, and then randomly picking up two sample points x_a, x_b from class ω_1 ensuring $x_a \prec x_b$ (if not, do it again), and then to obtain l by linking x_a and x_b. The feature line l may or may not trespass D_2. The probability, P_n, that l trespasses D_2 in the n-dimensional problem is

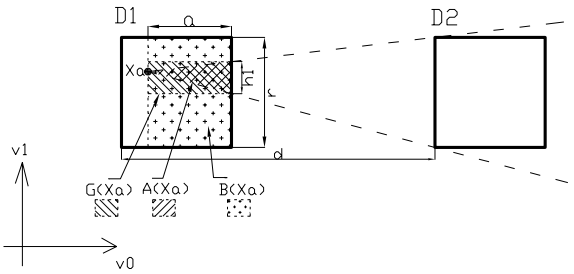
$$P_n = \int_{D_1} p(x_a|\omega_1) \cdot \frac{\int_{A(x_a)} p(x_b|\omega_1) dx_b}{\int_{B(x_a)} p(x_b|\omega_1) dx_b} \cdot dx_a \tag{37}$$

where $A(x_a)$ and $B(x_a)$ are

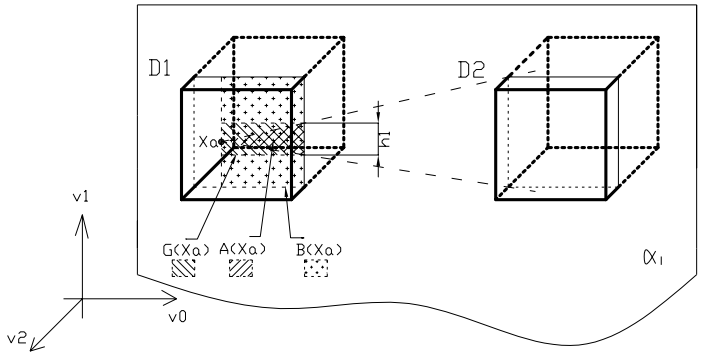
$$A(x_a) = \{x_b | x_b \in D_1 \wedge x_a \prec x_b \wedge \overline{x_a x_b} \cap D_2 \neq \phi\} \tag{38}$$

$$B(x_a) = \{x_b | x_b \in D_1 \wedge x_a \prec x_b\}. \tag{39}$$

An important property is that P_n will asymptotically come to 0 when n grows sufficiently large. The following analysis explains this property.



(a)



(b)

Figure 10: The two-cube problem to estimate the upperbound of the extrapolation inaccuracy. (a)A two-dimensional case. (b)The same model in three-dimensional feature space.

Suppose we have a non-overlapping two-category problem [10] in an n-dimensional feature space, where both $p(x|\omega_1)$ and $p(x|\omega_2)$ assume uniform densities on hypercube D_1 and D_2. The side length of each cube is r, all the edges are parallel to one of the coordinate axis, and D_2 is considered as a shifted version of D_1 along coordinate axis v_0 with distance d, satisfying

$$d > r. \tag{40}$$

We give the sequence number to the sample points in D_1 by sorting them using their geometrical positions, that is to specify the projection of one sample point on v_0, v_1 as the first and second sort key respectively. Fig.10(a) shows a two-dimensional case. Thus,

$$p(x|\omega_1) = 1/\mathcal{V}^{\S}[D_1], \quad for\ x \in D_1 \tag{41}$$

$$P_n = \frac{1}{\mathcal{V}[D_1]} \int_{D_1} \frac{\mathcal{V}[A(x_a)]}{\mathcal{V}[B(x_a)]} \cdot dx_a. \tag{42}$$

To calculate the upper-bound of $\mathcal{V}[A(x_a)]/\mathcal{V}[B(x_a)]$, a small bounding box will be constructed. The process begins with $n = 2$, where region $A(x_a)$ is a triangle, as shown in Fig.10(a). The corresponding bounding box is a rectangle denoted by $G(x_a)$. In the two-dimensional space,

$$\mathcal{V}[G(x_a)] = a \cdot h_1 \quad , \quad x_a \in \Re^2. \tag{43}$$

We see that $A(x_a)$ is in $G(x_a)$, so

$$\frac{\mathcal{V}[A(x_a)]}{\mathcal{V}[B(x_a)]} < \frac{\mathcal{V}[G(x_a)]}{\mathcal{V}[B(x_a)]}. \tag{44}$$

