Pattern Recognition 36 (2003) 25–34

Breast cancer detection using rank nearest neighbor classification rules

Subhash C. Bagui[a], Sikha Bagui[b], Kuhu Pal[c], Nikhil R. Pal[d,*]

[a]*Department of Mathematics and Statistics, The University of West Florida, Pensacola, FL 32514, USA*
[b]*Department of Computer Science, The University of West Florida, Pensacola, FL 32514, USA*
[c]*Pragatinagar, Chinsurah, Hooghly, WB, India*
[d]*Electronics and Communication Sciences Unit, Indian Statistical Institute, 203 B.T. Road, Calcutta 700 035, India*

Received 30 May 2001; accepted 10 December 2001

Abstract

In this article, we propose a new generalization of the rank nearest neighbor (RNN) rule for multivariate data for diagnosis of breast cancer. We study the performance of this rule using two well known databases and compare the results with the conventional k-NN rule. We observe that this rule performed remarkably well, and the computational complexity of the proposed k-RNN is much less than the conventional k-NN rule.

Keywords: Classification rules; Rank nearest neighbor rules; Nearest neighbor rules; Breast masses; Breast cancer detection; Cell nucleus; Mean texture; Worst mean area; Error rate; Bayes error rate

1. Introduction

Of all types of cancers, breast cancer is one of the leading causes of death among middle-aged and old women. According to an estimate by WHO, by the end of year 2000, worldwide death due to breast cancer alone would be 500,000/year (We do not know yet the exact figure of 2000). Thus, prevention and an early diagnosis of breast tumor are immediate demands from the society. The primary prevention is difficult as the causes of the disease are not well understood. But if it can be detected at its early stage, success rate of survival is quite high. Physical examination is one of the methods used for detection of breast tumor, but effectiveness of this technique is limited by the subjective ability of doctors. In addition to physical examination, mammograms are quite often used. Currently mammography is the best way to make an early diagnosis of the disease. A precise detection, however, often depends on the visibility of microcalcifications in the mammogram. It is still challenging for radiologists to differentiate between benign and malignant cases. The existence of breast tumor is usually reflected in the mammogram. Some of the important signs of malignancy are: clustered calcifications, poorly defined masses, isolated dilated ducts, etc. But all of these are not equally reflected in the mammograms. Experts (doctors) physically look at the mammograms to detect deformations that may be taken as an indicator of cancerous changes. Obviously this suffers from the human error and error with visual inspection, which may further be enhanced by poor quality of the mammogram images. Most importantly, with such subjective visual analysis, explicit use of any consistent diagnostic principle is often difficult. There is a demand for intelligent systems for early detection of tumors, assessment of their malignancy and monitoring of the same [1–3]. In this direction even some aiding tools would be of immense help. The efficiency and effectiveness of this process can be increased if tumors are detected and classified automatically through computers as benign or malignant.

* Corresponding author. Tel.: +91-33-577-3035; fax: +91-33-577-3035.
E-mail address: nikhil@isical.ac.in (N.R. Pal).

In this article our aim is to use the proposed k-RNN classifier to discriminate between benign or malignant masses, and to compare the results with the conventional k-NN rule.

2. k-RNN rule and k-NN rule

The nearest neighbor (NN) classification rule was first introduced by Fix and Hodges [4,5]. This rule is based on the density estimates using distance nearest neighbors. Cover and Hart [6] proposed and studied a slightly modified version of Fix and Hodges's NN rule. They termed the rule as the k-NN rule. This rule is very widely used and popular among computer scientists. This conventional k-NN rule may be described as follows:

The k-NN rule: Let $\{X_1, X_2, \ldots, X_{n_1}\}$ and $\{Y_1, Y_2, \ldots, Y_{n_2}\}$ be training samples from two given populations π_1 and π_2, respectively, and Z be an unknown observation known to be from either π_1 or π_2 to be classified between π_1 and π_2. Using a distance function d, order the distances of all the observations from Z. For a fixed integer k, the k-NN rule assigns the unknown observation Z to π_i if the majority of the k nearest neighbors (in a distance sense) of Z come from π_i, $i = 1, 2$.

