A Comparative Study of Bandwidth Choice in Kernel Density Estimation for Naive Bayesian Classification

Bin Liu, Ying Yang, Geoffrey I. Webb and Janice Boughton

Clayton School of Information Technology, Monash University, Australia
{bin.liu, ying.yang, geoff.webb, janice.boughton}@infotech.monash.edu.au

Abstract. Kernel density estimation (KDE) is an important method in nonparametric learning. While KDE has been studied extensively in the context of accuracy of density estimation, it has not been studied extensively in the context of classification. This paper studies nine bandwidth selection schemes for kernel density estimation in Naive Bayesian classification context, using 52 machine learning benchmark datasets. The contributions of this paper are threefold. First, it shows that some commonly used and very sophisticated bandwidth selection schemes do not give good performance in Naive Bayes. Surprisingly, some very simple bandwidth selection schemes give statistically significantly better performance. Second, it shows that kernel density estimation can achieve statistically significantly better classification performance than a commonly used discretization method in Naive Bayes, but only when appropriate bandwidth selection schemes are applied. Third, this study gives bandwidth distribution patterns for the investigated bandwidth selection schemes.

1 Introduction

A critical task in Bayesian learning is estimation of the probability distributions of attributes in datasets, especially when the attributes are numeric. Traditionally, the numeric attributes are handled by discretization [1]. These methods are usually simple and computationally efficient. However, they suffer from some basic limitations [2, 3]. An alternative to calculating probability estimates for numeric attributes using discretized intervals is to estimate the probabilities directly, using an estimate of the point-wise density distribution. Both parametric and nonparametric density estimation methods have been developed.

Parametric density estimation imposes a parametric model on the observations. For example, the parameters for a Gaussian model are its sufficient statistics, the mean and variance. Normally simple parametric models do not work very well with Bayesian classification [4], as the real distributions do not exactly fit specific parametric models.

Some estimation methods, including Gaussian mixture models, use subsets of the data to obtain local fitting models, then mix these models to obtain the density estimate for all observations. In contrast, Kernel Density Estimation estimates the probability density function by imposing a model function on every data point and then adding them together. The function applied to each data point is called a kernel function. For example, a Gaussian function can be imposed on every single data point, making the center of each Gaussian kernel function the data point that it is based on. The standard deviation of the Gaussian kernel function adjusts the dispersion of the function and is called a *bandwidth* of the function.

Given sufficiently large sample data, KDE can converge to a reasonable estimate of the probability density. As there are no specific finite parameters imposed on the observations, KDE is a nonparametric method.

The univariate KDE [5, 6] can be expressed as:

$$f(x) = \frac{1}{nh} \sum_{i=1}^{n} \mathbf{K}\left(\frac{x - X_i}{h}\right), \tag{1}$$

where $\mathbf{K}(.)$ is the density kernel; x is a test instance point; X_i is a training instance point, which controls the position of the kernel function; h is the bandwidth of the kernel, which controls the dispersion of each kernel; and n is the number of data points in the data. For a univariate Gaussian kernel $\mathbf{K}(\xi) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\xi^2}{2}}$.

Naive Bayes is a widely employed effective and efficient approach for classification learning, in which the class label $y(x)$ of a test instance x is evaluated by $y(x) = \mathbf{argmax}\ _c \left[P(c) \times \prod_{i=1}^{d} P(x_i \mid c)\right]$, where $P(c)$ is a class probability, d is the number of attributes, x_i is the i'th attribute of instance x, and $P(x_i \mid c)$ is the probability (or probability density) of x_i given the class. KDE (Equation (1)) can be used to estimate the class conditional probabilities for numeric attributes. Because the Naive Bayesian classifier considers each attribute independently, we use only univariate kernel density estimation in this paper.

It is known that the specific choice of kernel function $\mathbf{K}$ is not critical [7]. The key challenge is the choice of the bandwidth. A bandwidth value which is too small will give a too detailed curve and hence leads to an estimation with small bias and large variance. Large bandwidth leads to low variance at the expense of increased bias.

Many bandwidth selection schemes in kernel density estimation have been studied mainly for optimizing the mean squared error loss of the estimation which supports good density curve fitting. However, bandwidth selection schemes are still not extensively studied in the classification context applying 0-1 loss criteria.