When $n = 3$, consider a plane α_1 containing x_a, parallel to coordinate axis v_0, v_1 and perpendicular to v_2, shown in Fig.10(b). If a feature line l of class ω_1 trespasses D_2, the projection of l on plane α_1 must trespass the projection of D_2 on plane α_1. So when x_a is given, the projection of $A(x_a)$ in plane α_1 is in a triangle region, and a rectangle $G_1(x_a)$ whose edges are a and h_1 is large enough to cover it. Similarly, in the other degree of freedom, consider plane α_2 parallel to coordinate axis v_0, v_2 and perpendicular to v_1. Another rectangle $G_2(x_a)$ whose edges are a and h_2 is produced. After that, we construct the three dimensional bounding box $G(x_a)$ as the region, which is precisely the largest union of points in R^3 whose projections on the above two planes are in $G_1(x_a)$ and $G_2(x_a)$. So Equ.(44) is also satisfied in three dimensional case, where

$$\mathcal{V}[G(x_a)] = a \cdot h_1 \cdot h_2 \quad , \quad x_a \in \Re^3. \tag{45}$$

[§] $\mathcal{V}[\Psi]$ is the volume of region Ψ

In n-dimensional case, an n-dimensional bounding box $G(x_a)$ is built, and

$$\mathcal{V}[G(x_a)] = a \cdot \prod_{i=1}^{n-1} h_i \quad , \quad x_a \in \Re^n. \tag{46}$$

From Fig.10(a), we obtain

$$h_i \leq \frac{r^2}{d}, \quad (1 \leq i < n). \tag{47}$$

Since

$$\mathcal{V}[B(x_a)] = a \cdot r^{n-1}, \tag{48}$$

by Equ.(46), (47) and (48),

$$\frac{\mathcal{V}[G(x_a)]}{\mathcal{V}[B(x_a)]} = (\frac{r}{d})^{n-1}. \tag{49}$$

According to Equ.(40), (42), (44) and (49), we have

$$\lim_{n \to \infty} P_n = 0. \tag{50}$$

As a concrete example, let $p(x|\omega_1)$ be the uniform density function on hypercube D_1, and consider $d = 2r, 3r, 4r, 8r$ respectively. The trespassing probability P_n of a feature line in n-dimension problem is shown in Fig.11. Since the probability drops sharply with the dimensionality, it is reasonable to ignore the extrapolation inaccuracy if the classification task is in high dimension.

When the above two-category-problem is in a space of low dimensionality, the expected number of feature lines of class ω_1 trespassing D_2, however, will be comparable with the number of feature lines of class ω_2 so that the area of class ω_2 will be carved up by large amounts of feature lines of both class ω_1 and ω_2. Thus, a query that appears in D_2 which should have been classified to class ω_2 will have a considerable probability to be mistakenly labelled as class ω_1. This is a significant shortcoming because many other classifiers including Nearest Neighbor (NN) can easily achieve the correct classification rate of 100% for this non-overlapping distribution.

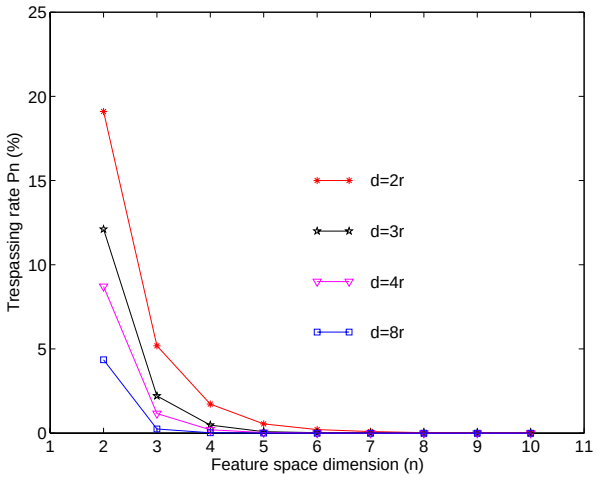


Figure 11: The extrapolation trespassing probability P_n, that is the probability of a feature line of class ω_1 trespassing the territory of class ω_2, is decreasing on the feature space dimension n.

