INDECISIVE CLASSIFIER

Zhenqiu Zhang *Xun Xu* *Thomas Huang*
zzhang6@uiuc.edu , xunxu@uiuc.edu, huang@ifp.uiuc.edu
University of Illinois at Urbana Champaign
Urbana, IL, USA 61801

ABSTRACT

Nearest neighbor classification expects the class conditional probabilities to be locally constant. The assumption becomes invalid in high dimension due to the curse-of-dimensionality. Severe bias can be introduced under this condition when using nearest neighbor rule. We propose an adaptive nearest neighbor classification method "indecisive classifier" to minimize bias and variance by avoiding decision making in some hard-decision region. As a result, better classification performance can be expected in some scenario such as video based face recognition.

1. INTRODUCTION

In a situation where a mathematical representation of the underlying probability distribution is difficult to obtain, the optimum or Bayes classifier is difficult to implement. Under this condition, a more fruitful approach is to use nonparametric methods to classify a pattern. In the nonparametric framework, the nearest neighbor approach was first introduced by Fix and Hodges [1] and later studied by Cover and Hart [2]. Nearest neighbor method is supported by the fact that the asymptotic of infinite sample size error has a value between the Bayes error and twice the Bayes error. However, in practice, we never have an infinite number of samples, and, due to the finite sample size, the NN estimates have large biases and variances. In such cases, the NN method does not always approach the asymptotic risk or the infinite sample size error. This is a major obstacle in many practical situations where the ratio of the training sample size to the dimensionality is small.

A number of locally adaptive methods have recently been proposed to address the small ratio of training sample size to the dimensionality issue [3]. In [4], Trevor and Robert propose a locally adaptive form of nearest neighbor classification to ameliorate the curse of dimensionality. They use a local linear discriminant analysis to estimate an effective metric for computing neighborhoods. In [5], Abdelhamid presented a method to produce a finite sample size risk close to the asymptotic one. It is based on an attempt to eliminate the first-order effects of the samples size, as well as all higher odd terms. Domeniconi et al. [6] describe a locally adaptive NN method by approximating the Chi-queared distance. The technique employs a "patient" averaging process to reduce variance.

While these methods have shown promise in a number of classification problems, they have a potential limitation. In a high dimensional space where data become spare, we are forced to look far away from the query to find a nearest neighborhood. Severe bias can be introduced under this condition. As seen in [4], the relative radius of the nearest-neighbor sphere groups like $r^{1/d}$, where *d* is the dimension and *r* the radius for *d*=1, resulting in severe bias at the target point x.

In this paper, we propose an adaptive nearest neighbor classification method "indecisive classifier" to minimize bias by avoiding decision making in some hard-decision region. As a result, better classification performance can be expected in some scenario such as video based face recognition.

The rest of this paper is organized as follows. In section 2 we describe bias-variance of NN method. Then in section 3, we describe 2-NN method. Indecisive classifier algorithm is presented in section 4. In section 5 we report the results of experiments with indecisive classifier. Finally, in section 6 we present conclusions and ideas for future research.

2. BIAS-VARIANCE OF NN ALGORITHMS

The NN algorithm has been a subject of both experimental and theoretical studies for many years. Experimental studies have shown that the predictive accuracy of NN algorithms is comparable to that of decision trees, rule learning systems, and neural net learning algorithms on many practical tasks. In addition, it has long been known that the probability of error of the NN rule is bounded above by twice (optimal) Bayes probability of error (Eq.1).

$$R^* \leq R \leq 2R^* \tag{1}$$

0-7803-9332-5/05/$20.00 ©2005 IEEE

However, when nearest-neighbor classification is carried out in a high-dimensional feature space, the nearest neighbors of a point can be very far away, causing bias and degrading the performance of the rule [7].

To quantify this [4], consider N data points uniformly distributed in the unit cube $\left[-0.5,0.5\right]^d$. Consider a spherical (one)-nearest neighborhood centered at the origin. Let R be the radius of the neighborhood. Then

$$\Pr ob(R \geq r) = (1 - v_d r^d)^N \tag{2}$$

Where $v_d r^d$ is the volume of the sphere of radius r in d dimensions. We can compute the median of R:

$$med(R) = {v_d}^{-1/d}(1 - 0.5^{1/N})^{1/d} \tag{3}$$

The median radius quickly approaches 0.5, the distance to the edge of the cube [4].