We look at the seven most commonly used bandwidth selection schemes in the statistical community plus two very simple schemes, using 52 datasets. It is shown that the choice of bandwidth dramatically affects the accuracy results of classification. An appropriate bandwidth selection scheme can archive statistically significantly better classification performance than a commonly used discretization method. Surprisingly, the two simple bandwidth selection schemes both achieved good performance, whereas the more sophisticated and computationally expensive schemes delivered no improvement in classification performance.

2 Bandwidth Selection Schemes

Background Intuitively, it is assumed that there is a positive correlation between the accuracy of the probability estimates and the accuracy of classification. Friedman [8] challenged this assumption and states that more accurate probability estimates do not necessarily lead to better classification performance and can often make it worse.

Unfortunately, most bandwidth selection research considers the assumption to be true and attempts to achieve the highest possible probability estimation accuracy. These schemes are often based on a mean squared error (MSE) criteria, instead of a 0-1 loss criteria.

To the best of our knowledge, there is no practical bandwidth selection scheme that focuses on improving the classification accuracy, rather than the accuracy of the probability estimates. A recent paper [9] explores the theory of bandwidth choice in classification under limited conditions. It states that the optimal size of the bandwidth for 0-1 loss based estimation is generally the same as that which is appropriate for squared error based estimation.

Generally speaking, KDE bandwidth choice in the context of classification under 0-1 loss is a more difficult one compared with bandwidth choice under MSE loss. For example, we consider using Cross-Validation to chose optimal bandwidths in Naive Bayes, using class labels as the supervised information. Every evaluation under 0-1 loss (according to the class label) should use all attributes in the dataset. This is a global optimization problem in which the optimal bandwidth for one attribute may interact with those for other attributs. It is different to the MSE criteria which only uses the attribute under consideration.

In this section we give some theoretical descriptions of the mean squared error criteria and describe 7 bandwidth selection schemes that are based on this criteria. We also discuss two schemes which are not theoretically related to MSE.

Mean Squared Error Criteria In probability density estimation, the Mean Squared Error (MSE) or Mean Integrated Squared Error (MISE) are the most used density estimation error criteria,

$$MISE(\hat{f}) = E\int[\hat{f}(x) - f(x)]^2 dx\,, \tag{2}$$

where integral is in the range of x, to measure how well the entire estimated curve $\hat{f}$ approximates the real curve f. The expectation operation averages over all possible samplings. From this equation, we can get $MISE(\hat{f}) = \int Bias^2[\hat{f}(x)]dx + \int Var[\hat{f}(x)]dx$, where $Bias[\hat{f}(x)] = E[\hat{f}(x)] - f(x)$ and $Var[\hat{f}(x)] = E[f^2(x)] - E^2[\hat{f}(x)]$. This equation is the starting point of the bandwidth selection scheme UCV we discuss below.

We process $E[\hat{f}(x)]$ first by using Equation (1). This leads to $E[\hat{f}(x)] = E\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}\mathbf{K}(\frac{x-X_i}{h})\right] = E\left[\frac{1}{h}\mathbf{K}(\frac{x-X}{h})\right] = \int\frac{1}{h}\mathbf{K}(\frac{x-y}{h})f(y)dy$, where for each test point x, we regard each X_i as an independent and identically distributed random variable with distribution f. Making a simple variable substitution $y = x - ht$, we obtain: $Bias[\hat{f}(x)] = \int\mathbf{K}(t)[f(x-ht) - f(x)]dt\,.$ A Taylor series expansion $f(x-ht) \approx f(x) - htf'(x) + \frac{1}{2}h^2t^2f''(x)$ can be substituted into this equation. The first term of $f(x-ht)$ is canceled out by the negative $f(x)$. The second term is also canceled out because the $K(t)$ in the integral is a symmetric function. So, $\int Bias^2[\hat{f}(x)]dx \approx \frac{1}{4}h^4(\int t^2K(t)dt)^2\int(f''(x))^2dx = \frac{1}{4}h^4\mu_2^2(K)R(f'')\,,$ where $R(g) = \int g^2(x)dx$ and $\mu_2(g) = \int x^2g(x)dx$.