4.3 Computational Complexity

For a given sample size n, the number of feature line segments in *RNFLS-subspace* is upper bounded by n^2. Thus, the proposed RNFLS method takes the same computation complexity of $O(n^2)$ as NFL for classifying one query, and in addition, since some of feature line segments are removed in the rectification process as described in Section 3.2, RNFLS usually works faster than NFL. On the other hand, the advantages of RNFLS in rectifying the representational subspace and concentrating the distribution come at a cost. The expense is the longer computation time-complexity $O(n^3)$ in preparing the *RNFLS-subspace*. However, preparing RNFLS-subspace is typically an offline, one-time operation, the increased time complexity is not a big issue.

5 Experiment Results and Discussions

The performance of the RNFLS method is compared with four classifiers - NN, k-NN, NFL and NNL - using two artificial datasets as well as a group of real-world benchmarks widely used to evaluate classifiers. The results on these datasets, representing various distributions and different dimensions, demonstrate that RNFLS possesses remarkably

stronger classification ability than the other methods.

5.1 The Two-Spiral Problem

The two-spiral problem was originally used to test multi-layered neural classifiers [11], [12], [13], and is now included by many authors as one of the benchmarks for evaluation of new classification algorithms. The two-spiral curves in a two-dimensional feature space is described as follows

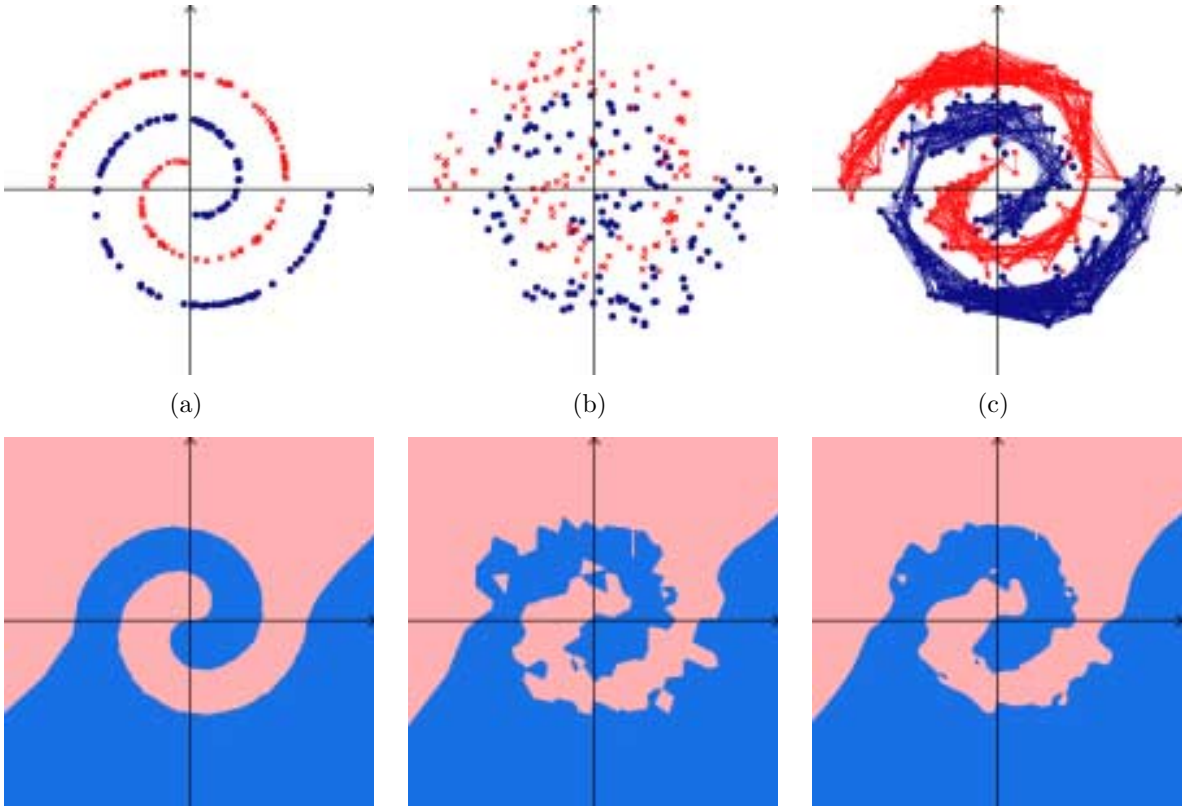
$$\begin{aligned} spiral1 &: \begin{cases} x = k\theta\cos(\theta) \\ y = k\theta\sin(\theta) \end{cases} \\ spiral2 &: \begin{cases} x = k\theta\cos(\theta+\pi) \\ y = k\theta\sin(\theta+\pi) \end{cases} \end{aligned} \tag{51}$$

