Improving Kernel Incapability by Equivalent Probability in Flexible Naïve Bayesian

James N. K. Liu
Department of computing,
The Hong Kong Polytechnic University,
Hung Hom, Kowloon, Hong Kong
Email: csnkliu@inet.polyu.edu.hk

Yu-Lin He
College of Mathematics and
Computer Science, Hebei University,
Baoding 071002, Hebei, China
Email: csylhe@gmail.com

Xi-Zhao Wang
College of Mathematics and
Computer Science, Hebei University,
Baoding 071002, Hebei, China
Email: xizhaowang@ieee.org

***Abstract*—In flexible naïve Bayesian (FNB), the excellent qualities of Gaussian kernel have been demonstrated by the theoretical analyses and experimental comparisons with normal naïve Bayesian (NNB). There are also several types of kernel functions commonly used for probability density estimation, i.e., uniform, triangular, epanechnikov, biweight, triweight and cosine. We call them discontinuous kernels. In this paper, we verify the feasibility and efficiency of applying these alternative kernels in FNB. Our works mainly focus on three aspects: firstly, we give the application conditions of these kernels for the given domain data by analyzing the structural difference between the discontinuous kernel and Gaussian kernel; secondly, the equivalent probability is proposed to improve the capabilities of discontinuous kernels when such problem of kernel incapability occurs; finally, we carry out the experimental demonstration of our proposed method based on 15 UCI datasets. The results show that the discontinuous kernels can obtain better classification accuracies with the help of equivalent probabilities.**

***Index Terms*—discontinuous kernel, equivalent probability, flexible naïve Bayesian, Gaussian kernel, kernel incapability**

I. Introduction

For the convenience of discussion, we firstly describe the notations used in our study. Let $X = \{x_1, x_2, \cdots, x_n\}$ be the given dataset, where $x_i = \{x_{i1}, x_{i2}, , x_{id}, c_i\}$ denotes the ith $(i = 1, 2, \cdots, n)$ training example in the given dataset X, n is the number of samples in X, x_{ij} is the jth $(j = 1, 2, \cdots, d)$ feature attribute of x_i, d is the number of feature attributes of x_i, c_i is the class attribute of x_i, $c_i \in \{w_1, w_2, \cdots, w_k\}$, k is the number of classes. In our study, we mainly apply NBC to deal with the classification problems of continuous (numerical) attributes.

Naïve Bayesian classifier (simply NBC) [1] is one kind of very popular classifiers. Its outstanding advantages have been demonstrated in many practical and theoretical fields. When NBC is used to carry out a classification task for a new instance $x = \{x_1, x_2, \cdots, x_d\}$, where x_j is the jth $(j = 1, 2, , d)$ feature attribute of x, the following Eq. (1) should be calculated. Let the class attribute of x be $c(x)$:

$$c(x) = \arg\max_{i=1,2,\cdots,k} p(w_i)\, p(x \,|w_i)\,. \tag{1}$$

NBC applies the Bayesian theory as the classification basis. In Eq. (1), for the consideration of computing time, we usually make $p(w_i) = \frac{1}{k}, (i = 1, 2, \cdots, k)$ which is the prior probability. The key point of NBC is the derivation of $p(x|w_i)$ which is called the class conditional probability. The joint probability $p(x|w_i) = p(x_1, x_2, \cdots, x_d|w_i), (i = 1, 2, \cdots, k)$ can be computed based on the assumption that all feature attributes of x are independent. This is the reason why this probability-like classifier is called naïve classifier. So, Eq. (1) can also be written as another form as shown in Eq. (2):

$$c(x) = \arg\max_{i=1,2,\cdots,k} p(w_i) \prod_{j=1}^{d} p(x_j \,|w_i\,). \tag{2}$$

For the continuous attributes, probability density estimation is used to provide the solution for $p(x_j|w_i)$ based on the given dataset X. In 1995, John and Langley [2] proposed the flexible Bayesian classifier to solve $p(x_j|w_i), (i = 1, 2, \cdots, k; j = 1, 2, \cdots, d)$. In their work, the authors explained why the classifier is more flexible than normal naïve Bayesian. The mainly reason is that the classifier is efficient for the attribute which is not normally distributed. FNB estimated $p(x_j|w_i)$ by using the Parzen window method [3]. The computing equation of $p(x_j|w_i)$ based on the given dataset X can be formulated as following Eq. (3):

$$p(x_j \,|w_i\,) = \frac{\sum_{p=1}^{n_{w_i}} \mathrm{K}\left(\frac{x_j - x_{pj}}{h_{ij}}\right)}{n_{w_i} h_{ij}}, \tag{3}$$

where n_{w_i} is the number of samples belonging to the class $w_i, (i = 1, 2, \cdots, k)$. The parameter $h_{ij} = \frac{1}{\sqrt{n_{w_i}}}$ is the bandwidth. The kernel function $\mathrm{K}(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$ used in John and Langley's work is called Gaussian kernel.

In the research filed of probability density estimation, it is well accepted that the choice of kernel function $\mathrm{K}(\cdot)$ is less important than the selection of bandwidth [4], [5], [6], [7], [8]. So, the studies of kernel selection are smaller in scope compared with the bandwidth determination. Meanwhile, under the framework of classification task, there are also very little researches about the kernel selection.

In this paper, we try to explore the influence exerted on flexible Bayesian classifier by different kernels. In the probability density estimation, there are also six commonly used kernels: uniform, triangular, epanechnikov, biweight, triweight and cosine kernels [9]. The efficiencies of these kernels have

U.S. Government work not protected by U.S. copyright

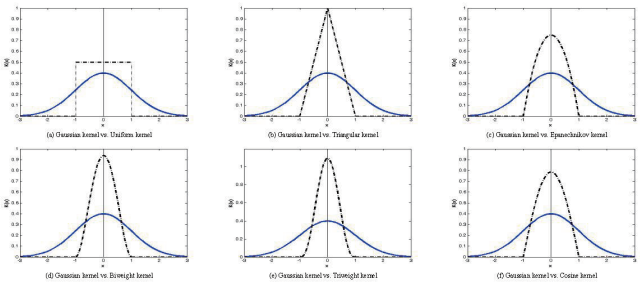


Fig. 1. Six different types of discontinuous kernels

TABLE I
SIX KERNEL FUNCTIONS STUDIED IN OUR RESEARCH

Kernel function $\mathrm{K}(x)$	$\lvert x\rvert \leq 1$	$\lvert x\rvert > 1$
Uniform	$\frac{1}{2}$	0
Triangular	$1-\lvert x\rvert$	0
Epanechnikov	$\frac{3\left(1-x^2\right)}{4}$	0
Biweight	$\frac{15\left(1-x^2\right)^2}{16}$	0
Triweight	$\frac{35\left(1-x^2\right)^3}{32}$	0
Cosine	$\frac{\pi \cos \frac{\pi}{2} x}{4}$	0

