HUITIEME COLLOQUE SUR LE TRAITEMENT DU SIGNAL ET SES APPLICATIONS

NICE du 1er au 5 JUIN 1981

Introduction parmi les échantillons de l'ensemble de référence d'échantillons non classés, et règle des k-plus proches voisins, améliorée par prise en compte de l'information locale sur le gradient

A design Sample Set Mixed With Unclassified Samples and The Improved K-nearest Neighbor Rule Considering The Local Gradient Information

Shoji Tatsumi, Shinya Kimura et Tadahiro Kitahashi
Shoji Tatsumi, Shinya Kimura and Tadahiro Kitahashi

Université de Technologie de Toyohashi
Tempaku-cho, Toyohashi, Aichi, 440 JAPON

Toyohashi University of Technology
Tempaku-cho, Toyohashi, Aichi, 440 JAPAN

RESUME

Parmi les nombreuses règles discriminatoires permettant de classer un échantillon inconnu X, la règle des k plus proches voisins(K-NNR) est bien connue. Cette règle prend comme etat de X l'etat qui apparaît le plus souvent pour les k échantillons classés $X_1,--,X_k$ les plus proches de X. Nous considerons ici le cas où à l' ensemble de référence composé d'échantillons classés sont mêlés des échantillons non classés. Dans ce cas, pour classer un échantillon inconnu, deux points de vue sont possible:

. inclure les échantillons non classés dans l'ensemble de référence,

. ne pas les inclure.

Nous montrons que l'erreur Bayes diffère selon le point de vue adoptè. Pour choisir les fonctions discriminatoires, il faut donc inclure les échantillons non classés. Toujours en utilisant les échantillons non classés, nous proposons aussi une règle K-NNR améliorée, qui prend en compte l' information locale sur le gradient. Enfin, à l'aide d'essais sur ordinateur, nous montrons que la règle K-NNR ameliorée est efficace pour de nombreux types de données differents.

SUMMARY

Among many discriminant rules which classify an unknown sample X, the k-nearest neighbor rule(K-NNR) is much known. This rule takes to be the state which occurs most often among the states of the k closest labelled samples $X_1,X_2,--,X_k$ to X. This paper considered the case where the design set of classified samples was mixed with unclassified samples. In this case, to classify an unknown sample, we will be able to take two views about the design set; as the design set, one takes all samples including the unclassified samples and the other takes the samples excluding the unclassified samples. It was shown that the Bayes error rate with considering the unclassified samples was different from the Bayes error rate without the unclassified samples. So in case of deciding the discrimination functions, as the design set, we should take all samples including the unclassified samples. Using the unclassified samples too, we also proposed the improved K-NNR considering the local gradient information. That is, for an unknown sample X to classify, we took the k closest samples(including classified and unclassified samples) to X. And next at the point X, the quasi-gradient was computed with these k samples. The sub-space where k samples spanned was cut with the hyperplane which was through the point X and orthogonal to the quasi-gradient so that we should take two half sub-spaces. As a result, we assigned to X the label which was most frequent label of classified samples in the half sub-space which was in the direction of the quasi-gradient. At last, in the computer experiments, it was shown that the improved K-NNR was effective for many different data.

Introduction parmi les échantillons de l'ensemble de référence d'échantillons non classés,et règle des k-plus proches voisins, améliorée par prise en compte de l'information locale sur le gradient
A design Sample Set Mixed With Unclassified Samples and The Improved K-nearest Neighbor Rule Considering The Local Gradient Information

I. Introduction

When the set of training samples is given, many kinds of learning machines have been studied to classify the samples. Among them there exist about two types of approaches. One approach is to constitute the discriminant functions and the other classifies an unclassified sample using the distributed configuration of the training samples. The method proposed in this paper belongs to the latter. As a classification method using the training(design) samples themselves, the nearest neighbor rule(NNR) is much known. This rule is very simple and a practical method that establishes the appropriate classification rate. But to take the appropriate classification rate, the rule requires a large number of samples so that it have the difficuties in the memory size. Besides, the configuration of the obtained samples must reflect the true distribution of the population impartially. When the design sample set is composed of known samples and unknown, we may have two behaviors; whether we reject unknown samples or not. If unkown samples were rejected, the configuration of the design samples resulted after the rejection should not reflect that of the true population impartially so that we should not be able to take the appropriate classification rate with the NNR.