In a similar way, we can get, $Var[\hat{f}(x)] = \frac{1}{nh}R(K)\,.$ The elementary Equation (2) becomes an asymptotic form, as the error term in Taylor expansion is the higher-order term of h, which monotonously decreases when samples grow. The asymptotic mean integrated squared error is,

$$AMISE = \frac{1}{nh}R(K) + \frac{1}{4}h^4\mu_2^2(K)R(f'')\,. \tag{3}$$

This equation is the starting point for the bandwidth selection schemes BCV, STE and DPI, which are discussed below.

Unbiased Cross-Validation (UCV) Scheme The method of Unbiased Cross-Validation [10] is based on the elementary Equation (2). It is also called least squares cross-validation. UCV obtains a score function to estimate the performance of candidate bandwidth. In practice, UCV minimizes the integrated square error, the Equation (4), which uses one realization of samples from underlaying distribution f.

$$ISE = \int[\hat{f}(x) - f(x)]^2dx = R(\hat{f}) - 2\int\hat{f}(x)f(x)dx + R(f)\,, \tag{4}$$

where $R(g)$ is similar to Equation (3).

Notice the first term in Equation (4) is only related to the estimated $\hat{f}(x)$, so it is easy to process given a specific bandwidth $\hat{h}$. The third term is independent of the estimated $\hat{h}$ and remains constant for all estimations, so it can be ignored. The second term can be written as $\int\hat{f}(x)f(x)dx = E[\hat{f}(x)]$, i.e., it is the statistic mean of $\hat{f}(x)$ with respect to x.

If we get n samples of x, for the sake of obtaining a stable estimation of $E[\hat{f}(x)]$, we can use a Leave-One-Out method to get an n-points estimation value of the $\hat{f}(x)$. The Leave-One-Out method estimates the value of $\hat{f}(x_i)$ by leaving the x_i out and using the other n-1 points of x. This is why this method is called a Cross-Validation. We use $\hat{f}_{-i}(x_i)$ to express this Leave-One-Out estimation, which is evaluated from Equation (1). Then, $E[\hat{f}(x)] = \frac{1}{n}\sum_{i=1}^{n}\hat{f}_{-i}(x_i)$. Substituting this to Equation (4) we construct a score function in the sense of ISE. Now for some specific candidate bandwidth $\hat{h}$, we can give a unbiased cross validation score for the candidate bandwidth $\hat{h}$ as,

$$UCV(\hat{h}) = R(\hat{f}) - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{-i}(x_i)\,.$$

We can use a start bandwidth as a reference estimation, and make a brute-force search near this reference bandwidth with respect to the minima of UCV score function.

Normal Reference Density (NRD-I, NRD and NRD0) schemes Normal Reference Density [5] scheme is also called the Rule of Thumb scheme. It is based on Equation (3). To minimize AMISE, a simple first order differential can be used on Equation (3) towards the bandwidth h and setting the differential to zero. The optimal bandwidth is:

$$\hat{h}_{AMISE} = \left[\frac{R(K)}{\mu_2^2(K)R(f'')}\right]^{1/5} n^{-1/5}\,. \tag{5}$$

This result still depends on the unknown density derivative function $f''(x)$, which will depend on h recursively again. Normal Reference Density scheme simplifies this problem by using a parametric model, say, a Gaussian to estimate $f''(x)$. Compared with the Cross-validation selection, this is a straightforward method and can lead to an analytical expression of bandwidth $\hat{h} = 1.06\,\hat{\sigma}n^{-1/5}$, where n is the number of samples and $\hat{\sigma}$ is the estimated normal distribution standard deviation of the samples.

This bandwidth selection scheme is a classic one. We use this bandwidth as a standard bandwidth in our experiments. We call this scheme *NRD-I*.

A more robust approach [5] can be applied by considering the interquartile range (IQR). The bandwidth is calculated from the minimum of standard deviation and standard IQR: $\hat{h} = 1.06\,\mathbf{min}\,(\hat{\sigma}, IQR/1.34)\,n^{-1/5}$. This procedure [5] helps to lessen the risk of oversmoothing. We call the bandwidth the NRD bandwidth in this paper. A smaller version of NRD suggested in R [11] is $\hat{h} = 0.9\,\mathbf{min}\,(\hat{\sigma}, IQR/1.34)\,n^{-1/5}$. We call this bandwidth the NRD0.