where $\theta \geq \pi/2$ is the parameter. If the probability density of each class is uniform along the corresponding curve, an instance of such distribution is shown in Fig.12(a).

In our experiment, Gaussian noise is added to the samples so that the distribution regions of the two classes may overlap each other, as shown in Fig.12(b). If the prior distribution density were known, according to the optimal Bayesian rule, Fig.12(d) should be the optimal decision boundary. This, however, can hardly be achieved because the only information we have is a finite number of sample points.

The original NFL is not a good choice for this classification problem. We may imagine how fragmented the decision region is carved up because of its interpolation and extrapolation inaccuracy. The decision boundary created by NN rule is shown in Fig.12(e). When it comes to RNFLS, Fig.12(c) is the RNFLS-subspaces and Fig.12(f) is the corresponding decision boundaries. Compared with the decision boundary created by NN, RNFLS performs much better where the boundary is smoother and some incorrect regions caused by isolated noise points is smaller. This significant enhancement can be attributed to the concentration property.

As a concrete test, let $\theta \in [\pi/2, 3\pi]$ and the Gaussian noise is of a variance $\sigma = 1.7$ and an expectation $\mu = 0$. We generate 500 points according to this distribution, where 250 belong to class ω_1 and the other 250 belong to class ω_2. Then, half of them are



(a) (b) (c)

(d) (e) (f)

Figure 12: (a)Two-spiral problem. (b)Two-spiral problem with Gaussian noise. (c)RNFLS subspaces. (d)Bayesian decision boundary. (e)NN classification result. (f)RNFLS classification result.

randomly chosen to form the sample set and the remaining half constitute the test set. The classifiers, NN, k-NN(k=3), NFL, NNL and RNFLS, are applied to this task for 10 times, and Table 1 shows the results.

Table 1: Performance evaluation on the two-spiral problem using NN, 3-NN, NFL[1], NNL[7] and RNFLS. (CCR: correct classification rate, percentage)

Classifier	CCR (average)	CCR (min)	CCR(max)
NN	83.2	80.4	85.3
k-NN(k=3)	85.3	83.2	87.3
NFL	53.2	49.8	56.7
NNL	72.4	69.0	78.0
RNFLS	**86.1**	**84.0**	**88.2**

Table 2: CCR(%) for NN, 3-NN, NFL, NNL and RNFLS on the real-world datasets

	Dataset	#Classes	#Instances	#Attributes	NN	3NN	NFL	NNL	RNFLS
1	hepatitis	2	80	19	**92.5**	91.3	91.3	76.3	91.3
2	iris	3	150	4	94.7	94.7	88.7	94.7	**95.3**
3	housing	6	506	13	70.8	73.0	71.1	67.6	**73.5**
4	pima	2	768	8	70.6	**73.6**	67.1	62.8	73.0
5	wine	3	178	13	95.5	95.5	92.7	78.7	**97.2**
6	bupa	2	345	6	63.2	65.2	63.5	57.4	**66.4**
7	ionosphere	2	351	34	86.3	84.6	85.2	87.2	**94.3**
8	wpbc	2	194	32	72.7	68.6	72.7	54.1	**75.8**
9	wdbc	2	569	30	95.1	96.5	95.3	64.0	**97.2**
10	glass	6	214	9	70.1	72.0	66.8	65.4	**72.0**