Cover and Hart [6] showed that for 1-NN rule, bounds for the limiting risk R_1 satisfy $R^* \leqslant R_1 \leqslant 2R^*(1 - R^*)$, where R^* is the (minimum) Bayes error rate. That is

$$R^* = \int \min(\xi_1 f_1(z), \xi_2 f_2(z))\,\mathrm{d}z,$$

where ξ_i and f_i are prior probability and the density function for the class i, respectively, $i = 1, 2$. Devroye [7] obtained the following upper bound on the asymptotic risk of the k-NN rule R_k:

$$R_k \leqslant (1 + a_k)R^*, \quad a_k = \frac{\alpha\sqrt{k}}{k - 3.25}\left(1 + \frac{\beta}{\sqrt{k-3}}\right),$$

k odd, $k \geqslant 5$,

where $\alpha = 0.3399$ and $\beta = 0.9749$ are universal constants. This bound is the best possible in a certain sense. Next we describe the k-rank nearest neighbor (k-RNN) rule.

The k-RNN rule: The k-RNN rule for univariate populations was first introduced by Anderson [8]. This rule may be described as follows:

Pool the observations X_i's, Y_j's and Z, and rank them in ascending order; then count down k observations to the left-hand side of Z and count up k observations to the right-hand side of Z; (i) if there are more X's than Y's among $2k$ rank nearest neighbors, classify Z into the X-population π_1; (ii) if there are more Y's than X's, classify Z into the Y-population π_2; (iii) if there are exactly k X's and k Y's, classify Z into either of the two populations with probability $\frac{1}{2}$ each (to break the tie); and (iv) if on any side of Z k observations are not available then use as many as available.

Dasgupta and Lin [9] studied the 1-RNN rule. They derived the asymptotic risk (r_1) of the 1-RNN rule and showed that $R^* \leqslant r_1 \leqslant 2R^*$, where R^* is the (minimum) Bayes error rate defined earlier. In fact the value of this asymptotic risk r_1 is exactly the same as the asymptotic risk of the conventional 1-NN rule. Bagui and Vaughn [10] investigated the k-RNN and obtained the asymptotic risk r_k of this rule. They derived an upper bound of this risk which is parallel to the bound obtained by Devroye [7]. The upper bound on r_k is

$$r_k \leqslant (1 + c_k)R^*, \quad c_k = \frac{\alpha\sqrt{2k-1}}{2k - 4.25}\left(1 + \frac{\beta}{\sqrt{2k-4}}\right),$$

$k \geqslant 3$.

Bagui and Vaughn [10] also demonstrated that this risk converges to the Bayes risk twice as fast as the conventional k-NN rule. Thus, the k-RNN rule has good asymptotic properties. But one drawback with this rule is that there is no natural extension to the multivariate populations, since multivariate observations cannot be ranked uniquely. But the majority of the real life problems occur in a multivariate form. Bagui and Pal [11] suggested a k-RNN rule for multivariate data. This rule uses the idea of component wise classification using univariate k-RNN rule, then the majority vote rule is applied on the feature level decisions to classify an unknown multivariate observation.

In this article, we propose an efficient way of ranking multivariate observations. This new procedure ranks the multivariate observations taking into account the variability between mean vectors and the covariance matrices of the populations. Below we describe the multivariate k-RNN rule.

2.1. The k-RNN rule for multivariate data

Suppose we have two multivariate populations, say a X-population π_1 and a Y-population π_2. We also assume that $\mathbf{X} = (x_1, x_2, \ldots, x_p)' \in \mathbf{R}^p$ follows a multivariate distribution with a mean of $\boldsymbol{\mu}_1 \in \mathbf{R}^p$ and covariance matrix of Σ_1 of size $p \times p$, and $\mathbf{Y} = (y_1, y_2, \ldots, y_p)' \in \mathbf{R}^p$ follows a multivariate distribution with a mean of $\boldsymbol{\mu}_2 \in \mathbf{R}^p$ and covariance matrix of Σ_2 of size $p \times p$. An object $\mathbf{Z} = (z_1, z_2, \ldots, z_p)' \in \mathbf{R}^p$ is known to originate either from π_1 or π_2, is to be classified into one of π_1 or π_2. Suppose that only training data (past samples) are available from both populations. Let $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{n_1}\}$ and $\{\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_{n_2}\}$ be training samples from the two given multivariate populations π_1 and π_2, respectively. In order to use the above mentioned k-RNN rule we need to have pooled ranks of $\mathbf{X}_i$'s, $\mathbf{Y}_j$'s and $\mathbf{Z}$. For this purpose we propose the following *score* function to obtain the combined ranks of $\mathbf{X}_i$'s, $\mathbf{Y}_j$'s and $\mathbf{Z}$