Biased Cross-Validation (BCV) Scheme Biased cross-validation uses Equation (3) as the basis of the score function. Scott and Terrel [12] develop an estimation of $R(f'')$ in Equation (3), using $\hat{R}(f'') = R(\hat{f}'') - \frac{1}{nh^5}R(K'')$, where f'', $\hat{f}''$ and K'' are second-order derivatives of distribution and kernel respectively. The right hand side of this estimation can be evaluated given a specific bandwidth $\hat{h}$. Substituting the $\hat{R}(f'')$ to Equation (3), we can get a new score function,

$$BCV(\hat{h}) = \frac{1}{nh}R(K) + \frac{1}{4}h^4\mu_2^2(K)[R(\hat{f}'') - \frac{1}{nh^5}R(K'')]\,.$$

A exhaustive search procedure similar to UCV scheme can be applied to find the optimal bandwidth.

Direct-Plug-In (DPI) Scheme and Solve-The-Equation (STE) Scheme The Direct-Plug-In scheme [13] is a more complicated version of the Normal Reference Density scheme. It seeks $R(f'')$ by estimation of $R(f'''')$. This problem continues because the $R(f^{(s)})$ will depend on $R(f^{(s+2)})$. Normally, for a specific s, $R(f^{(s+2)})$ is estimated by a simple parametric method, to obtain $R(f^{(s)})$ and so on. We call Direct-Plug-In Scheme the DPI in our experiments.

Notice that Equation (5) is a fixed point equation $h = F(h)$, where $F(h) = \left[\frac{R(K)}{\mu_2^2(K)R(f'')}\right]^{1/5} n^{-1/5}$. and $R(f'')$ is a function of h. Solve-The-Equation Scheme [6, 13] is applied by solving the fixed point of $F(h)$. We call Solve-The-Equation scheme the STE in our experiments.

Two Very Simple (WEKA and SP) Schemes We use two very simple bandwidth selection schemes. These two schemes are both based on the range of data divided by a measure of the size of the samples. There is less theoretical consideration [4, 14, 15] of these methods compared with the other methods discussed above. They merely conform to the basic requirement in KDE that when the number of samples approaches infinity, the bandwidth approaches zero.

One scheme uses $\sqrt{n}$ as a division factor [4], so the bandwidth approaches zero as n increases,

$$\hat{h} = \frac{range(x)}{\sqrt{n}} ,$$

where n is the number of samples, $range(x)$ is the range of values of x in training data. This scheme is used in WEKA [14], with some calibration that $\hat{h}$ should be no less than $\frac{1}{6}$ of the average data interval, which avoids $\hat{h}$ becoming too small compared with the average data interval. We call this scheme WEKA.

The other scheme is a very old scheme[16].

$$\hat{h} = \frac{range(x)}{2(1 + log_2 n)} .$$

The basic principle of this equation does not have very strong theoretical basis [15]. However it was widely used in the old version of S-PLUS statistic package (up to version 5.0) [17, page 135]. We call it the SP scheme.

3 Experiments

3.1 Data and Design

In addition to the nine bandwidth selection schemes described in Section 2, the widely used MDL discretization method [1] was also used as a performance reference. The Naive Bayesian classifier was the classifier used for all schemes being evaluated. Every statistic sample (every dataset, every experiment trial and fold) and every piece of classifier code is the same for all schemes. The only difference between each scheme in the classifier algorithm is the bandwidth of the kernel.

Table 1. The 52 experimental datasets, with the numbers of instances, classes, attributes and numeric attributes.