5.2 Two Gaussian Distributions

Of the various distributions that have been investigated in the literature, Gaussian density has received great attention. Suppose that there are two classes ω_1 and ω_2 in a two-dimensional feature space according to the density functions,

$$p_{\omega_1}(x,y) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma(b-a)} exp\Big[-\frac{1}{2}\Big(\frac{x}{\sigma}\Big)^2 \Big] & , \quad a \le y \le b \\ 0 & , \quad otherwise \end{cases}$$

$$p_{\omega_2}(x,y) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma(b-a)} exp\Big[-\frac{1}{2}\Big(\frac{x-d}{\sigma}\Big)^2 \Big] & , \quad a \le y \le b \\ 0 & , \quad otherwise \end{cases} \tag{52}$$

In the experiment, given the distance d between the two Gaussian centers, 200 points are generated for each of the two classes based on the above density function. The sample set is formed by randomly selecting out half of the points from each class, and the test set contains the remaining half. Experiments with different d are practiced, and the result is shown in Fig.13. It shows that RNFLS always reaches the top performance among the five classifiers, NN, 3NN, NFL, NNL, RNFLS, and it achieves a correct classification rate(CCR) closest to the optimal Bayesian rule.

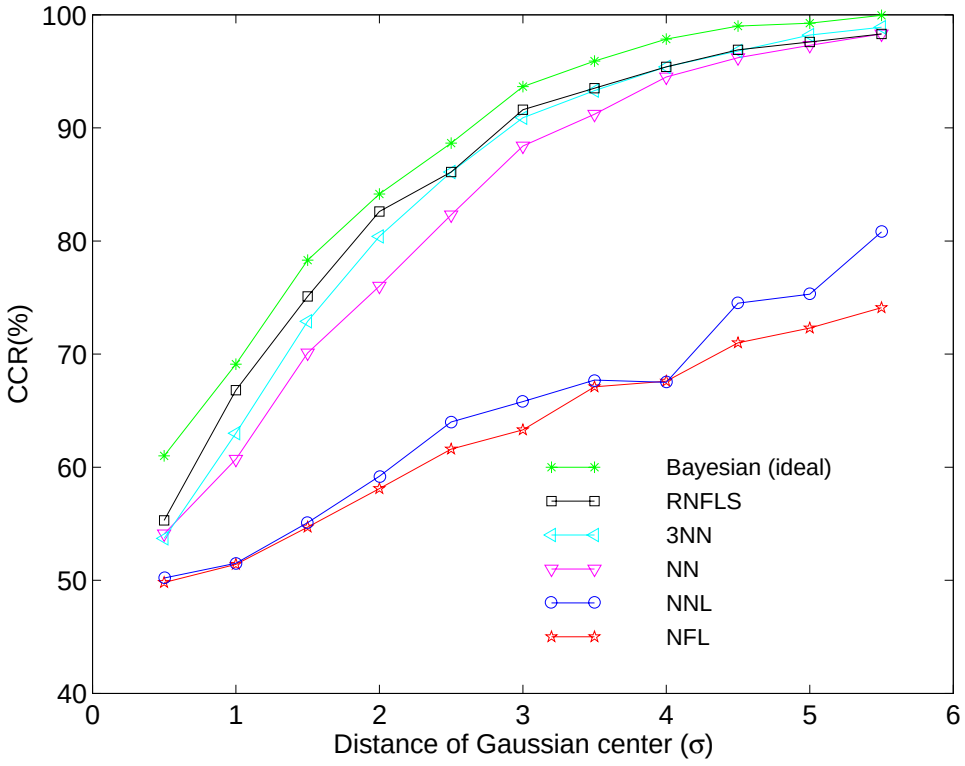


Figure 13: The CCR of the two-Gaussian distribution in two-dimensional feature space. The test is run using different classification methods under different distance of the two Gaussian centers.

The above experiment results show that the original NFL and NNL methods are not suitable in this situation. From the analysis of the extrapolation inaccuracy in Section 2 and 4, it is seen that the main reason lies in the fact of the large areas of the territory of one class being carved up by the extrapolating part of feature lines of other classes in a low dimensional feature space. When it comes to higher dimensions, as confirmed by

related work [9], this type of error is greatly mitigated.

5.3 Real-World Classification Problems

We test the RNFLS classifier on a group of real-world datasets as listed in Table 2. All of the datasets are obtained from the U.C. Irvine repository [14]. Since we do not deal with the issue of missing data, instances with missing values are removed. And for the fairness of the procedure, attributes of the instances are standardized (normalized) by their means and standard deviations before submitted to the classifiers. The performance in CCR is obtained using the leave-one-out [15] procedure.