For two classes problem, $Y \in \{0,1\}$, we have the conditional probability $p\left(Y = 1 \mid x\right)$ and $p\left(Y = 0 \mid x\right)$. At a point x_0, we find the nearest neighbor X with class $Y = C(X)$ (X is a random variable), which equal to 0 or 1, and the estimation of $C(x_0)$ is simply $C(X)$. The bias and variance of $C(X)$ are:

$$\begin{aligned} Bias &= E[C(X)] - C(x_0) \\ &= E[p(Y = 1 \mid X)] - p(Y = 1 \mid x_0) \end{aligned} \tag{4}$$

$$\begin{aligned} Var &= \mathrm{var}[E(C(X) \mid X)] + E[\mathrm{var}(C(X) \mid X] \\ &= E[p(Y = 1 \mid X)](1 - E[p(Y = 1 \mid X]) \end{aligned} \tag{5}$$

The expectations of bias and variance are with respect to the distribution of the nearest neighbor X. Nearest neighbor classification expects the class conditional probabilities to be locally constant. When dimension increases, the distance between nearest neighbor X and x_0 increases, so the difference between $p(Y \mid X)$ and $p(Y \mid X_0)$ will increase and the bias $E[C(X)] - C(x_0)$ will increase [4].

For variance, we could find it will increase as well when dimension increases [4]. Suppose we have equal numbers in each class. $E[p(Y = 1 \mid X)](1 - E[p(Y = 1 \mid X)])$ takes its maximum at $p(Y = 1 \mid X) = 1/2$. If x_0 is in a pure region, where $p(Y = 1 \mid X_0)$ is near 1 or 0, and x is very near x_0, the variance will be small. However, when x_0 is not in a pure region (equal (6)), or x is far away from x_0 (when dimension increases), $p(Y = 1 \mid X)$ is tend to be 0.5, the variance will approach its maximum.

$$p(Y = 1 \mid X) \approx p(Y = 0 \mid X) \tag{6}$$

3. 2-NN ESTIMATES

Also let's consider two class problem, $Y \in \{0,1\}$, with the conditional probabilities $p\left(Y = 1 \mid x\right)$ and $p\left(Y = 0 \mid x\right)$. The conditional probability of error (Bayes risk) when X is classified according to the Bayes decision rule is

$$r(X) = \min\{p(Y = 1 \mid X), p(Y = 0 \mid X)\} \tag{7}$$

In [8], 2-NN is proposed to estimate the Bayes risk locally. Consider the 2-NN estimate at a point X. Let X^1 and X^2 be the first and second NN to X and Y^1 and Y^2 their respective classes. If $Y^1 \neq Y^2$, the conditional risk $r(X)$ is more likely to be large, and $r(X)$ is small if $Y^1 = Y^2$. We estimate the posterior probabilities as

$$\begin{aligned} &p(Y = i \mid X) = 0 \quad\; if \quad Y^1 = Y^2 \neq i \\ &p(Y = i \mid X) = 1/2 \;\; if \quad Y^1 \neq Y^2 \\ &p(Y = i \mid X) = 2/2 \;\; if \quad Y^1 = Y^2 = i \end{aligned} \tag{8}$$

Then the 2-NN estimate of $r(X)$, may be taken as

$$r(x) = \min\{p(Y = 0 \mid X), p(Y = 1 \mid X)\} \tag{9}$$

So, if $Y^1 \neq Y^2$, $r(X)$ equals to 1/2, otherwise $r(X)$ equals to 0.

4. INDECISIVE CLASSIFIER

Now, let's consider K class problem, $Y \in \{1,2,...K\}$. Same as 2-NN method, at a point x_0, we find two nearest neighbors X^1 and X^2 of X, with Y^1 and Y^2 be their respective classes. The difference is that X^1 and X^2 must belong to two difference classes, which means $Y^1 \neq Y^2$. Indeed, if we define the distance of the point x_0 to a class k as the distance between point x_0 and nearest neighbor X^k of x_0, which belong to class *k,* the distance between X and X^1, $D_1 = D(X, X^1)$, stands for the distance of X to its nearest class, and $D_2 = D(X, X^2)$ is the distance of X to its second nearest class, as seen in Figure 1. The strategy of indecisive classifier is that, when

$$D(X, X^1) > Threshold1 \tag{10}$$

AND

$$D(X, X^2) - D\left(X, X^1\right) < Threshold2 \tag{11}$$

we make no decision (Threshold 1 and 2 can be set by experience), otherwise, X belongs to class Y^1.