Data	Ins.	Cls.	Att.	NAtt.	Data	Ins.	Cls.	Att.	NAtt.
Abalone	4177	3	8	8	Letter	20000	26	16	16
Adult	48842	2	14	6	Liver-disorders	345	2	6	6
Anneal	898	6	38	6	Lymph	148	4	18	3
Arrhythmia	452	16	279	206	Mfeat-factors	2000	10	216	216
Autos	205	7	25	15	Mfeat-fourier	2000	10	76	76
Backache	180	2	32	6	Mfeat-karhunen	2000	10	64	64
Balance-scale	625	3	4	4	Mfeat-morphological	2000	10	6	6
Biomed	209	2	8	7	Mfeat-zernike	2000	10	47	47
Cars	406	3	7	6	New-thyroid	215	3	5	5
Cmc	1473	3	9	2	Optdigits	5620	10	64	64
Collins	500	15	23	20	Page-blocks	5473	5	10	10
German	1000	2	20	7	Pendigits	10992	10	16	16
Crx(credit-a)	690	2	15	6	Prnn-synth	250	2	2	2
Cylinder-bands	540	2	39	18	Satellite	6435	6	36	36
Diabetes	768	2	8	8	Schizo	340	2	14	12
Echocardiogram	131	2	6	5	Segment	2310	7	19	19
Ecoli	336	8	7	7	Sign	12546	3	8	8
Glass	214	7	9	9	Sonar	208	2	60	60
Haberman	306	2	3	2	Spambase	4601	2	57	57
Heart-statlog	270	2	13	13	Syncon	600	6	61	60
Hepatitis	155	2	20	6	Tae	151	3	5	3
Horse-colic	368	2	21	8	Vehicle	846	4	18	18
Hungarian	294	2	13	6	Vowel	990	11	13	10
Hypothyroid	3772	4	29	7	Waveform-5000	5000	3	40	40
Ionosphere	351	2	34	34	Wine	178	3	13	13
Iris	150	3	4	4	Zoo	101	7	17	1

The fifty-two datasets used in the experiments were drawn from the UCI machine learning repository [18] and the web site of WEKA [14]. We use all the datasets that we could identify from these places, given the dataset has at least one numeric attribute and has at least 100 instances. Table 1 describes these datasets. Any missing values occurring in the data for numeric attributes were replaced with the mean average value for that attribute.

Each scheme was tested on each dataset using a 30-trial 2-fold cross validation bias-variance decomposition. A large number of trials was chosen because bias-variance decomposition has greater accuracy when a sufficiently large number of trials are conducted [19]. Selecting two folds for the cross-validation maximizes the variation in the training data from trial to trial.

Thirty trials and two folds yields sixty Naive Bayesian classification evaluations for each dataset. For these evaluations we recorded the mean training time, mean error rate, mean bias and mean variance. Kohavi and Wolpert's method [20] of bias and variance decomposition was employed to determine the bias and variance based on the obtained error rate.

Since there are nine alternative KDE classifiers and one discretization classifier, we get ten comparators of the performance measure for each dataset.

After the classification performance comparison, we also produce a statistic for the bandwidth distribution for alternative bandwidth selection schemes. The fifty-two datasets contain 1294 numeric attributes collectively. Every numeric attribute has at least two and at most 26 class labels. Since we evaluate the KDE for every class conditional probability, there are 10967 class conditional probability evaluation objects. Each of these evaluation objects produces 60 different realization samples by the 30 trails 2 fold cross-validation. Every bandwidth selection scheme is applied to each realization of the conditional probability evaluation objects, and produces an estimated bandwidth for that realization.

These bandwidths are transformed to a ratio to a standard bandwidth. We use the NRD-I bandwidth as the standard. By using these bandwidth ratios, we get a statistical distribution of the bandwidth size for each scheme.

3.2 Observations and Analysis

Classification Error, Bias and Variance We use Friedman's method [21] to rank classification error, bias and variance. The scheme that performs the best is ranked 1, the second best is ranked 2 and so forth. The mean rank of classification accuracy and time measure (real time) are summarized in Figure 1 as the shaded bars. Since the bandwidth calculations are carried out during training, the computational time for the test stage is essentially the same for all schemes and therefore is not reported.

A win/tie/lose record (w/t/l) is calculated for each pair of competitors A and B with regard to a performance measure M. The record represents the number of datasets in which A wins loses or ties with B on M. The win/tie/loss records are summarized in Table 2.

We also apply statistical comparison methods of multiple classifiers over multiple data sets recommended by Demsar [22].

The null hypothesis was rejected for all Friedman tests (using the 0.05 critical level) conducted on error, bias and variance, so we can infer that there exists significant difference among all ten schemes.

Having determined that a significant difference exists, the post-hoc Nemenyi test was used to identify which pairs of schemes differ significantly. The results of this test(using the 0.05 critical level) are illustrated by the line segments accompanying each bar in the graph in Figure 1. The length of these lines indicate the critical difference, and the performance of two methods are considered to be significantly different if the difference between their mean rank is greater than the critical difference (i.e. their vertical line segments do not overlap).