$$D(\mathbf{Z}; \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma_1, \Sigma_2) = \left(\boldsymbol{\mu}_1'\Sigma_1^{-1} - \boldsymbol{\mu}_2'\Sigma_2^{-1}\right)\mathbf{Z} - \tfrac{1}{2}\mathbf{Z}'(\Sigma_1^{-1} - \Sigma_2^{-1})\mathbf{Z}, \tag{1}$$

where $\boldsymbol{\mu}_i'$ denote the transpose of the mean vector $\boldsymbol{\mu}_i$ and Σ_i^{-1} denote the inverse of the covariance matrix Σ_i for $i=1,2$.

The score function $D(.)$ maps from $\mathbf{R}^p$ to $\mathbf{R}^1$. It is also a continuous function since it is a difference between a linear and quadratic functions. Depending on the values of the score function $D(.)$ we get the combined ranks of $\mathbf{X}_i$'s, $\mathbf{Y}_j$'s and $\mathbf{Z}$. A nice thing about $D(.)$ is that it ranks observations by taking into account the variability between $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, Σ_1, and Σ_2 and it is like a quadratic discriminant function between two populations. We can consider two other variations of $D(.)$ as discussed below

(i) If it is known that $\Sigma_1=\Sigma_2=\Sigma$, then (1) takes a simpler form (2) that may be used for ranking observations:

$$D(\mathbf{Z};\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\Sigma)=(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\Sigma^{-1}\mathbf{Z}. \quad (2)$$

(ii) If $\boldsymbol{\mu}_1=\boldsymbol{\mu}_2=\boldsymbol{\mu}$ is suspected then the score function (1) reduces to

$$D(\mathbf{Z};\boldsymbol{\mu},\Sigma_1,\Sigma_2)=\tfrac{1}{2}(2\boldsymbol{\mu}-\mathbf{Z})'(\Sigma_1^{-1}-\Sigma_2^{-1})\mathbf{Z}. \quad (3)$$

If the parameters $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, Σ_1, and Σ_2 are unknown, which would generally be the case, then they may be replaced by their corresponding unbiased sample estimates $\bar{\mathbf{X}},\bar{\mathbf{Y}},\mathbf{S}_1$, and $\mathbf{S}_2$, respectively. Thus the estimated score function is

$$\hat{D}(\mathbf{Z};\bar{\mathbf{X}},\bar{\mathbf{Y}},\mathbf{S}_1,\mathbf{S}_2)$$
$$=(\bar{\mathbf{X}}'\mathbf{S}_1^{-1}-\bar{\mathbf{Y}}'\mathbf{S}_2^{-1})\mathbf{Z}-\tfrac{1}{2}\mathbf{Z}'(\mathbf{S}_1^{-1}-\mathbf{S}_2^{-1})\mathbf{Z},$$

where

$$\bar{\mathbf{X}}=\frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{X}_i,$$

$$\bar{\mathbf{Y}}=\frac{1}{n_2}\sum_{j=1}^{n_2}\mathbf{Y}_j\mathbf{S}_1=\frac{1}{n_1-1}\sum_{i=1}^{n_1}(\mathbf{X}_i-\bar{\mathbf{X}})(\mathbf{X}_i-\bar{\mathbf{X}})'$$

and

$$\mathbf{S}_2=\frac{1}{n_2-1}\sum_{i=1}^{n_2}(\mathbf{Y}_i-\bar{\mathbf{Y}})(\mathbf{Y}_i-\bar{\mathbf{Y}})',$$

see Johnson and Wichern [12].

This type of score functions are previously used by Randles et al. [22] to rank multivariate observations between two populations. Once the ranking of $\mathbf{X}_i$'s, $\mathbf{Y}_j$'s and $\mathbf{Z}$ is done, classification of $\mathbf{Z}$ can be done easily using the *k-RNN Algorithm* defined earlier.