Next the relation between the Bayes rule and the NNR is discussed. Usually it is difficult to compute the error rate of Bayes rule accurately so that the Bayes error rate is often estimated with using NNR. For above reasons some improved methods of NNR have been studied. Among them, the k-nearest neighbor rule (K-NNR), a K-NNR with a rejection option and a NNR with an editing rule were proposed to take the more severe error rate. In other sides a condensed NNR was proposed to reduce the memory size. This paper will discuss the problems in the case where a design sample set is mixed with unclassified samples. Besides the improved k-nearest neighbor rule with considering the unknown design samples will be proposed and this rule will be shown to be effective for many different data in computer experiments.

II. Bayes Rule and Design Sample Set Mixed With Unknown Design Samples

This section considers the case where we have obtained the design sample set mixed with unclassified samples. That is, we discuss the problems which occur when we abandon unclassified samples and use the Bayes rule.

2.1 Bayes Rule

Let $\omega_1,\omega_2,\cdots,\omega_s$ be s states of nature and the feature vector X be a n-component. In two-category case, the operation of the Bayes classifier is shown in Fig. 1. An unknown sample X is decided as followed.

$$\begin{aligned} X &\leftarrow \omega_1, \quad \text{if } p(X|\omega_1)P(\omega_1) > p(X|\omega_2)P(\omega_2) \\ X &\leftarrow \omega_2, \quad \text{if } p(X|\omega_1)P(\omega_1) < p(X|\omega_2)P(\omega_2) \end{aligned} \tag{1}$$

The sum of two "triangle areas" equals the Bayes error rate.

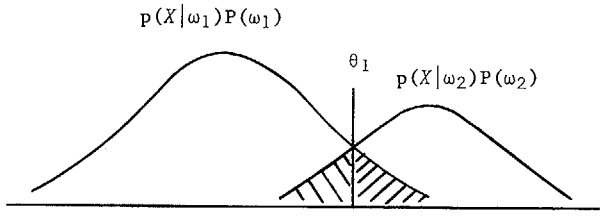


Fig.1 Bayes rule and its error rate

2.2 Problems to occur in the case of rejecting unknown samples

We consider the two-category case shown in Fig.2. $p(X|\omega_i)$ is defined as the conditional probability function, $p_s(X|\omega_i)$ as that of classified samples, and $p_N(X|\omega_i)$ as that of unknown samples. And suppose that $p(X|\omega_i)$ is composed of $p_s(X|\omega_i)$ and $p_N(X|\omega_i)$.

$$p(X|\omega_i) = p_s(X|\omega_i) + p_N(X|\omega_i) \tag{2}$$

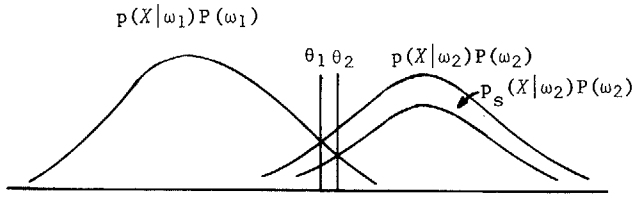


Fig.2 The threshold θ_2 in case of rejecting unknown design samples and the threshold θ_1 in case of not rejecting

To speak briefly, we assume that $p(X|\omega_1)$ is composed of known samples and $p(X|\omega_2)$ is composed of known samples and unkown samples. For these situation, we may take two points of views in order to classify an unknown sample when we use the NNR. That is,

Introduction parmi les échantillons de l'ensemble de référence d'échantillons non classés, et règle dès k-plus proches voisins, améliorée par prise en compte de l'information locale sur le gradient
A design Sample Set Mixed With Unclassified Samples and The Improved K-nearest Neighbor Rule Considering The Local Gradient Information

(A) an unknown sample is classified using $p(X|\omega_1)$ and $p(X|\omega_2)$, and

(B) an unknown sample is classified using $p(X|\omega_1)$ and $p_s(X|\omega_2)$ except $p_N(X|\omega_2)$.