been demonstrated by the better mean integrated squared error (MISE) when probability density estimation is considered. However, these six kernel functions are not compared based on the 0-1 loss (misclassification rate) [10]. We firstly introduce these kernels into FNB to replace the Gaussian kernel in order to verify whether there is any difference of classification performance between Gaussian kernel and other kernels. In our study, these six alternative kernels are called discontinuous kernels. Our works mainly focus on the following three aspects. Firstly, we analyze the structural difference between the discontinuous kernel and the Gaussian kernel. And, based on this analysis, the application conditions of discontinuous kernels are given. Secondly, we explore the kernel incapability which is the phenomenon that the FNB equipped with the discontinuous kernels could not classify some instances. Focused on the kernel incapability, the equivalent probability is proposed to improve the capabilities of discontinuous kernels. Finally, we carry out the experimental techniques to verify the queries proposed above based on 15 UCI benchmark datasets [11]. Our experiments include two parts: the comparison between Gaussian kernel and the discontinuous kernels, and the comparison between the discontinuous kernels before and after applying the equivalent probability. The evaluation method of classifiers is classification accuracy [12] in this paper. Our testing results are all average values of 10 runs of 10-fold cross-validation. The statistical method, two-tailed t-test with a 95 percent confidence level [13], is used to compare the rival kernel's $win/tie/lose$ records. The empirical observations and analyses on comparative results show that the Gaussian kernel can not achieve statistically better classification accuracy than discontinuous kernels. And, the equivalent probability method is feasible which can indeed improve the kernel incapability effectively and enhance the classification performances of discontinuous kernels significantly.

The rest of the paper is organized as follows: In Section II, we summarize the existing kernel functions and classify these kernels. Section III analyzes the structural difference between Gaussian kernel and discontinuous kernel. In Section IV, we introduce the kernel incapability and propose the equivalent probability to improve the performances of discontinuous kernels. The experimental setup and results are described in Section V. Section VI makes a conclusion and outlines the main directions for future research.

II. KERNEL FUNCTIONS

The kernel is widely used in many areas, for example, SVM, density estimation, and so on. It is a function which is non-negative, real-valued and integrable. For a given kernel $\mathrm{K}(x)$, it should satisfy two requirements [14]:

$$\int_{-\infty}^{+\infty} \mathrm{K}(x) = 1 \text{ and for } \forall x, \mathrm{K}(x) = \mathrm{K}(-x).$$

There are other six common kernels used in areas of density estimation except Gaussian kernel. Table I shows the detailed descriptions of these six kernels. For observing these kernels more intuitively, Fig. 1 gives the comparative pictures between Gaussian kernel and every kernel listed in Table I. From the expressions of these kernels, we can see that all these six kernels are discontinuous. So, we call these kernels discontinuous kernels. In Fig. 1, the blue solid line denotes the Gaussian kernel and the black dashdot lines denote six different types of discontinuous kernels. The subfigure (a) depicts the sharp of uniform kernel. When $x \in [-1, 1]$, the graph of uniform kernel is a line which is parallel with the x-axis. We call uniform kernel discontinuous-line kernel. While the subfigures (b)-(f) plot another figurate kernel: when $x \in [-1, 1]$, the graphs of these kernels are curved (when $x \in [-1, 1]$, we deem the graph of triangular kernel as a curve which is constituted by two oblique lines). Then, these five kernels are called discontinuous -curve kernels. Based on the above classification of kernels, we will give the application conditions of discontinuous-line kernel (DLK) and discontinuous-curve kernels (DCK).

III. THE APPLICATION CONDITIONS OF DIFFERENT KERNEL FUNCTIONS

Without loss of generality, the classification is considered to be two classes, denoted by $+$ and $-$ respectively. The instances in class $+$ are $\{x_{i1}, x_{i2}, \cdots, x_{id}\}$, $(i = 1, 2, \cdots, n_+)$, where n_+ is the number of instances in class $+$. And, the instances in class $-$ are $\{y_{i1}, y_{i2}, \cdots, y_{id}\}$, $(i = 1, 2, \cdots, n_-)$, where n_- is the number of instances in class $-$. For a given new sample $x = \{x_1, x_2, \cdots, x_d\}$, we assume $\exists j \in \{1, 2, \cdots, d\}$, such that Eq. (4) holds when Gaussian kernel is used in FNB:

$$p_{\text{Gaussian}}(x_j \mid +) < p_{\text{Gaussian}}(x_j \mid -), \tag{4}$$

where, $p_{\text{Gaussian}}(\cdot)$ is the probability density function estimated by using Gaussian kernel. Combining Eqs. (3) and (4), we can derive Eq. (5) as follows:

$$\frac{\sum_{p=1}^{n_+}\left[\mathrm{K}_{\mathrm{Gaussian}}\left(\frac{x_j-x_{pj}}{h_{+j}}\right)\right]}{n_+h_{+j}} < \frac{\sum_{q=1}^{n_-}\left[\mathrm{K}_{\mathrm{Gaussian}}\left(\frac{x_j-y_{qj}}{h_{-j}}\right)\right]}{n_-h_{-j}}, \quad (5)$$

where, $\mathrm{K}_{\mathrm{Gaussian}}(\cdot)$ denotes Gaussian kernel function. Then, our work is to find the conditions under which the character $<$ in Eq. (5) could be changed when DLK or DCK is used in FNB.

A. The Application Conditions of Discontinuous-Line Kernel

From the subfigure (a) in Fig. 1, we can see that the functions of Gaussian kernel and uniform kernel are symmetrical with respect to $x = 0$. So, we only study the situation of $x > 0$. Let $\alpha_p = \frac{|x_j - x_{pj}|}{h_{+j}}$, $(p = 1, 2, \cdots, n_+)$ and $\beta_q = \frac{|x_j - y_{qj}|}{h_{-j}}$, $(q = 1, 2, \cdots, n_-)$.

1) $0 \le \alpha_p < 1$, $\forall p \in \{1, 2, \cdots, n_+\}$ *and* $0 \le \beta_q < 1$, $\forall q \in \{1, 2, \cdots, n_-\}$*:* Under this situation, DLK in FNB will not classify the jth $(j = 1, 2, \cdots, d)$ feature attribute of x. It is because all $\mathrm{K}_{\mathrm{DLK}}(\alpha_p)$ and $\mathrm{K}_{\mathrm{DLK}}(\beta_q)$ equal $\frac{1}{2}$, $p_{\mathrm{Gaussian}}(x_j|+) = p_{\mathrm{Gaussian}}(x_j|-)$.

2) $\exists q \in \{1, 2, \cdots, n_-\}$, s.t. $\beta_q > 1$*:* The right part of (5) can be reduced when DLK is used under this condition. It is because when $\beta_q > 1$, $\mathrm{K}_{\mathrm{DLK}}(\beta_q) = 0$. Now, a special example is given to explain this situation.