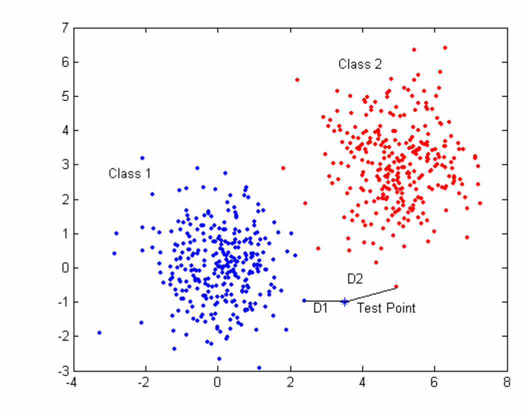

Figure 1: Distance to two nearest classes.

The assumption is that in high dimensional feature space, the point X and its nearest neighbor can be far away (Eq.10), so that the bias and variance of nearest neighbor method at point X is high. In another situation, when X_0 is not in a pure region (Eq.11), the variance of nearest neighbor method also can be high. Indecisive classifier avoids making decision if both of these situations happen, which means really high bias and variance locally (see Figure 1).

5. EXPERIMENT

Performance of indecisive classifier is tested on two databases: MNIST digit handwritten database and Yamaha AGV database.

5.1. Handwritten digits recognition

In this experiment, we applied the proposed indecisive classifier on MNIST database of handwritten digits. In this dataset, each class represents one of 10 handwritten numerals. The MNIST database of handwritten digits, available from the web [9], has a training set of 60,000 examples, and test set of 10,000 example.

In this experiment, we only used the test set of 10,000 examples. We randomly selected part of the 10,000 examples for training and the left for testing. We decreased the number of training samples from 5,000 to 300 and use nearest neighbor with Euclidean distance for classification. The recognition rate decreased from 95% to 86.5%. The result shows that training samples become sparser in the high dimensional space when decreasing the number of training samples, and bias of nearest neighbor classifier becomes more serious.

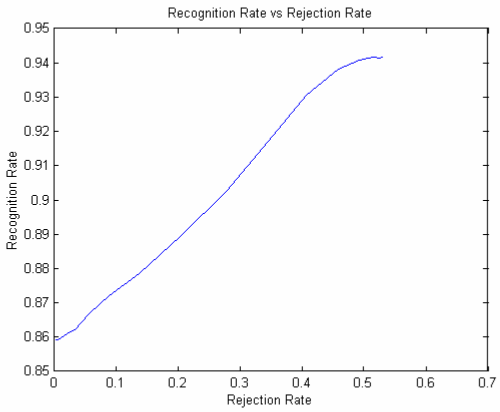

Figure 2: Roc curve for handwritten digits recognition

In case of small training sample size, we applied indecisive classifier to obtain higher recognition rate within some rejection rate. It is a tradeoff between recognition rate and rejection rate.

Figure 2 shows the ROC curve while 300 samples are used for training, which means we only have 30 training samples for each digit. Compared with the dimension of the feature space (28*28), the training samples are really sparse in the high dimensional feature space. We keep threshold 1 as the median of the distance of nearest neighbor in the training set and adjust value of threshold 2. As seen in figure 2, recognition rate increases from 86% to 94%. At the meantime, rejection rate increases from 0% to 50%.

5.2. VIDEO BASED FACE RECOGNITION

In this experiment, we applied indecisive classifier to video based face recognition. In this scenario, face recognition is based on voting among a sequence of video frames. Instead of making decision on every frame, we could make no decision on some frames that we are not very sure, which is the idea of indecisive classifier.

The database we used is the Yamaha AGV database, in which, there are totally 22 persons, sitting in the golf car. For each person, several minutes of video sequence were captured in the moving golf car. Since these video sequences were captured in the outdoor environment and people moved their head arbitrarily during capturing, pose and illumination change of people's face is big in this database (see figure 3).

For every 4 to 5 seconds, ground truth of face location (two outer eye corners) is provided for one frame of the video sequence. In each video sequence, there are 80 frames provided with ground truth of face location, so each person has 80 face images with ground truth. From the first 30 frames of each sequence, we randomly select 5 frames for training as seen in figure 3 and the remain 50 images per person is used for testing. Face images were cropped out, normalized to 30*30 and histogram equalization was applied to decrease the illumination variance. We constructed a PCA subspace [10] based on these training face images (5*22=110), and the dimension of this PCA subspace was chose to be 30.