In Fig.2, it is clear that a θ_1 conducted from the method(A) is different from a θ_2 from the method(B). An test sample to be classified should occur from the mixed distribution of $p(X|\omega_1)$ and $p(X|\omega_2)$ so that, to take the minimum error rate, the θ_1 computed from the method(A) should be selected to classify an unknown sample. If the configuration of $p_N(X|\omega_2)$ is the similar figure of $p_s(X|\omega_2)$ or $p_N(X|\omega_2)$ is definite, the threshold θ_1 is easy to compute. But $p_N(X|\omega_i)$ is not usually plain so that, while obtaining the design sample set mixed with unknown samples, unknown samples were always excluded to constitute a classifier.

To discuss analytically, suppose that $p_s(X|\omega_i)$ and $p_N(X|\omega_i)$ are normal distributions and we shall consider the case where $p_N(X|\omega_i)$ is the similar figure of $p_s(X|\omega_i)$ as shown in eq.(3). That is,

$$p_N(X|\omega_i) = k_i p_s(X|\omega_i), \tag{3}$$

where k_i is any value but $0 \leq k_i \leq 1$.

We assume that

$$p_s(x|\omega_i) = \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp[-\frac{1}{2}(\frac{x-\mu_i}{\sigma_i})^2]. \tag{4}$$

The Bayes error rates using both the method(A) and (B) will be accumulated.

(A)
$$p(x|\omega_1) = \frac{1}{\sqrt{2\pi}\,\sigma_1} \exp[-\frac{1}{2}(\frac{x-\mu_1}{\sigma_1})^2] + k_1 \frac{1}{\sqrt{2\pi}\,\sigma_1} \exp[-\frac{1}{2}(\frac{x-\mu_1}{\sigma_1})^2]$$
$$= (1+k_1) \frac{1}{\sqrt{2\pi}\,\sigma_1} \exp[-\frac{1}{2}(\frac{x-\mu_1}{\sigma_1})^2] \tag{5}$$

$$p(x|\omega_2) = (1+k_2) \frac{1}{\sqrt{2\pi}\,\sigma_2} \exp[-\frac{1}{2}(\frac{x-\mu_2}{\sigma_2})^2] \tag{6}$$

The threshold θ_1 can be taken with the equation

$$P(\omega_1)p(x|\omega_1) = P(\omega_2)p(x|\omega_2). \tag{7}$$

For simplicity, we assume that $\sigma_1=\sigma_2$ and $P(\omega_2) = P(\omega_1) = \frac{1}{2}$, and then compute the error rates. Then eq.(7) is

$$\frac{1}{\sigma_1^2}(\mu_1-\mu_2)x - \frac{1}{2}\frac{1}{\sigma_1^2}(\mu_1^2-\mu_2^2) - \log\frac{k_2+1}{k_1+1} = 0. \tag{8}$$

As a result,

$$\theta_1 = \theta_o + \theta_1', \tag{9}$$

where

$$\theta_o = \frac{1}{2}(\mu_1+\mu_2) \tag{10}$$

$$\theta_1' = \frac{\sigma_1^2}{\mu_1-\mu_2} \log\frac{k_2+1}{k_1+1}. \tag{11}$$

(B)
$$p_s(x|\omega_1) = \frac{1}{\sqrt{2\pi}\,\sigma_1} \exp[-\frac{1}{2}(\frac{x-\mu_1}{\sigma_1})^2] \tag{12}$$

$$p_s(x|\omega_2) = \frac{1}{\sqrt{2\pi}\,\sigma_2} \exp[-\frac{1}{2}(\frac{x-\mu_2}{\sigma_2})^2] \tag{13}$$

We compute the threshold θ_2 in the same way as (A). So we gain

$$\theta_2 = \theta_o. \tag{14}$$

Bayes error rate can be computed as followed,

$$P(\text{error}) = \int_{\theta}^{\infty} p(x|\omega_1)P(\omega_1)dx + \int_{-\infty}^{\theta} p(x|\omega_2)P(\omega_2). \tag{15}$$

The calculated results in the case where $\mu_1=0$, $\mu_2=2$, $k_1=0$, $\sigma_1=\sigma_2=1$, and $P(\omega_1)=P(\omega_2)=1/2$ are shown in Fig. 3, where P_{eA} means the error rate with the method(A), and P_{eB} the error rate with the method(B).