Example 1: Let $x_{1j} = 0.335$, $x_{2j} = 0.656$, $y_{1j} = 0.392$, $y_{2j} = 0.627$, $y_{3j} = 0.699$, and $y_{4j} = 0.397$. For the given $x_j = 0.932$, we can get $h_{+j} = \frac{1}{\sqrt{2}} = 0.707$, and $h_{-j} = \frac{1}{\sqrt{4}} = 0.500$. And, $\frac{|x_j - x_{1j}|}{h_{+j}} = 0.843$, $\frac{|x_j - x_{2j}|}{h_{+j}} = 0.390$, $\frac{|x_j - y_{1j}|}{h_{-j}} = 1.079$, $\frac{|x_j - y_{2j}|}{h_{-j}} = 0.609$, $\frac{|x_j - y_{3j}|}{h_{-j}} = 0.465$, and $\frac{|x_j - y_{4j}|}{h_{-j}} = 1.069$. So, $p_{\mathrm{Gaussian}}(x_j|+) = 0.437$, $p_{\mathrm{Gaussian}}(x_j|-) = 0.547$, $p_{\mathrm{Gaussian}}(x_j|+) < p_{\mathrm{Gaussian}}(x_j|-)$. Then, $p_{\mathrm{DLK}}(x_j|+) = 0.707$, $p_{\mathrm{DLK}}(x_j|-) = 0.500$, $p_{\mathrm{DLK}}(x_j|+) > p_{\mathrm{DLK}}(x_j|-)$, where $p_{\mathrm{DLK}}(\cdot)$ is the probability density estimated by using DLK function. ∎

B. The Application Conditions of Discontinuous-Curve Kernel

We can easily find that there are considerable similarities in the subfigures (b)-(f) of Fig. 1. Similarly, our study only focuses on the situation of $x > 0$ because of the symmetry. The common qualities in the subfigures (b)-(f) of Fig. 1 can be summarized as follows:

- There is one intersection point between DCK and Gaussian kernel. We call it *critical point* (*cp*);
- For $\forall x \in [0, cp)$, $\mathrm{K}_{\mathrm{Gaussian}}(x) < \mathrm{K}_{\mathrm{DCK}}(x)$;
- For $\forall x \in [cp, 1)$, $\mathrm{K}_{\mathrm{Gaussian}}(x) > \mathrm{K}_{\mathrm{DCK}}(x) \neq 0$;
- For $\forall x \in [1, +\infty)$, $\mathrm{K}_{\mathrm{Gaussian}}(x) > \mathrm{K}_{\mathrm{DCK}}(x) = 0$.

Based on the above common qualities, we will conclude the application conditions of DCK:

1) $0 \le \alpha_p < cp$, $\forall p \in \{1, 2, \cdots, n_+\}$ *and* $cp \le \beta_q < 1$, $\forall q \in \{1, 2, \cdots, n_-\}$*:* If $p_{\mathrm{Gaussian}}(x_j|+) > p_{\mathrm{Gaussian}}(x_j|-)$, $p_{\mathrm{DCK}}(x_j|+) > p_{\mathrm{DCK}}(x_j|-)$ must hold. It is because when $\alpha_p \in [0, cp)$, $\mathrm{K}_{\mathrm{Gaussian}}(\alpha_p) < \mathrm{K}_{\mathrm{DCK}}(\alpha_p)$ and $\beta_q \in [cp, 1)$, $\mathrm{K}_{\mathrm{Gaussian}}(\beta_q) > \mathrm{K}_{\mathrm{DCK}}(\beta_q)$. Then,

$$\mathrm{K}_{\mathrm{DCK}}(\alpha_p) > \mathrm{K}_{\mathrm{Gaussian}}(\alpha_p) > \mathrm{K}_{\mathrm{Gaussian}}(\beta_q) > \mathrm{K}_{\mathrm{DCK}}(\beta_q)$$

and

$$p_{\mathrm{DCK}}(x_j|+) > p_{\mathrm{Gaussian}}(x_j|+) > p_{\mathrm{Gaussian}}(x_j|-) > p_{\mathrm{DCK}}(x_j|-),$$

where, $p_{\mathrm{DCK}}(\cdot)$ is the probability density estimated by using DCK function. Under this condition, DCK can obtain the same determining result as Gaussian kernel.

If $p_{\mathrm{Gaussian}}(x_j|+) < p_{\mathrm{Gaussian}}(x_j|-)$, $p_{\mathrm{DCK}}(x_j|+) > p_{\mathrm{DCK}}(x_j|-)$ may hold. Under this condition, DCK can be used. Now, we use a special example to illustrate this conclusion.

Example 2: We select Epanechnikov kernel as testing kernel (other DCKs can also be used). Let $x_{1j} = 0.776$, $x_{2j} = 0.273$, $y_{1j} = 0.795$, $y_{2j} = 0.794$, $y_{3j} = 0.781$, and $y_{4j} = 0.791$. For the given $x_j = 0.368$, we can also get $h_{+j} = 0.707$, and $h_{-j} = 0.500$. And, $\frac{|x_j - x_{1j}|}{h_{+j}} = 0.576$, $\frac{|x_j - x_{2j}|}{h_{+j}} = 0.107$, $\frac{|x_j - y_{1j}|}{h_{-j}} = 0.853$, $\frac{|x_j - y_{2j}|}{h_{-j}} = 0.851$, $\frac{|x_j - y_{3j}|}{h_{-j}} = 0.826$, and $\frac{|x_j - y_{4j}|}{h_{-j}} = 0.845$. So, $p_{\mathrm{Gaussian}}(x_j|+) = 0.530$, $p_{\mathrm{Gaussian}}(x_j|-) = 0.559$, $p_{\mathrm{Gaussian}}(x_j|+) < p_{\mathrm{Gaussian}}(x_j|-)$. Then, $p_{\mathrm{DCK}}(x_j|+) = 0.878$, $p_{\mathrm{DCK}}(x_j|-) = 0.432$, $p_{\mathrm{DCK}}(x_j|+) > p_{\mathrm{DCK}}(x_j|-)$. In this example, the critical point $cp \approx 0.779$. We can find that $0 < \frac{|x_j - x_{1j}|}{h_{+j}} < cp$, $0 < \frac{|x_j - x_{2j}|}{h_{+j}} < cp$, $cp < \frac{|x_j - y_{1j}|}{h_{-j}} < 1$, $cp < \frac{|x_j - y_{2j}|}{h_{-j}} < 1$, $cp < \frac{|x_j - y_{3j}|}{h_{-j}} < 1$, and $cp < \frac{|x_j - y_{4j}|}{h_{-j}} < 1$. ∎

2) $0 \le \alpha_p < cp$, $\forall p \in \{1, 2, \cdots, n_+\}$ *and* $\exists q \in \{1, 2, \cdots, n_-\}$, s.t. $0 \le \beta_q < cp$*:* The DCK can also reverse the result obtained with Gaussian kernel and it can be used by FNB. The following example explains this.