Figure 1 and Table 2 show that the more sophisticated bandwidth selection schemes investigated do not yield improved performance over simpler schemes, although they are far more computationally expensive. The poorest performer was BCV, which was statistically significantly worse than the more simplistic SP scheme (with w/t/l record 15/1/36) and WEKA scheme (with w/t/l record 11/0/41). UCV was also determined to be statistically significantly worse than the SP scheme (with /w/t/l record 18/0/36). The computational time costs of the four sophisticated schemes are far more than the others.

The UCV scheme achieved low bias, but high variance, as stated by its name. Conversely, BCV achieved low variance, but high bias. Neither the SP scheme's bias nor its variance was particularly high or low, and it was found to be statistically significantly better than the discretization method and the two worst sophisticated bandwidth selection schemes, UCV and BCV. This analysis shows

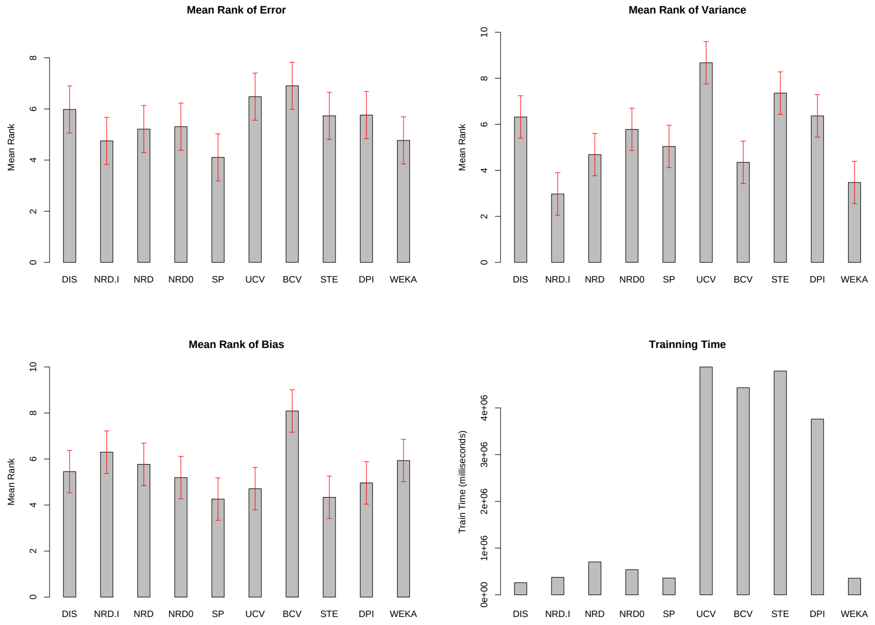

Fig. 1. Comparison of alternative methods' mean ranks of classification accuracy. Classification error can be decomposed into bias and variance. The shaded bars illustrate the mean rank and the smaller rank has the better performance. The line segments accompanying each bar indicate the Nemenyi test results. The performance of two methods are statistically significantly different if their vertical line segments are not overlapping. The mean training time is real time of computation.

that the choice of bandwidth dramatically affects the accuracy results of classification. The more sophisticated schemes can not guarantee good classification performance. Trade-off between bias and variance performance is essential to improve upon classification accuracy.

This analysis also shows that only one bandwidth selection scheme (the SP scheme) gives statistically better performance than a classical discretization method. It suggests that KDE can achieve statistically significantly better performance in classification, but the bandwidth selection schemes in classification behave different with traditional sophisticated bandwidth selection schemes. More theoretical researches are needed for kernel density estimation in classification.

Distribution of the Bandwidth The distribution of the bandwidth size for each scheme is illustrated in Figure 2. By comparing Figure 1 and Figure 2 we can see that the bandwidth of BCV and WEKA is statistically larger than others. This gives them a small variance and large bias in classification. By contrast,

Table 2. Comparison of rival schemes' win/tie/lose records with regard to classification error, bias and variance. Each three-number entry indicates the number of times the scheme named in the row wins, ties and loses against the scheme named in the column. A statistically significant record (at the 0.05 critical level) is indicated in a bold face.