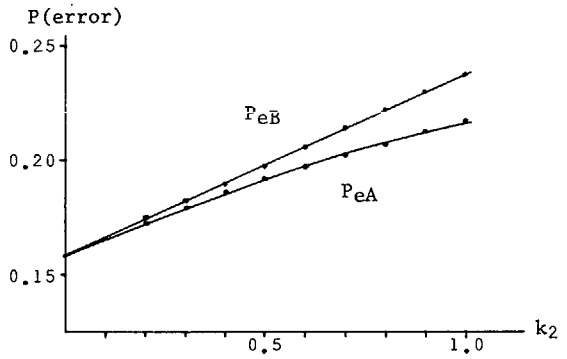


Fig.3 Error rate
($\mu_1=0, \mu_2=2, k_1=0, \sigma_1=\sigma_2=1$, and $P(\omega_1)=P(\omega_2)=1/2$)

From Fig.2 and Fig.3, It is clear that we should not reject unknown design samples, when we have obtained the design sample set mixed with unknown samples.

As a method which considers the case where the probability of available design samples is different from a prior probability, the weighted nearest neighbor rule was proposed in ref.(3). But the defect of this method is that we need a prior probability $P(\omega_i)$, which is impossible to take before. And this method is unsuitable for the case where the distribution is locally biased.

III. A Improved K-nearest Neighbor Rule Considering The Unknown Samples Of The Design Set

Though the NNR and the K-NNR are very simple discrimination rules, they have good performance only in the case of the infinite number of samples. This is why the distribution of the infinite number of samples has completely reflected that of the population. We

Introduction parmi les échantillons de l'ensemble de référence d'échantillons non classés, et règle des k-plus proches voisins, améliorée par prise en compte de l'information locale sur le gradient
A design Sample Set Mixed Unclassified Samples and The Improved K-nearest Neighbor Rule Considering The Local Gradient Information

shall propose the method which decides the class of an unknown sample without leaving the information which the unclassified sample has in the distributed design set. In Fig.4, it is reasonable that an unknown sample X should take the state which occurs most often among the states of samples lying in the direction of an ascent. When many unknown design samples lie in the direction of an ascent, the removal of these unknown samples will make a mistake if the label which occurs most often in the direction of an descent is major in the region of the k closest labelled samples to X. So we will propose the improved k-NNR considering the local gradient information.

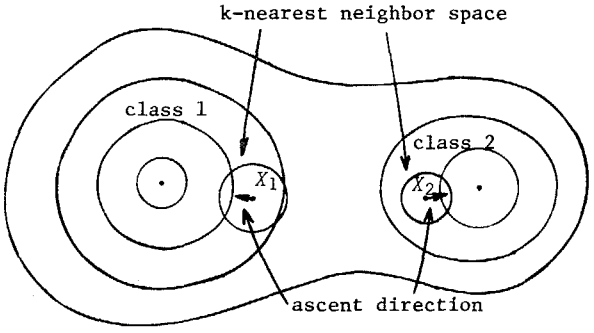


Fig.4 k-nearest neighborhood of an unknown sample X

[K-NNR with using a local gradient value]

Step 1: Take up the L-nearest neighbor sub-space in which an unknown sample X to classify is the central point. While doing, unclassified samples Y_i are included within L samples.

Step 2: The L-nearest neighbor sub-space in Step 1 is partitioned with the hyperplane which is through the point X, and we gain a pair of half sub-spaces.

Step 3: Among many pairs of half sub-spaces in Step 2, we select the half sub-space where the most design samples(including unclassified samples) are distributed.

Step 4: If there exist more than two half sub-spaces having the most design samples, compute eq(16) as followed.

$$V_x = \frac{1}{m_j} \sum_{Z_i \in N_{x,L'}} f(d)\, e_{(Z_i - X)}, \qquad (16)$$

where

$N_{x,L'} = [\, Y \mid |Y-X| < L' \,]$: neighborhood of X,

m_j : number of elements in $N_{x,L'}$,

$e_{(Z_i - X)}$: unit vector directed from X to Z_i,

f(d) : certain function of d, and

d : distance between X and Z_i.