2.1.1. Computational complexity of k-NN and k-RNN rules

Note that if there are N observations in the training data, then the total cost of finding the rank of $\mathbf{Z}$ is equal to the cost of score calculation for $\mathbf{Z}$ and the cost of finding the rank of $\mathbf{Z}$. The score function has two parts, the first part requires p multiplications and $p-1$ additions, while the second part requires (p^2+p) multiplications and $(p+1)(p-1)$ additions and then 1 multiplication for 0.5 and 1 addition for combining the two parts. So the total cost of score calculation is $2p^2+p+1$ operations (additions and multiplications); while the cost of finding the rank of $\mathbf{Z}$ is $\log_2 N$ comparisons, where $N=n_1+n_2$. So the total cost is $\log_2(N)+2p^2+2p+1$ operations. On the other hand, for the conventional k-NN rule, the cost of finding k-neighbors is $(N-1)(N-2)\cdots(N-k)$ comparisons and N distance calculations involving a cost of $N(3p-1)$ operations as each square Euclidean distance involves p subtraction, p multiplications and $p-1$ additions, which is much higher than that of k-RNN rule.

2.2. Some asymptotic properties

In case of large samples from π_1 and π_2, we may apply large sample theory on $\bar{\mathbf{X}}$, $\bar{\mathbf{Y}}$, $\mathbf{S}_1$, $\mathbf{S}_2$ and $\hat{D}(.)$ and show that $\bar{\mathbf{X}}\to\boldsymbol{\mu}_1$, $\bar{\mathbf{Y}}\to\boldsymbol{\mu}_2$, $\mathbf{S}_1\to\Sigma_1$, $\mathbf{S}_2\to\Sigma_2$ and $\hat{D}(.)\to D(.)$ in probability.

Let $n=\min(n_1,n_2)$. Note by Chebyshev's inequality that

$$P\{|\bar{x}_i-\mu_i|>\varepsilon, \quad \text{for all } i=1,2,\ldots,p\}$$
$$\leqslant P\{|\bar{x}_i-\mu_i|>\varepsilon,\}\leqslant\frac{\mathrm{Var}(X_i)}{n\varepsilon^2}.$$

Now as $n\to\infty$, $P\{|\bar{x}_i-\mu_i|>\varepsilon$, for all $i=1,2,\ldots,p\}\to 0$, we may conclude that $\bar{\mathbf{X}}\to\boldsymbol{\mu}_1$ in probability. Similarly, we can prove that $\bar{\mathbf{Y}}\to\boldsymbol{\mu}_2$ in probability, $\mathbf{S}_1\to\Sigma_1$ in probability, and $\mathbf{S}_2\to\Sigma_2$ in probability. From $\mathbf{S}_1\to\Sigma_1$ in probability and $\mathbf{S}_2\to\Sigma_2$ in probability, we obtain $\mathbf{S}_1^{-1}\to\Sigma_1^{-1}$ in probability and $\mathbf{S}_2^{-1}\to\Sigma_2^{-1}$ in probability, since the probability limits of sums, differences, products, and quotients of random variables are sums, differences, products, and quotients of the probability limits as long as the probability limit of each denominator is different from zero, Cramer [13, p. 254]. Furthermore,

$$(\bar{\mathbf{X}}'\mathbf{S}_1^{-1}-\bar{\mathbf{Y}}'\mathbf{S}_2^{-1})\to\left(\boldsymbol{\mu}_1'\Sigma_1^{-1}-\boldsymbol{\mu}_2'\Sigma_2^{-1}\right) \text{ in probability}$$

and

$$(\mathbf{S}_1^{-1}-\mathbf{S}_2^{-1})\to\left(\Sigma_1^{-1}-\Sigma_2^{-1}\right) \text{ in probability.}$$

Thus, we may conclude that $\hat{D}(.)\to D(.)$ in probability.

3. Implementations

We implement our k-RNN rule on two well known breast cancer databases namely: (i) Wisconsin diagnostics breast cancer (WDBC) database; (ii) Wisconsin breast cancer (WBC) database. The performance of the k-RNN rule is compared with the conventional k-NN rule.