Figure 3: Yamaha AGV database

In the testing stage, face images were cropped out based on provided ground truth, normalized to 30*30, projected to the PCA subspace of dimension 30, and person ID was obtained by nearest neighbor rule for each frame (recognition result is given for each testing image separately). The ROC curve of single image based face recognition is shown in Figure 4. Here we keep threshold 1 as the median distance of nearest neighbor in the training set and adjust value of threshold 2. Recognition rate increases from 72% to 93% and rejection rate increases from 0% to 50%.

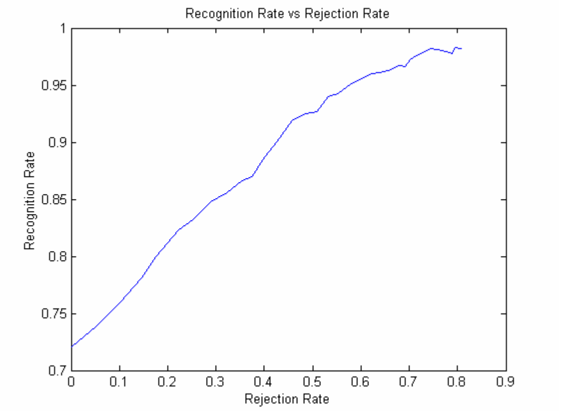

Figure4: Roc curve for face recognition

Then we divide the 50 testing images of each person into nine overlapped subsets, each subset contained 11 consecutive frames with ground truth, such as frame 31 to frame 41, frame 35 to frame 45 and frame 41 to frame 51, and there were totally 198 (9*22) this kind of subset. In stead of recognizing the person by single image seperately, the final face recognition result for each subset is based on the voting result among the eleven frames in this subset.

We adjust threshold 2 to obtain 30% rejection rate, and vote only on the frames which are not rejected. We select the person ID with maximum voting score. In case of tie, which means two person ID have same maximum voting score, we select the ID with smaller distance to the testing subset. 170 sub-sequences of these 198 sub-sequences obtain correct person ID, so the recognition rate is 86%, using this indecisive classifier based voting method.

6. CONCLUSION

We have proposed an indecisive classifier for the NN classifier. The performance of the NN classifier based on indecisive classifier was demonstrated on two databases. It is shown that the proposed classifier outperforms the conventional NN classifier. Experimental results suggest that the use of our indecisive classifier is an effective means of reducing the bias and variance of the NN error.

11. REFERENCES

[1] E. Fix and J.L. Hodges, "Discriminatory Analysis: Nonparametric Discrimination: Small Sample Performance", Report No.11, USAF School of Aviation Medicine, Randolph Field, Texas, Aug. 1952..

[2] T.M. Cover and P.E.Hart, "Nearest Neighbor Pattern Classification", Information Theory, vol. 13, no. 1, pp.21-27, Jan.1967.

[3] Jing Peng, Douglas R. Heisterkamp, and H.K. Dai, "Adaptive Quasiconformal Kernel nearest Neighbor Classification", PAMI, VOL. 26, NO. 5, May 2004.

[4] Trevor Hastie and Robert Tibshirani, "Discriminant Adaptive Nearest Neighbor Classification", PAMI, VOL. 18, NO. 6, JUNE 1996..

[5] Abdelhamid Djouadi, "On the Reduction of the Nearest Neighbor Variation for More Accurate Classification and Error Estimates", PAMI, VOL. 20, NO.5, May 1998.

[6] C. Domeniconi, J.Peng, and D. Gunopulos, "Locally Adaptive Metric nearest Neighbor Classification", PAMI, Vol.24, No. 9, PP. 1281-1285, Sept. 2002.

[7] Trevor Hastie, Robert Tibshirani and Jerome Friedman, "The elements of statistical learning", Springer 2001, pp. 427-432.

[8] Keinosuke Fukunaga and David Kessell, "Nonparametric Bayes error estimation using unclassified samples", Information theory, Vol. IT-19,No. 4, July 1973.

[9] http://yann.lecun.com/exdb/mnist/, The MNIST database of handwritten digits.

[10] M.A. Turk and A.P. Pentland, "Face recognition using eigenfaces," *Proc. Int. Conf. on Patt. Recog.*, pp. 586–591, 1991.