Step 5: Among many half sub-spaces, we pick up the half sub-space which has the maximum value of $|V_x|$ computed in Step 4. As a result, we assign to X the label which is the most frequent label of classified samples in the selected half sub-space.

Step 6: In the half sub-space picked up in Step 5, if there exist more than two labels which have the most frequency, the classification of X is withheld.

Ⅳ. Computer Experiments

To compare the improved K-NNR with the K-NNR, computer experiments for two-category bivariate gaussian distribution were performed.

[Process of computer experiments]

Process 1: Two kinds of bivariate gaussian random numbers are generated.

Process 2: The gaussian random numbers are separated into the set of classified design samples and the set of unclassified design samples.

Process 3: All unclassified design samples are classified with both the K-NNR and the proposed method.

Among many separation ways devised in Process 2, one separation way we adopted is that the samples of the gaussian distribution are separated in such a way that the configuration of the unclassified design set is similar to that of the classified design set. Let M be the number of all samples, and m unknown samples are separated in the following way. In section[0,1], m(<M) random numbers of the uniform distribution are generated but not overlapped. Next, m numbers of the gaussian distribution which correspond to m integer numbers of the uniform distribution converted in [1,M] are picked up to constitute the design set of unclassified samples.

[Process of the proposed K-NNR]

Step 1: Compute all distances d_j(j=1,--,M) between an unknown sample X to classify and each design sample.

Step 2: Among all distances computed in Step 1, select L design samples closest to X.

Step 3: Among many pairs of half sub-spaces, we choose the half sub-space which takes the most samples. If there exist more than two half sub-spaces which take the most samples, we compute eq(16). In eq(16), we used

$$f(d) = \frac{1}{d + a}. \quad (a=0.005) \qquad (17)$$

As a result, we choose the half sub-space which has the maximum value of eq(16).

Step 4: Perform the same processes of Step 5,6 in the section Ⅲ.

[Experiment 1]

Many samples were generated from each two bivariate gaussian distributions with the following parameters(M_i denotes the mean, and Σ_i the covariance matrix for class i),

$$M_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad M_2 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \text{ and}$$
$$\Sigma_1 = \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{18}$$

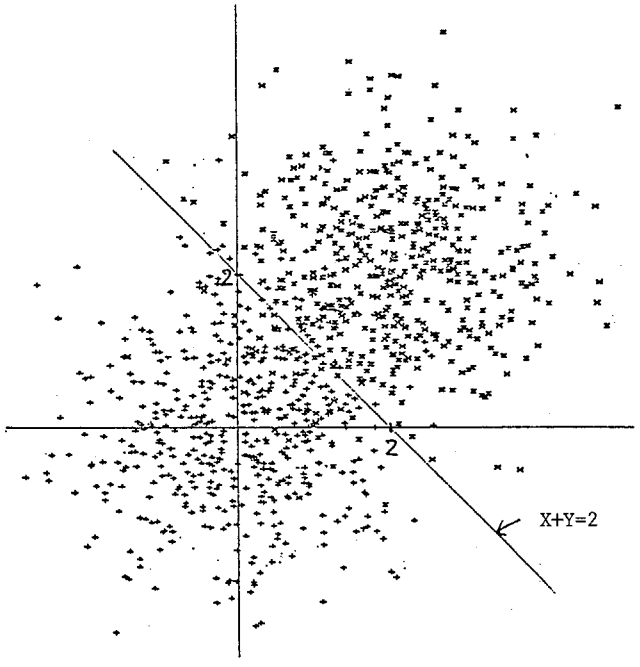
The set of class 1 was composed of 500 classified design samples(sign +), and the set of class 2 was composed of 200 classified design samples(sign x)and of 300 unclassified design samples(sign ⧻). The configuration of unclassified samples is similar to that of classified samples. In Fig.5, the distributions and their classification results were shown, and the sign + means error, the sign x correct and the sign ⧻ withheld. In Fig.5(b), we set that k=7, and in Fig.5(c) L=15. The Bayes discrimination function is