It can be seen that RNFLS performs well on both two-category and multi-category classification problems in both low and high dimensional feature spaces. This is encouraging since these datasets represent real-world problems and none of them is specially designed to suit a specific classifier. Since one common characteristic of real-world problems is distribution scattering caused by noise, the concentration property of RNFLS helps improving the correct classification rate.

6 Conclusions

This paper introduces a new feature subspace RNFLS to enhance the representational capacity of the original sample points. RNFLS constitutes a substantial improvement to NFL. It works well and its performance is independent of the distribution shape and the feature-space dimension. In particular, we have shown that RNFLS is able to generate an RNFLS feature distribution that is more concentrated than the initial distribution of the sample points and offers a higher correct classification rates for a wide range of classification applications.

Further investigation into RNFLS seems warranted. In the rectification process, it would be helpful to define the territory of one class in detail, and to treat the trespassing feature line segments more specifically, perhaps finding a way to cut off a part of a

trespasser instead of eliminating the whole feature line segments. Also worth more investigation is the concentration property, which might be of great potential and appears a good research direction.

Acknowledgments

The research work presented in this paper is supported by National Natural Science Foundation of China, Grant No. 60575022, and Science and Technology Commission of Shanghai Municipality, Grant No. 04JC14014.

References

[1] S. Z. Li and J. Lu, "Face recognition using the nearest feature line method," *IEEE Trans. Neural Networks*, vol. 10, pp. 439–443, Mar. 1999.

[2] J. T. Chien and C. C. Wu, "Discriminant waveletfaces and nearest feature classifiers for face recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, pp. 1644–1649, Dec. 2002.

[3] S. Z. Li, "Content-based audio classification and retrieval using the nearest feature line method," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 619–625, Sept. 2000.

[4] S. Z. Li, K. L. Chan, and C. L. Wang, "Performance evaluation of the nearest feature line method in image classification and retrieval," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 1335–1339, Nov. 2000.

[5] K. Chen, T. Y. Wu, and H. J. Zhang, "On the use of nearest feature line for speaker identification," *Pattern Recognition Letters*, vol. 23, pp. 1735–1746, Dec. 2002.

[6] J. H. Chen and C. S. Chen, "Object recognition based on image sequences by using inter-feature-line consistencies," *Pattern Recognition*, vol. 37, pp. 1913–1923, Sept. 2004.

[7] W. Zheng, L. Zhao, and C. Zou, "Locally nearest neighbor classifiers for pattern classification," *Pattern Recognition*, vol. 37, pp. 1307–1309, June 2004.

[8] Z. Zhou and C. K. Kwoh, "The pattern classification based on the nearest feature midpoints," in *17th International Conference on Pattern Recognition ICPR*, vol. 3, Aug. 2000, pp. 446–449.

[9] Z. Zhou, S. Z. Li, and K. L. Chan, "A theoretical justification of nearest feature line method," in *Proc. 15th ICPR International Conference on Pattern Recognition*, vol. 2, Sept. 2000, pp. 759–762.

[10] S. R. Kulkarni, "Learning pattern classification – a survey," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2178–2206, Oct. 1998.

[11] T. Denoeux and R. Lengelle, "Initializing back-propagation networks with prototypes," *Neural Networks*, vol. 6, pp. 351–363, 1993.

[12] S. Sin and R. J. P. Defigueiredo, "Efficient learning procedures for optimal interpolative nets," *Neural Networks*, vol. 6, pp. 99–113, 1993.

[13] Y. Q. Chen, D. W. Thomas, and M. S. Nixon, "Generating-shrinking algorithm for learning arbitrary classification," *Neural Networks*, vol. 7, pp. 1477–1489, 1994.

[14] C. L. Blake and C. J. Merz, "UCI repository of machine learning databases," 1998. [Online]. Available: http://www.ics.uci.edu/∼mlearn/MLRepository.html

[15] K. Fukunaga and D. M. Hummels, "Leave-one-out procedures for nonparametric error estimates," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 11, pp. 421–423, Apr. 1989.