Example 3: We also select Epanechnikov kernel as the testing kernel. Let $x_{1j} = 0.877$, $x_{2j} = 0.969$, $y_{1j} = 0.113$, $y_{2j} = 0.987$, $y_{3j} = 0.960$, and $y_{4j} = 0.605$. For the given $x_j = 0.490$, we can also get $h_{+j} = 0.707$, and $h_{-j} = 0.500$. And, $\frac{|x_j - x_{1j}|}{h_{+j}} = 0.547$, $\frac{|x_j - x_{2j}|}{h_{+j}} = 0.107$, $\frac{|x_j - y_{1j}|}{h_{-j}} = 0.754$, $\frac{|x_j - y_{2j}|}{h_{-j}} = 0.994$, $\frac{|x_j - y_{3j}|}{h_{-j}} = 0.940$, and $\frac{|x_j - y_{4j}|}{h_{-j}} = 0.229$. So, $p_{\mathrm{Gaussian}}(x_j|+) = 0.477$, $p_{\mathrm{Gaussian}}(x_j|-) = 0.763$, $p_{\mathrm{Gaussian}}(x_j|+) < p_{\mathrm{Gaussian}}(x_j|-)$. Then, $p_{\mathrm{DCK}}(x_j|+) = 0.659$, $p_{\mathrm{DCK}}(x_j|-) = 0.565$, $p_{\mathrm{DCK}}(x_j|+) > p_{\mathrm{DCK}}(x_j|-)$. In this example, the critical point $cp \approx 0.779$. We can find that there are two samples y_{1j} and y_4j satisfying the conditions $0 < \frac{|x_j - y_{1j}|}{h_{-j}} < cp$ and $0 < \frac{|x_j - y_{4j}|}{h_{-j}} < cp$. ∎

3) $0 \le \alpha_p < cp$, $\forall p \in \{1, 2, \cdots, n_+\}$, $\exists q \in \{1, 2, \cdots, n_-\}$, s.t. $\beta_q > 1$*; or* $0 \le \alpha_p < cp$, $\forall p \in \{1, 2, \cdots, n_+\}$, $\beta_q \ge cp$, $\forall q \in \{1, 2, \cdots, n_-\}$*:* When $\beta_q > 1$, $\mathrm{K}_{\mathrm{DCK}}(\beta_q) = 0$. So, $p_{\mathrm{Gaussian}}(x_j|-) > p_{\mathrm{DCK}}(x_j|-)$. And, $p_{\mathrm{Gaussian}}(x_j|+) < p_{\mathrm{DCK}}(x_j|+)$. Because when $0 \le \alpha_p < cp$, $\mathrm{K}_{\mathrm{Gaussian}}(\alpha_p) < \mathrm{K}_{\mathrm{DCK}}(\alpha_p)$. That is to say, when DCK is used, $p_{\mathrm{DCK}}(x_j|-)$ will decrease and $p_{\mathrm{DCK}}(x_j|+)$ will increase. Only the number of β_qs that $\mathrm{K}_{\mathrm{DCK}}(\beta_q) = 0$ is enough, the relationship $p_{\mathrm{DCK}}(x_j|+) > p_{\mathrm{DCK}}(x_j|-)$ will hold.

The above analyses and examples provide the opportunity for the usages of DLK and DCK. It tells us that under the

different application conditions, DLK and DCK can change the results obtianed by using traditional Gaussian kernel. However, the discontinuous kernel will limit the performance of FNB when classification task needs to be implemented. We call this limitation as the kernel incapability.

IV. The Kernel Incapability

The existing study [9] shows that DLK and DCK can be used to estimate the probability density function efficiently. However, the classification is different from the probability density estimation. The classification accuracy is always used as the evaluation criterion of a classification algorithm. So, in our study, we apply the classification accuracy to evaluate the performance of different kernels. In order to compare these kernels with Gaussian kernel, we need to use these six discontinuous kernels to classify the known sample firstly.

From Table I and Fig. 1, we can know that DLK and DCK can only calculate the values belonging to the interval $[-1, 1]$. For DLK and DCK, such kernel incapability will impose restriction on the determination performance of FNB. Now, we will describe the meaning of kernel incapability. Let $\mathrm{K_{DK}}(\cdot)$ denote the discontinuous kernel. For the new sample $x = \{x_1, x_2, \cdots, x_d\}$ in the two-class classification problem ($+$ class and $-$ class), the probability belonging to class $+$ is $p(x|+)$ (the prior probability $p(+) = p(-) = \frac{1}{2}$, and the probability belonging to class $-$ is $p(x|-)$. According to Eq. (3), we can get Eqs. (6) and (7) as follows:

$$p(x|+) = \prod_{j=1}^{d} p(x_j|+) = \prod_{j=1}^{d} \sum_{p=1}^{n_+} \frac{\mathrm{K_{DK}}\left(\frac{x_j - x_{pj}}{h_{+j}}\right)}{n_+ h_{+j}}, \tag{6}$$

and

$$p(x|-) = \prod_{j=1}^{d} p(x_j|-) = \prod_{j=1}^{d} \sum_{q=1}^{n_-} \frac{\mathrm{K_{DK}}\left(\frac{x_j - y_{qj}}{h_{-j}}\right)}{n_- h_{-j}}. \tag{7}$$

In $p(x|-)$, if $\exists j_1 \in \{1, 2, \cdots, d\}$, $s.t.$ $\frac{|x_{j_1} - x_{pj}|}{h_{+j}} \geq 1$ for all $p \in \{1, 2, \cdots, n_+\}$, then $\mathrm{K_{DK}}\left(\frac{|x_{j_1} - x_{pj}|}{h_{+j}}\right) = 0$ for all $p \in \{1, 2, \cdots, n_+\}$. So, $p(x_{j_1}|+) = 0$ and $p(x|+) = 0$. In the same way, for $p(x|-)$, if $\exists j_2 \in \{1, 2, \cdots, d\}$, $s.t.$ $\frac{|x_{j_2} - x_{qj}|}{h_{-j}} \geq 1$ for all $q \in \{1, 2, \cdots, n_-\}$, then $\mathrm{K_{DK}}\left(\frac{|x_{j_2} - x_{qj}|}{h_{-j}}\right) = 0$ for all $q \in \{1, 2, \cdots, n_-\}$. So, $p(x_{j_2}|+) = 0$ and $p(x|-) = 0$. For the new instance x, we get that the probabilities belonging to different classes are equal. FNB can not determine the class label for the new sample based on the obtained result $p(x|+) = p(x|-) = 0$. The result of $p(x|+) = p(x|-) = 0$ will never happen for Gaussian kernel because $\mathrm{K_{Gaussian}}(u)$ can always assign a sole, non-negative and real-valued real number for any $u \in [-\infty, +\infty]$. In order to solve the kernel incapability effectively, we propose an improved strategy named the equivalent probability. The procedures of this improved method can be described as following Algorithm 1. Assume the probability of a new sample x belonging to class w is represented as $p(x|w)$. Let X denote the current training dataset.