$$X + Y = 2, \tag{19}$$

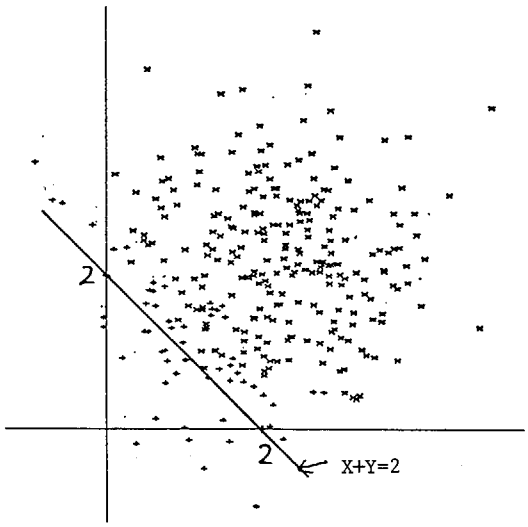
and using this rule made 20 error numbers. Using K-NNR(k=7) made 55 errors and using the proposed method (L=15) made 46 errors and 12 withheld points.

[Experiment 2]

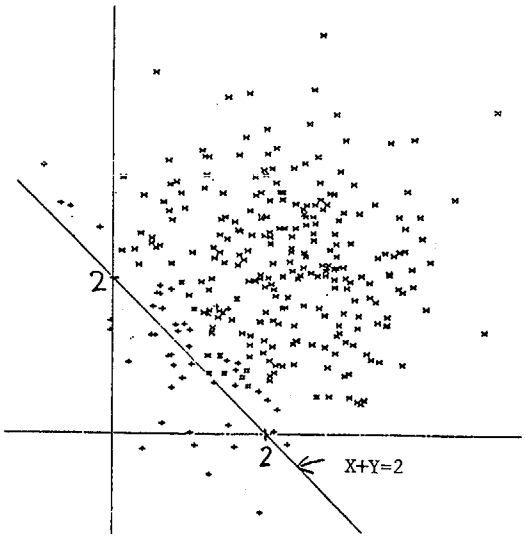
In the Experiment 2, 200 classified samples of class 1 and 484 known samples and 16 unknown samples of class 2 were generated with the same parameters in Experiment 1. The unknown samples of class 2 were, among the samples of class 2 generated in Experiment 1, the samples which satisfy the inqual equation $x/4+y < 1$. As a result, the error number with the Bayes rule is 9, the error number with using the improved K-NNR(L=15) is 9, and the error number with the K-NNR(k=7) is 10.



(a) two-category gaussian distribution



(b) K-NNR (k=7)



(c) improved K-NNR (L=15)

Fig.5 Experiment 1 (Similar case) two-category gaussian distribution and the classification results with the K-NNR (k=7) and the improved K-NNR(L=15)

Ⅴ. Discussions and Conclusions

When we used the improved K-NNR for any data, we took some undecided samples. This is why the number of the samples labelled class 1 is equal to that of the samples labelled class 2 in the nearest neighbor subspace. To decide the class of undecided samples, the radius of the neighborhood space is increased and the number of classified samples may be limited to be odd. In Fig.6, while comparing K-NNR with the improved K-NNR, it is expected that the proposed method will do

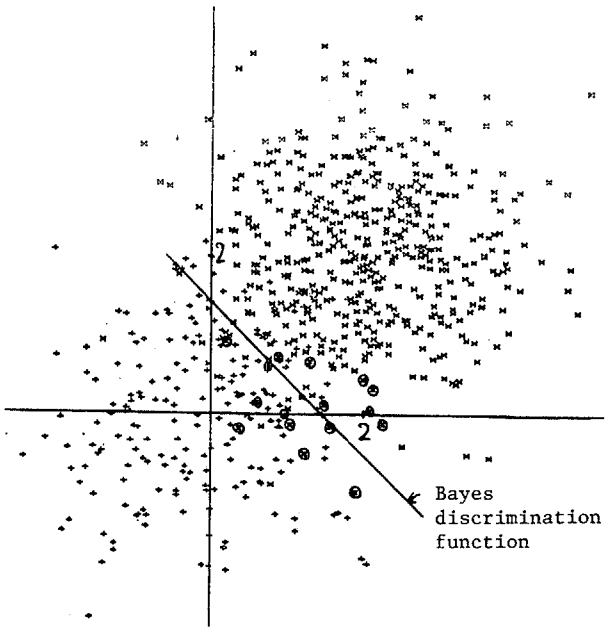
Introduction parmi les échantillons de l'ensemble de référence d'échantillons non classés,et règle des k-plus proches voisins, améliorée par prise en compte de l'information locale sur le gradient
A design Sample Set Mixed With Unclassified Samples and The Improved K-nearest Neighbor Rule Considering The Local Gradient Information

more performance than the K-NNR. But we need the error estimation in the case of the infinite samples.