(a) ERROR

w/t/l	DIS	NRD-I	NRD	NRD0	SP	UCV	BCV	STE	DPI
NRD-I	32/0/20								
NRD	30/0/22	22/1/29							
NRD0	28/0/24	22/0/30	25/1/26						
SP	**33/0/19**	32/0/20	34/2/16	34/0/18					
UCV	24/0/28	21/0/31	17/0/35	17/0/35	**14/0/38**				
BCV	26/0/26	9/0/43	15/0/37	16/1/35	**15/1/36**	23/0/29			
STE	25/1/26	19/0/33	23/0/29	25/0/27	18/0/34	33/1/18	29/0/23		
DPI	28/1/23	23/1/28	21/0/31	22/1/29	16/1/35	31/1/20	30/0/22	24/1/27	
WEKA	32/0/20	26/0/26	31/0/21	28/1/23	23/1/28	30/0/22	41/0/11	30/1/21	29/1/22

(b) BIAS

w/t/l	DIS	NRD-I	NRD	NRD0	SP	UCV	BCV	STE	DPI
NRD-I	22/0/30								
NRD	25/1/26	31/1/20							
NRD0	26/0/26	34/1/17	33/0/19						
SP	28/0/24	39/0/13	37/0/15	33/0/19					
UCV	27/0/25	32/0/20	29/0/23	31/1/20	27/0/25				
BCV	**19/0/33**	12/0/40	**12/0/40**	**9/0/43**	**8/0/44**	**13/0/39**			
STE	28/0/24	35/0/17	32/0/20	30/0/22	28/1/23	26/0/26	**45/1/6**		
DPI	29/0/23	35/1/16	33/0/19	31/0/21	20/0/32	20/1/31	**44/0/8**	18/0/34	
WEKA	27/0/25	26/0/26	25/0/27	21/0/31	15/0/37	20/0/32	**40/0/12**	16/1/35	21/0/31

(c) VARIANCE

w/t/l	DIS	NRD-I	NRD	NRD0	SP	UCV	BCV	STE	DPI
NRD-I	**37/1/14**								
NRD	34/0/18	8/1/43							
NRD0	34/0/18	**8/0/44**	11/0/41						
SP	32/0/20	**14/0/38**	26/0/26	30/0/22					
UCV	**20/0/32**	**3/0/49**	**5/0/47**	**5/0/47**	**7/0/45**				
BCV	32/0/20	16/3/33	31/0/21	36/0/16	29/1/22	**46/0/6**			
STE	24/0/28	**6/0/46**	**10/0/42**	13/0/39	**13/0/39**	42/0/10	**9/0/43**		
DPI	27/0/25	**8/0/44**	11/0/41	22/1/29	18/1/33	**43/0/9**	**11/1/40**	37/1/14	
WEKA	36/0/16	23/0/29	36/0/16	**39/0/13**	35/2/15	**48/0/4**	33/1/18	**46/0/6**	**42/0/10**

NRD0, SP, STE and DPI tend to have smaller bandwidths. This gives them a relatively small bias and large variance in classification. We can see that there is a transition range (from approximately 0.5 to 1.5 times of NRD-I bandwidth) that indicates a change in tendency of bias-variance trade off, from a low-bias high-variance to a high-bias low-variance profile. This transition range is narrow. This relatively narrow distribution range shows that classification performance is more sensitive to the size of the bandwidth than was first thought.

4 Conclusions

The more simplistic and less computationally intensive bandwidth selection schemes performed significantly better compared to some of the more sophisticated schemes in Naive Bayesian Classification. A kernel density estimation method can significantly outperform a classical discretization method, but only when appropriate bandwidth selection schemes are applied.

Our experiments and analysis also show that an unsuitable bandwidth value can easily give poor classification performance. In a relatively narrow distribution range, we find that the bias-variance trade off changes, from low-bias and