3.1. Description of the databases

Wisconsin diagnostics breast cancer (WDBC) database: The WDBC database is created by Wolberg et al., University of Wisconsin [14,15]. This database contains 569 observations among which 357 are benign cases and 212 are

Table 1
Summary of results for first partition of X_{WDBC}

k	k-RNN rule				k-NN rule			
	Confusion matrix		Prob. of false positive false negative	Avg. error rate	Confusion matrix		Prob. of false positive false negative	Avg. error rate
1	246	11	0.045	0.041	238	19	0.074	0.092
	4	108	0.036		15	97	0.134	
2	246	11	0.045	0.041	249	8	0.031	0.094
	4	108	0.036		25	87	0.223	
3	244	13	0.051	0.041	241	16	0.062	0.081
	2	110	0.018		14	98	0.125	
4	244	13	0.051	0.041	248	9	0.035	0.081
	2	110	0.018		21	91	0.188	
5	244	13	0.051	0.041	242	15	0.058	0.079
	2	110	0.018		14	98	0.125	
6	244	13	0.051	0.041	245	12	0.047	0.079
	2	110	0.018		17	95	0.152	

Table 2
Summary of results for second partition of X_{WDBC}

k	k-RNN rule				k-NN rule			
	Confusion matrix		Prob. of false positive false negative	Avg. error rate	Confusion matrix		Prob. of false positive false negative	Avg. error rate
1	249	8	0.031	0.049	239	18	0.070	0.089
	10	102	0.089		15	97	0.058	
2	241	16	0.062	0.059	250	7	0.027	0.070
	6	106	0.053		19	93	0.169	
3	241	16	0.062	0.059	244	13	0.051	0.068
	6	106	0.053		12	100	0.107	
4	241	16	0.062	0.059	249	8	0.031	0.059
	6	106	0.053		14	98	0.125	
5	241	16	0.062	0.059	242	15	0.058	0.076
	6	106	0.053		13	99	0.116	
6	241	16	0.062	0.059	249	8	0.031	0.065
	6	106	0.053		16	96	0.143	

malignant cases. For each instance, there are 30 featured variables. These features are computed from digital images of fine needle aspirates (FNA) of breast masses [16,17]. These features describe the characteristics of the cell nuclei in the image. The authors of this database considered 10 real-valued features for each cell nucleus:

(i) radius (mean of distances from center to points on perimeter); (ii) texture (standard deviation of gray-scale values); (iii) perimeter; (iv) area; (v) smoothness (local variation in radius lengths); (vi) compactness (perimeter^2/(area—1.0)); (vii) concavity (severity of concave portions of the contour); (viii) concave points (number of concave portions of the contour); (ix) symmetry; and (x) fractal dimension (*coastline approximation*—1.0).

They computed the *mean*, *standard error*, and *worst mean* (the mean of the three largest values) of each feature. This process resulted in 30 feature variables for each image. Bennett and Mangasarian [18] arrived at three best features from the above data by creating separating hyperplane that uses multisurface method-tree (MSM-T) and a classification method that uses linear programming to construct a decision tree. These three features are *mean texture*, *worst mean area*, and *worst mean smoothness*. Based on these three features they reported an estimated (best) correct classification percentage of 97.5 [16,17]. This estimate was obtained using a repeated 10-fold cross-validation method.

Wisconsin breast cancer (*WBC*) *database*: The source of this database is the University of Wisconsin Hospital, Madison [19]. There are 699 data points in this database of which 458 are benign and 241 are malignant. In 16 instances there are some missing values. So we decided to remove these 16 incomplete observations from the original database

Table 3
Summary of results for third partition of X_{WDBC}

k	k-RNN rule				k-NN rule			
	Confusion matrix		Prob. of false positive false negative	Avg. error rate	Confusion matrix		Prob. of false positive false negative	Avg. error rate
1	242	15	0.058	0.059	231	26	0.101	0.087
	7	105	0.063		6	106	0.054	
2	242	15	0.058	0.057	247	10	0.039	0.062
	6	106	0.053		13	99	0.116	
3	242	15	0.058	0.057	236	21	0.082	0.073
	6	106	0.053		6	106	0.054	
4	242	15	0.058	0.057	246	11	0.043	0.059
	6	106	0.053		11	101	0.098	
5	244	13	0.051	0.052	240	17	0.066	0.068
	6	106	0.053		8	104	0.071	
6	244	13	0.051	0.052	244	13	0.051	0.059
	6	106	0.053		9	103	0.080	