Algorithm 1 Computing the equivalent probability

```
p(x|w) = 1;
for j = 1 to d do
    p(x_j|w) = 0;
    h_j = 1/sqrt(n_w);
    for i = 1 to n_w do
        p(x_j|w) = p(x_j|w) + K_DK((|x_j - x_ij|)/h_j);
    end for
    p(x_j|w) = p(x_j|w)/(n_w h_j);
    if p(x_j|w) == 0 then
        d_min = min_{i=1,2,...,n_w} {|x_j - x_ij|};
        d_sum = sum_{i=1,2,...,n_w} {|x_j - x_ij|};
        p(x_j|w) = 1/(d_min × d_sum);
    end if
    p(x|w) = p(x|w) × p(x_j|w);
end for
```

In Algorithm 1, n_w is the number of samples in class w, and $p(x_j|w) = \frac{1}{d_{\min} \times d_{\mathrm{sum}}}$ is equivalent probability of x_j about dataset X. The usage of equivalent probability can deal with the kernel incapability to some extent. We apply a numeric experiment to exhibit the estimation performance of discontinuous kernel when equivalent probability is used. Firstly, we generate 10 data points $x_1, x_2, \cdots, x_{10}$ (all $x_i > 0$, $i = 1, 2, \cdots, 10$) which obey the distribution $N(0, 1)$. These points are used as the training samples in class $+$. The other 10 points $y_1, y_2, \cdots, y_{10}$ (all $y_i < 0$, $i = 1, 2, \cdots, 10$) are also generated from $N(0, 1)$ which are served as the training samples in class $-$. Then, the testing set is generated which includes 2 samples (1 sample x_+ belonging to class $+$ and 1 samples y_- belonging to class $-$). In order that the experiment can express our intention, we select the data points with the following constraints: all $|x_+ - x_j|$ and $|x_+ - y_j|$, $(j = 1, 2, \cdots, 10)$ are large than $\frac{1}{\sqrt{10}}$. And, all $|y_- - x_j|$ and $|y_- - y_j|$, $(j = 1, 2, \cdots, 10)$ are also larger than $\frac{1}{\sqrt{10}}$.

Our experiment is setup as follows: we want to use the equivalent probabilities to classify the testing data points x_+ and y_- based on the training data points $x_1, x_2, \cdots, x_{10}$ and $y_1, y_2, \cdots, y_{10}$. The probability of x_+ belonging to class $+$ is $p(x_+|+)$ and the probability belonging to class $-$ is $p(x_+|-)$. And, the probability of y_- belonging to class $+$ is $p(y_-|+)$ and the probability belonging to class $-$ is $p(y_-|-)$ so that $p(x_+|+) > p(x_+|-)$ and $p(y_-|+) < p(y_-|-)$. The above experiments are repeated five times independently. The detailed experimental results are summarized in Table II. Through the experiments, we can get that the equivalent probability is feasible when the kernel incapability occurs. The learning algorithm is able to classify the sample correctly by using the equivalent probability.

Next, we will discuss the computational complexity of using discontinuous kernels with the equivalent probability to carry

TABLE II
THE EXAMPLES FOR THE USAGE OF EQUIVALENT PROBABILITY

	Training										**Testing**
x	0.002	0.253	0.420	0.471	0.240	0.490	0.202	0.162	0.054	0.261	0.835
y	-0.886	-0.983	-0.014	-0.960	-0.917	-0.862	-0.671	-0.813	-0.007	-0.780	-0.350
x	0.370	0.383	0.482	0.523	0.406	0.394	0.545	0.313	0.139	0.234	0.898
y	-0.827	-0.644	-0.746	-0.671	-0.698	-0.643	-0.874	-0.752	-0.761	-0.727	-0.229
x	0.022	0.088	0.113	0.123	0.029	0.388	0.094	0.335	0.270	0.237	0.829
y	-0.956	-0.792	-0.124	-0.931	-0.899	-0.104	-0.951	-0.947	-0.772	-0.053	-0.443
x	0.331	0.299	0.395	0.151	0.179	0.314	0.269	0.423	0.234	0.361	0.745
y	-0.243	-0.053	-0.006	-0.132	-0.042	-0.070	-0.043	-0.348	-0.163	-0.327	-0.725
x	0.282	0.972	0.273	0.141	0.280	0.242	0.976	0.008	0.147	0.039	0.602
y	-0.082	-0.063	-0.952	-0.956	-0.141	-0.080	-0.027	-0.873	-0.929	-0.918	-0.501

out a new classification task. There are n training samples belonging to class c which has d feature attributes. When the probability of new sample x belonging to class c is calculated, the time consumption of Gaussian kernel is $O(nd)$. In the best situation, the discontinuous kernel can also reach this complexity $n \times d$. It indicates that the kernel incapability does not appear. The discontinuous kernel can determine the probability of new sample x that belongs to class c without the help of equivalent probability. The worst situation will happen when all calculations of $p(x_j|c)$ $(j = 1, 2, \cdots, d)$ have to rely on the equivalent probabilities. Under such circumstances, the complexity of discontinuous kernel is $3\times n\times d$, due to the extra processes needed for finding the minimum and computing the sum. In conclusion, the average complexity of discontinuous kernel with equivalent probability is also $O(nd)$. The usage of discontinuous kernel and introduction of equivalent probability do not dramatically increase the computational complexity of FNB.

V. THE EXPERIMENTAL OBSERVATIONS AND ANALYSES

In this part, we will discuss our experimental setup and results. And, based on the comparative results, the statistical analyses are also carried out.

A. The Data Preparation

In our comparative experiment, 15 UCI datasets [11] are selected which represent a wide range of domains and data characteristics. The detailed descriptions of datasets are listed in Table III. To the 15 UCI datasets, we adopted the following three preprocessing steps in our experiment:

1) Delete the nominal attributes: In our work, we mainly apply FNB to deal with classification problems of continuous attributes. We only want to investigate the effect imposed by different kernels on density estimation in FNB;
2) Replace the missing values: Any missing values in each dataset are replaced by running the unsupervised filter named *ReplaceMissingValues* in Weka. Its operation is: *weka.filters.unsupervised.attribute.ReplaceMissingValues* [15]. It replaces all missing values with the modes and means from the training data;
3) Reduce the large datasets. For saving the time of running experiments, the large dataset Magic Telescope is reduced randomly by using the unsupervised filter named *Resample* with the *sampleSizePercent* 10 in Weka. The implementation of this unsupervised filter is: *weka.filters.unsupervised.instance.Resample* [15].

B. The Experimental Procedures and Results

In order to eliminate the effect generated by splitting dataset randomly, we use 10 runs of 10-fold cross-validation procedure to implement our experiment. The experimental procedures are arranged as the following descriptions: Every dataset is randomly divided into 10 disjoint subsets, and the size of each subset is $N/10$, where N is the number of samples in this dataset. This procedure is run 10 times, each time using the different one of these subsets as the testing set and combining the other nine subsets for the training set. The testing accuracies are then averaged as the final classification accuracy. Every run of FNB with the different kernels is carried out on the same training sets and evaluated on the same testing sets. In particular, the folds of cross-validation are same for FNB (before and after the application of equivalent probability) on each dataset.