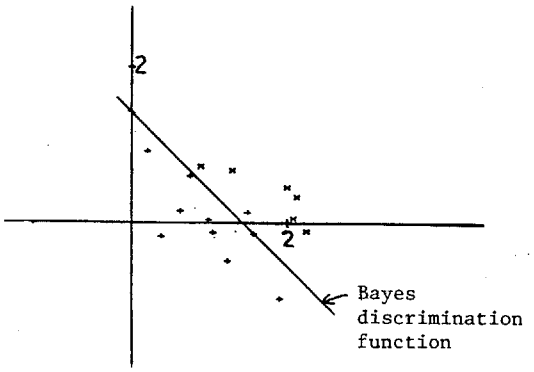
In this paper we have discussed the problems which ocurred when the design set was mixed with some unknown samples and we have had the conclusion that we should not abandon any unknown design samples but utilize the information of the location which unknown samples have. And we have proposed the improved K-NNR using the local gradient information and it has been made sure to be efficient for many data in computer experiments.

Reference

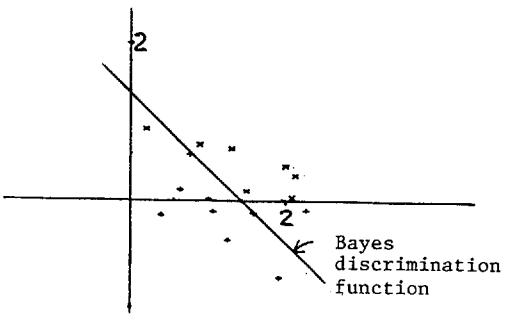
(1) T.Cover and P.Hart,"Nearest neighbor pattern classification," IEEE Trans. Inform. Theory, vol IT-13,pp.21-27,Jan. 1962
(2) T.Cover,"Estimation by the nearest neighbor rule," IEEE Trans. Inform. Theory, vol.IT-14,pp.50-55,Jan. 1968
(3) T.A.Brown and J.Koplowitz,"The weighted nearest neighbor rule for class dependent sample sizes," IEEE Trans. Inform. Theory, vol.IT-25,No.5, pp.617-619,1979
(4) S.Tatsumi,Y.Nagata,H,Sanada and Y.Tezuka,"A Non-parametric clustering using the quasi-gradient," IECE of JAPAN, vol.E 61,pp.50-51,Jan. 1978
(5) R.O.Duda and P.E.Hart, "Pattern Classification and Scene Analysis," John Wiley & Sons,1973
(6) D.L.Wilson,"Asymptotic properties of NN rules using edited data," IEEE Trans. Syst., Man,Cybern., vol.SMC-2,pp.408-421,July,1972
(7) I.Tomek,"An experiment with the edited nearest neighbor rule," IEEE Trans. Syst.,Man,Cybern.,vol. SMC-6, pp.448-452,June,1976
(8) P.E.Hart,"The condensed nearest neighbor," IEEE Trans. Inform. Theory, vol.IT-14,pp.515-516,May,1968
(9) M.E.Hellman,"The nearest neighbor classification rule with a reject option," IEEE Trans. Syst.,Sci., Cybern., vol.SSC-6, pp.179-185, July, 1970
(10) W.H.Rogers and T.J.Wangner,"A finite sample distribution-free performance bound for local discrimination rules," The Annals of Statistics vol.6, No.3 pp.506-514,1978



(a) two-category gaussian distribution



(b) K-NNR (k=7)



(c) improved K-NNR (L=15)

Fig.6 Experiment 2
two-category gausian distribution and the classification results with the K-NNR (k=7) and the improved K-NNR(L=15)
(sign + means error and sign x correct)