Table 4
Summary of results for fourth partition of X_{WDBC}

k	k-RNN rule				k-NN rule			
	Confusion matrix		Prob. of false positive false negative	Avg. error rate	Confusion matrix		Prob. of false positive false negative	Avg. error rate
1	233	24	0.093	0.087	237	20	0.078	0.057
	6	106	0.054		11	101	0.098	
2	233	24	0.093	0.087	246	11	0.043	0.065
	6	106	0.054		13	99	0.116	
3	233	24	0.093	0.087	239	18	0.070	0.076
	6	106	0.054		10	102	0.089	
4	233	24	0.093	0.087	244	13	0.051	0.071
	6	106	0.054		13	99	0.116	
5	233	24	0.093	0.087	240	17	0.062	0.073
	6	106	0.054		10	102	0.089	
6	233	24	0.093	0.087	241	16	0.066	0.073
	6	106	0.054		11	101	0.098	

and work with the remaining 683 data points (444 benign cases and 239 malignant cases). WBC is a nine-dimensional data set with the following features:

(i) Clump thickness; (ii) Uniformity of cell size; (iii) Uniformity of cell shape; (iv) Marginal adhesion; (v) Single epithelial cell size; (vi) Bare nucli; (vii) Bland chromatin; (viii) Normal nucleoli; and (ix) Mitoses.

As of 1990 this data set had 369 instances. Wolberg and Mangasarian [20] used this data set for two different methods and their reported correct classification percentages are 93.5 and 95.9, respectively. Zhang [21] also studied this data set for classification purposes by two different methods. His reported percentages of correct classifications are 93.7 and 92.2.

3.2. Results

3.2.1. WDBC datadase

Let the random variable $\mathbf{X} \in \mathbf{R}^{30}$ denote the benign population and follow a multivariate distribution with a mean vector of $\boldsymbol{\mu}_1$ and a covariance matrix of Σ_1, and the random variable $\mathbf{Y} \in \mathbf{R}^{30}$ denote the malignant population with a mean vector of $\boldsymbol{\mu}_2$ and a covariance matrix of Σ_2. Let us denote the WDBC data set by $X_{WDBC} = X_B \cup X_M$ where X_B is the set of benign cases and X_M is the set of malignant cases. The database X_{WDBC} contains 569 data points of which 357 are benign cases and 212 are malignant cases. We partition X_{WDBC} into X_{Tr} and X_{Te} such that $X_{WDBC} = X_{Tr} \cup X_{Te}$, $X_{Tr} \cap X_{Te} = \phi$, where X_{Tr} is called the training data set and

Table 5
Summary of results on X_{WDBC} with the three best features for first partition

k	k-RNN rule				k-NN rule			
	Confusion matrix		Prob. of false positive false negative	Avg. error rate	Confusion matrix		Prob. of false positive false negative	Avg. error rate
1	254	3	0.012	0.035	217	40	0.156	0.138
	10	102	0.089		9	103	0.080	
2	254	3	0.012	0.035	240	17	0.066	0.087
	10	102	0.089		15	97	0.134	
3	251	6	0.023	0.033	224	33	0.128	0.111
	6	106	0.054		8	104	0.071	
4	251	6	0.023	0.033	234	23	0.089	0.092
	6	106	0.054		11	101	0.098	
5	250	7	0.027	0.030	225	32	0.125	0.108
	4	108	0.036		8	104	0.071	
6	252	5	0.019	0.033	237	20	0.078	0.084
	7	105	0.063		11	101	0.098	

Table 6
Summary of results on X_{WDBC} with the three best features for second partition

k	k-RNN rule				k-NN rule			
	Confusion matrix		Prob. of false positive false negative	Avg. error rate	Confusion matrix		Prob. of false positive false negative	Avg. error rate
1	249	8	0.031	0.073	212	45	0.175	0.141
	19	93	0.170		7	105	0.063	
2	249	8	0.031	0.046	231	26	0.101	0.111
	19	103	0.080		15	97	0.134	
3	250	7	0.027	0.038	226	31	0.121	0.103
	7	105	0.063		7	105	0.063	
4	252	5	0.019	0.027	241	16	0.062	0.073
	5	107	0.045		11	101	0.098	
5	252	5	0.019	0.027	227	30	0.117	0.106
	5	107	0.045		9	103	0.080	
6	252	5	0.019	0.027	234	23	0.089	0.125
	5	107	0.045		13	99	0.116	