Our experiments include the following three parts. Firstly, we compare the classification performances of Gaussian kernel with the discontinuous kernels before using equivalent probability and after using equivalent probability respectively. The results are summarized in Table IV. In Table IV, the italic accuracies and standard derivations represent the performances of corresponding discontinuous kernels before using equivalent probability. Then, the corresponding kernels' $win/tie/lose$ records are compared by using the two-tailed t-test with 95 percent confidence level. Our tests include two parts: Gaussian kernel vs. discontinuous kernels before using the equivalent probability and Gaussian kernel vs. discontinuous kernels after using equivalent probability. The comparative results can be shown in Table V. Each three-number unit [16] $w/t/l$ in Table V means that the kernel in the corresponding row wins in w

TABLE III
THE DESCRIPTIONS OF DATASETS USED IN OUR EXPERIMENT

Datasets	The number of attributes	The number of classes	The distribution of classes	The number of samples
Blood transfusion	5	3	245/79/68	392
Credit approval	15	2	383/307	690
Cylinder bands	20	2	312/228	540
Ecoli	5	8	143/77/52/35/20/5/2/2	336
Glass identification	9	7	76/70/29/17/13/9/0	214
Heart disease	13	2	150/120	270
Iris	4	3	503	150
Magic telescope	10	2	12332/6688	19020(10%)
New thyroid gland	5	3	150/35/30	215
Parkinsons	22	2	147/48	195
Pima Indian diabetes	8	2	500/268	768
Sonar	60	2	111/97	208
Vehicle silhouettes	18	4	218/217/212/199	846
Vowel recognition	10	11	48×11	528
Wine	13	3	91/59/48	178

$$p\left(x\,|+\right)=\frac{1}{d_{\min}^{(+)}\times d_{\text{sum}}^{(+)}}=\frac{1}{\min\limits_{p=1,2,\cdots,n_+}\{|x-x_p|\}\times\sum\limits_{q=1,2,\cdots,n_+}\{|x-x_p|\}},\left(\forall p\in\{1,2,\cdots,n_+\},|x-x_p|>\frac{1}{\sqrt{n_+}}\right), \tag{8}$$

$$p\left(x\,|-\right)=\frac{1}{d_{\min}^{(-)}\times d_{\text{sum}}^{(-)}}=\frac{1}{\min\limits_{q=1,2,\cdots,n_-}\{|x-y_q|\}\times\sum\limits_{q=1,2,\cdots,n_-}\{|x-y_q|\}},\left(\forall q\in\{1,2,\cdots,n_-\},|x-x_q|>\frac{1}{\sqrt{n_-}}\right). \tag{9}$$

datasets, ties in t datasets, and loses in l datasets, against the kernel in the corresponding column. The italic $w/t/l$ records are the comparisons of discontinuous kernels before using equivalent probability. Finally, we validate the accuracy increments on these 15 datasets before and after equivalent probability is applied in discontinuous kernel by comparing with Gaussian kernel. Based on the above experimental results, the intuitive observations and theoretical analysis are given.

From Tables IV and V we can see that when the equivalent probabilities are not used, the discontinuous kernels can not obtain the significant classification performances compared with Gaussian kernel (with $w/t/l$ records 6/1/8, 7/4/4, 6/3/5, 6/3/6, 6/2/7 and 6/3/6). However, after using the equivalent probabilities, the performances of discontinuous kernels are significantly better than Gaussian kernel (with $w/t/l$ records 7/5/3, 11/3/1, 11/3/1, 9/3/3, 9/2/4 and 9/4/2). From this comparison we can know that the usage of equivalent probabilities can obviously improve the classification accuracy of discontinuous kernel.

Based on the above comparisons in Tables IV and V, we also compute the increment of accuracy of different kernels on 15 UCI datasets. On a specific dataset, if the discontinuous kernel is statistically better than Gaussian, then the accuracy difference between the discontinuous kernel and Gaussian kernel will be added into the increment of the discontinuous kernel; otherwise, the accuracy difference between Gaussian kernel and the discontinuous kernel will be added into the increment of Gaussian kernel. The comparative increments are summarized in Fig. 2. The left picture shows what happened when the equivalent probabilities are not used. And, the right one tells us the comparative results after the equivalent probabilities are adopted. From the left picture listed in Fig. 2, we find an interesting result: the accuracy increments of discontinuous kernels compared with Gaussian kernel are obviously superior except uniform kernel. According to statistical result, we know that when equivalent probability is not used, the performances of discontinuous kernels and Gaussian kernel are basically comparable (6 vs. 8, 7 vs. 4, 6 vs. 5, 6 vs. 6, 6 vs. 7, and 6 vs. 6). The increase of discontinuous kernels compared with Gaussian kernel is obviously larger than the increase of Gaussian kernel compared with discontinuous kernels. That is to say, even though the equivalent probability is not used, Gaussian kernel also cannot obtain the best classification accuracy. When the equivalent probabilities are used by the discontinuous kernels, the right picture in Fig. 2 tells us that the increments of discontinuous kernels become more obvious. For instance, before the equivalent probability is used, the comparative increment of Gaussian vs. epanechnikov is 0.123 vs. 0.267. While the equivalent probability is used by epanechnikov kernel, this result becomes 0.006 vs. 0.328. This comparison shows us that most commonly used Gaussian kernel can not reach the satisfactory classification accuracy.

Now, we will give the following explanation for the advantage of equivalent probability. Let the equivalent probabilities of new sample x belonging to class $+$ and class $-$ be $p(x|+)$ and $p(x|-)$. Accordingly, Eqs. (8) and (9) list the mathematical expressions of $p(x|+)$ and $p(x|-)$.