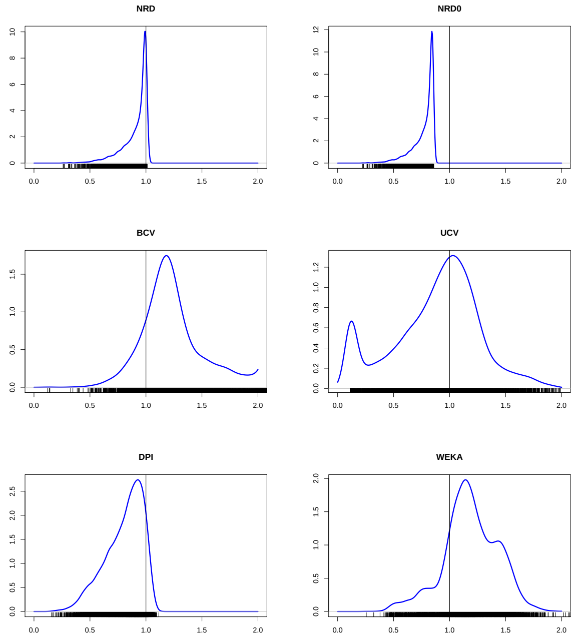

Fig. 2. Distribution of the size of bandwidth. X-axis is the ratio of alternative bandwidth to a standard bandwidth. Y-axis is the density of the ratio distribution. Standard bandwidth is NRD-I.

high-variance to high-bias and low-variance. Comparison of the bandwidth distribution patterns with error performance suggests that bandwidths within the range of 0.5 to 1.5 times NRD-I standard bandwidth are preferable.

5 Acknowledgements

The authors thank Dr. Eibe Frank and Dr.Leonard Trigg for the helpful discussion and Mr. Colin Enticott for the support of cluster computing. This work is supported by Australian Research Council grant DP0770741.

References

[1] Fayyad, U.M., Irani, K.B.: Multi-interval discretization of continuous-valued attributes for classification learning. Proceedings of the 13th International Joint Conference on Artificial Intelligence **2** (1993) 1022–1027

[2] Yang, Y., Webb, G.: Discretization for naive-bayes learning: managing discretization bias and variance. Machine Learning (2008) Online First

[3] Bay, S.D.: Multivariate discretization for set mining. Knowledge and Information Systems **3**(4) (2001) 491–512

[4] John, G.H., Langley, P.: Estimating continuous distributions in bayesian classifiers. Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (1995) 338–345

[5] Silverman, B.W.: Density Estimation for Statistics and Data Analysis. 1st edn. Chapman & Hall/CRC (1986)

[6] Wand, M.P., Jones, M.C.: Kernel Smoothing. Chapman & Hall/CRC (1994)

[7] Epanechnikov, V.A.: Non-parametric estimation of a multivariate probability density. Theory of Probability and its Applications **14**(1) (1969) 153–158

[8] Friedman, J.H.: On bias, variance, 0/1-loss, and the curse-of-dimensionality. Data Mining and Knowledge Discovery **1**(1) (1997) 55–77

[9] Hall, P., Kang, K.H.: Bandwidth choice for nonparametric classification. Annals of Statistics **33**(1) (2005) 284–306

[10] Bowman, A.W.: An alternative method of cross-validation for the smoothing of density estimates. Biometrika **71**(2) (1984) 353–360

[11] R Development Core Team: R: A Language and Environment for Statistical Computing. http://www.R-project.org, Vienna, Austria (2008)

[12] Scott, D.W., Terrell, G.R.: Biased and unbiased cross-validation in density estimation. Journal of the American Statistical Association **82**(400) (1987) 1131–1146

[13] Sheather, S.J., Jones, M.C.: A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society. Series B **53**(3) (1991) 683–690

[14] Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, Second Edition. Morgan Kaufmann (2005)

[15] Hyndman, R.J.: The problem with sturge's rule for constructing histograms. Available from http://www-personal.buseco.monash.edu.au/~hyndman/papers (1995)

[16] Sturges, H.A.: The choice of a class interval. Journal of the American Statistical Association **21**(153) (1926) 65–66

[17] Venables, W.N., Ripley, B.D.: Modern Applied Statistics with S-PLUS, Third Edition. Springer-Verlag Telos (1999)

[18] Asuncion, A., Newman, D.J.: UCI Machine Learning Repository. http://www.ics.uci.edu/~mlearn/MLRepository.html (2007)

[19] Webb, G.I.: Multiboosting: A technique for combining boosting and wagging. Machine Learning **40**(2) (2000) 159–196

[20] Kohavi, R., Wolpert, D.H.: Bias plus variance decomposition for zero-one loss functions. Machine Learning: Proceedings of the Thirteenth International Conference **275** (1996) 283

[21] Friedman, M.: The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association **32**(200) (1937) 675–701

[22] Demsar, J.: Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research **7** (2006) 1–30