The class attribute of new sample x is deemed as $+$. When

TABLE IV
THE DETAILED EXPERIMENTAL RESULTS ON THE CLASSIFICATION ACCURACY AND STANDARD DEVIATION

Datasets	Kernel functions						
	Gaussian	Uniform	Triangular	Epanechnikov	Biweight	Triweight	Cosine
Blood transfusion	0.704±0.003	*0.742±0.004*	*0.708±0.006*	*0.704±0.005*	*0.694±0.005*	*0.693±0.004*	*0.702±0.007*
		0.743±0.004	0.709±0.006	0.705±0.005	0.694±0.005	0.695±0.004	0.702±0.006
Credit approval	0.711±0.003	*0.751±0.003*	*0.761±0.004*	*0.756±0.002*	*0.761±0.004*	*0.761±0.003*	*0.757±0.003*
		0.752±0.003	0.762±0.004	0.757±0.002	0.762±0.004	0.761±0.003	0.758±0.003
Cylinder bands	0.711±0.006	*0.692±0.009*	*0.706±0.009*	*0.703±0.009*	*0.701±0.009*	*0.701±0.009*	*0.705±0.007*
		0.711±0.007	0.729±0.007	0.728±0.006	0.725±0.005	0.726±0.005	0.729±0.003
Ecoli	0.850±0.004	*0.862±0.006*	*0.859±0.004*	*0.858±0.006*	*0.856±0.004*	*0.853±0.005*	*0.858±0.004*
		0.864±0.005	0.861±0.004	0.861±0.006	0.858±0.004	0.852±0.006	0.860±0.004
Glass identification	0.592±0.016	*0.535±0.023*	*0.624±0.012*	*0.589±0.011*	*0.620±0.010*	*0.631±0.013*	*0.591±0.010*
		0.533±0.023	0.623±0.013	0.592±0.013	0.617±0.010	0.631±0.013	0.594±0.011
Heart disease	0.841±0.005	*0.829±0.007*	*0.826±0.010*	*0.830±0.006*	*0.823±0.010*	*0.821±0.010*	*0.829±0.007*
		0.837±0.005	0.834±0.009	0.839±0.006	0.831±0.010	0.829±0.009	0.837±0.006
Iris	0.957±0.005	*0.916±0.007*	*0.913±0.007*	*0.913±0.008*	*0.913±0.008*	*0.910±0.008*	*0.914±0.009*
		0.954±0.005	0.947±0.007	0.951±0.005	0.948±0.006	0.941±0.005	0.951±0.005
Magic telescope	0.761±0.004	*0.772±0.005*	*0.769±0.009*	*0.772±0.008*	*0.772±0.007*	*0.764±0.009*	*0.770±0.007*
		0.773±0.004	0.771±0.009	0.773±0.007	0.773±0.007	0.765±0.007	0.771±0.007
New thyroid gland	0.912±0.000	*0.935±0.004*	*0.954±0.006*	*0.948±0.004*	*0.954±0.006*	*0.959±0.004*	*0.948±0.004*
		0.941±0.002	0.959±0.005	0.954±0.002	0.960±0.005	0.964±0.003	0.954±0.002
Parkinsons	0.813±0.006	*0.825±0.011*	*0.804±0.012*	*0.816±0.013*	*0.802±0.010*	*0.794±0.006*	*0.815±0.013*
		0.834±0.012	0.809±0.013	0.822±0.014	0.807±0.012	0.798±0.009	0.821±0.013
Pima Indian diabetes	0.742±0.005	*0.704±0.004*	*0.743±0.005*	*0.743±0.005*	*0.743±0.005*	*0.742±0.006*	*0.744±0.005*
		0.707±0.004	0.748±0.005	0.748±0.005	0.749±0.006	0.749±0.006	0.749±0.004
Sonar	0.768±0.008	*0.722±0.013*	*0.720±0.016*	*0.718±0.015*	*0.720±0.018*	*0.724±0.015*	*0.719±0.013*
		0.775±0.011	0.767±0.012	0.766±0.011	0.768±0.013	0.772±0.014	0.767±0.009
Vehicle silhouettes	0.517±0.005	*0.525±0.006*	*0.542±0.006*	*0.532±0.007*	*0.541±0.007*	*0.548±0.006*	*0.534±0.008*
		0.525±0.005	0.543±0.006	0.533±0.008	0.542±0.006	0.549±0.006	0.535±0.009
Vowel recognition	0.573±0.011	*0.551±0.012*	*0.776±0.012*	*0.730±0.008*	*0.785±0.012*	*0.809±0.012*	*0.743±0.008*
		0.552±0.012	0.778±0.012	0.732±0.008	0.787±0.012	0.810±0.012	0.745±0.008
Wine	0.958±0.003	*0.944±0.004*	*0.952±0.003*	*0.948±0.004*	*0.952±0.003*	*0.953±0.004*	*0.948±0.004*
		0.962±0.005	0.971±0.002	0.967±0.004	0.971±0.002	0.968±0.004	0.967±0.004

$d_{\text{sum}}^{(+)} = d_{\text{sum}}^{(-)}$, the class of x will be determined by $d_{\text{min}}^{(+)}$ and $d_{\text{min}}^{(-)}$. If $p(x|+) > p(x|-)$, that is to say the new sample can be classified correctly. Then $d_{\text{min}}^{(+)} < d_{\text{min}}^{(-)}$ can be derived by observing the above Eqs. (8) and (9). It indicates that the sum of distances between x and x_p, $(p = 1, 2, \cdots, n_+)$ is smaller than the sum of distances between x and y_q, $(q = 1, 2, \cdots, n_-)$. The global similarity between x and the samples belonging to class $+$ is higher. So, it is feasible that the new sample x is classified in the class $+$. When $d_{\text{min}}^{(+)} = d_{\text{min}}^{(-)}$, we can easily get $d_{\text{sum}}^{(+)} < d_{\text{sum}}^{(-)}$ from Eqs. (8) and (9). It indicates that the minimum of distances between x and x_p, $(p = 1, 2, \cdots, n_+)$ is smaller than the minimum of distances between x and y_q, $(q = 1, 2, \cdots, n_-)$. Therefore, The local similarity between x and the samples belonging to class $+$ is higher. The equivalent probability $\frac{1}{d_{\text{min}} \times d_{\text{sum}}}$ can guarantee the global and local similarities between x and the samples in class $+$ are all optimal. When the kernel incapability occurs, it is possible that the global and local optimizations of equivalent probability can correctly classify the unknown instance.

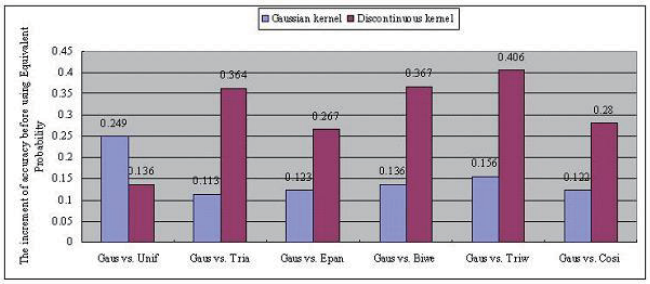
From the comparative results, we can also find another fact that the Gaussian kernel can not obtain the significantly better accuracies than that of the discontinuous kernels when the equivalent probabilities are used. The reason is that the discontinuous kernels can classify the unknown sample in a more flexible manner. The flexibility of discontinuous kernel can be expressed from the numeric experiments summarized in Table II. When Gaussian kernel does not determine the class attribute for the unknown sample, the discontinuous kernels using equivalent probabilities can classify it correctly. There are several possibilities that the discontinuous kernels can correct the classification result of Gaussian kernel as described in Section III.

VI. CONCLUSION

In this paper, the discontinuous kernels (uniform, triangular, epanechnikov, biweight, triweight and cosine) are firstly introduced into FNB and by analysing the disadvantages of discontinuous kernels an efficient strategy named equivalent probability is proposed to improve the classification accuracies

TABLE V
THE DETAILED EXPERIMENTAL RESULTS ON THE CLASSIFICATION ACCURACY AND STANDARD DEVIATION

$w/t/l$	**Gaussian**	**Uniform**	**Triangular**	**Epanechnikov**	**Biweight**	**Triweight**
Uniform	(*6/1/8*) 7/5/3					
Triangular	(*7/4/4*) 11/3/1	(*8/3/4*) 8/3/4				
Epanechnikov	(*6/3/5*) 11/3/1	(*7/6/2*) 8/4/3	(*1/8/6*) 1/8/6			
Biweight	(*6/3/6*) 9/3/3	(*7/2/5*) 8/1/5	(*1/13/1*) 1/13/1	(*5/7/3*) 7/5/3		
Triweight	(*6/2/7*) 9/2/4	(*7/3/5*) 8/2/5	(*3/9/3*) 3/8/4	(*5/7/3*) 4/6/5	(*3/12/0*) 3/10/2	
Cosine	(*6/3/6*) 9/4/2	(*7/5/3*) 8/3/4	(*1/9/5*) 1/9/5	(*1/14/0*) 1/14/0	(*3/7/5*) 3/7/5	(*3/7/5*) 5/6/4



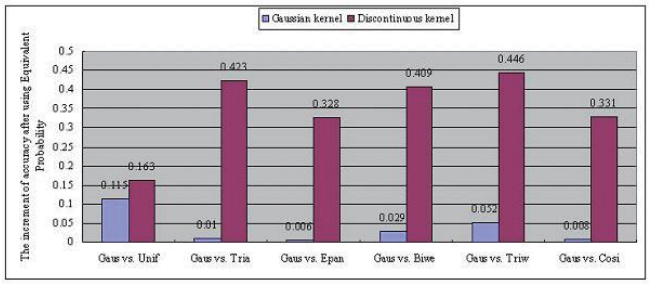


Fig. 2. The comparisons of accuracy increment on 15 UCI datasets before and after using equivalent probability

of discontinuous kernels in FNB. The experimental results show that (1) the most frequently used Gaussian kernel can not obtain the statistically best accuracy; (2) the equivalent probabilities indeed improve the classification accuracies of discontinuous kernels significantly. In the future study, the appropriate data distribution character for every discontinuous kernel will be investigated and the discontinuous kernels will also be introduced into the other data mining systems (e.g., [17], [18], [19], [20], [21], [22], and [23]) in order to improve their learning performances.

ACKNOWLEDGMENT

The authors would like to thank three anonymous reviewers for their constructive comments on the earlier version of this manuscript. This work is in part supported by GRF grant 5237/08E, CRG grant G-U756 of The Hong Kong Polytechnic University, The National Natural Science Foundation of China 61170040.

REFERENCES

[1] T. M. Mitchell, Machine learning, McGraw Hill, 1997.
[2] G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," In Proc. 1995 Conf. Uncertain. Artif. Intell., San Mateo, pp. 338-345, 18-20 Aug. 1995.
[3] E. Parzen, "On estimation of a probability density function and mode," Ann. Math. Stat., vol. 33, no. 3, pp. 1065-1076, 1962.
[4] G. R. Terrell and D. W. Scott, "Oversmoothed nonparametric density estimates," J. Amer. Statist. Assoc., vol. 80, no. 389, pp. 209-214, Mar. 1985.
[5] G. R. Terrell, "The maximal smoothing principle in density estimation," J. Amer. Statist. Assoc., vol. 85, no. 410, pp. 470-477, Jun. 1990.
[6] M. P. Wand and M. C. Jones, "Comparison of smoothing parameterizations in bivariate kernel density estimation," J. Amer. Statist. Assoc., vol. 88, no. 422, pp. 520-528, Jun. 1993.
[7] A. Pérez, P. Larrañaga, and I. Inza, "Bayesian classifiers based on kernel density estimation: Flexible classifiers," Int. J. Approx. Reason., vol. 50, no. 2, pp. 341-362, Feb. 2009.
[8] B. Liu, Y. Yang, G. I. Webb, and J. Boughton, "A comparative study of bandwidth choice in kernel density estimation for naïve Bayesian classification, Lect. Notes Comput. Sci., vol. 5476, pp. 302-313, 2009.
[9] M. P. Wand and M. C. Jones, Kernel smoothing. Chapman and Hall, 1995.
[10] P. Domingos and M. Pazzani, "On the optimality of the simple Bayesian classifier under zero-oneloss," J. Mach. Learn., vol. 29, pp. 103-130, 1997.
[11] The UC Irvine Machine Learning Repository, [Online]. Available: http://archive.ics.uci.edu/ml/datasets.html.
[12] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, f-score and ROC: a family of discriminant measures for performance evaluation, Lect. Notes Comput. Sci., vol. 4304, pp. 1015-1021, 2006.
[13] J. E. Freund, Modern elementary statistics, Prentice Hall, 1984.
[14] M. G. Genton, "Classes of kernels for machine learning: a statistics perspective," J. Mach. Learn. Res., vol. 2, pp. 299-312, Dec. 2001.
[15] I. H. Witten and E. Frank, Data mining: practical machine learning tools and techniques, second edition, Morgan Kaufmann, 2005.
[16] J. Demšar, "statistical comparisons of classifiers over multiple data sets," J. Mach. Learn. Res., vol. 7, pp. 1-30, Jan. 2006.
[17] X. Z. Wang and C. R. Dong, "Improving generalization of fuzzy if-then rules by maximizing fuzzy entropy," IEEE Trans. Fuzzy Syst., vol. 17, no. 3, pp. 556-567, Jun. 2009.
[18] X. Z. Wang, L. C. Dong, and J. H. Yan, "Maximum ambiguity based sample selection in fuzzy decision tree induction," IEEE Trans. Knowl. Data Eng., DOI: 10.1109/ TKDE.2011.67, 2011.
[19] X. Z. Wang, Y. L. He, L. C. Dong, and H. Y. Zhao, "Particle swarm optimization for determining fuzzy measures from data," Inf. Sci., vol. 181, no. 19, pp. 4230-4252, 2011.
[20] L. J. Wang, "An improved multiple fuzzy NNC system based on mutual information and fuzzy integral," Int. J. Mach. Learn. & Cyber., vol. 2, no. 1, pp. 25-36, Mar. 2011.
[21] Q. He and C. X. Wu, "Separating theorem of samples in banach space for support vector machine learning," Int. J. Mach. Learn. & Cyber., vol. 2, no. 1, pp. 49-54, Mar. 2011.
[22] G. B. Huang, D. H. Wang, and Y. Lan, "Extreme learning machines: a survey," Int. J. Mach. Learn. & Cyber., vol. 2, no. 2, pp. 107-122, 2011.
[23] M. Dobrska, H. Wang and W. Blackburn, "Ordinal regression with continuous pairwise preferences," Int. J. Mach. Learn. & Cyber., vol. 3, no. 1, pp. 59-70, Mar. 2012.