X_{Tre} the test data set. Again X_{Tr} and X_{Te} may be written as $X_{Tr} = X_{Tr,B} \cup X_{Tr,M}$ and $X_{Te} = X_{Te,B} \cup X_{Te,M}$, respectively. Here $X_{Tr,B}$ and $X_{Tr,M}$, respectively, denote the benign and malignant training points; and $X_{Te,B}$ and $X_{Te,M}$, respectively represent the benign and the malignant test data points. We use X_{Tr} for designing the k-RNN and k-NN classifiers. Designing the k-RNN classifier involves estimation of the parameters $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, Σ_1 and Σ_2 of the populations and then computation of scores of the points in X_{Tr} via $\hat{D}(.)$ using the estimated parameters. The classifier is then tested on X_{Te}.

In the present case we used $|X_{Tr,B}| = |X_{Tr,M}| = 100$ and the remaining 369 (112 malignant and 257 benign) points as X_{Te}. That is $|X_{Te}| = 369$ with $|X_{Te,B}| = 257$ and $|X_{Te,M}| = 112$. We have tested both k-RNN and k-NN for different random partitions of X_{WDBC} and the results are pretty consistent across different partitions. We report here four typical cases in Tables 1–4 for $k = 1$–6.

The tables include the confusion matrices exhibiting the number of correct classifications along the diagonal elements and number of false positives and false negatives along the off diagonal elements. We also calculate the probability of false negatives, probability of false positives, and the total (average) probability of misclassifications.

For the first set of training samples the results are summarized in Table 1. For the other three different sets of training samples, results are summarized in Table 2, Table 3, and Table 4, respectively.

Tables 1–4 reveal that the k-RNN rule performs better than the conventional k-NN rule in the majority of the cases. In fact, in the first three tables the k-RNN rule performs

Table 7
Summary of results on X_{WDBC} with the three best features for third partition

k	k-RNN rule				k-NN rule			
	Confusion matrix		Prob. of false positive false negative	Avg. error rate	Confusion matrix		Prob. of false positive false negative	Avg. error rate
1	246	11	0.043	0.051	223	34	0.132	0.141
	8	104	0.071		18	94	0.161	
2	251	6	0.023	0.030	249	8	0.031	0.095
	5	107	0.045		27	85	0.241	
3	251	6	0.023	0.030	235	22	0.086	0.106
	5	107	0.045		17	95	0.152	
4	251	6	0.023	0.030	244	13	0.051	0.089
	5	107	0.045		20	92	0.179	
5	251	6	0.023	0.030	232	25	0.097	0.108
	5	107	0.045		15	97	0.134	
6	246	11	0.043	0.038	239	18	0.070	0.098
	3	109	0.026		18	94	0.161	

Table 8
Summary of results on X_{WDBC} with the three best features for fourth partition

k	k-RNN rule				k-NN rule			
	Confusion matrix		Prob. of false positive false negative	Avg. error rate	Confusion matrix		Prob. of false positive false negative	Avg. error rate
1	255	2	0.008	0.033	221	36	0.140	0.119
	10	102	0.089		8	104	0.071	
2	254	3	0.012	0.022	250	7	0.027	0.057
	5	107	0.045		14	98	0.125	
3	252	5	0.019	0.019	231	26	0.101	0.100
	2	110	0.018		10	102	0.089	
4	251	6	0.023	0.019	239	18	0.070	0.089
	1	111	0.009		15	97	0.134	
5	252	5	0.019	0.022	229	28	0.109	0.106
	3	109	0.027		11	101	0.098	
6	252	5	0.019	0.022	243	14	0.054	0.084
	3	109	0.027		17	95	0.152	

uniformly better than the k-NN rule. The average error rate for the k-RNN rule in these four tables is 0.060 with a standard deviation of 0.017. The average error rate for the k-NN rule in these four tables is 0.073 with a standard deviation of 0.011. The t-test reveals that the performance of the k-RNN rule is significantly better than the k-NN rule, since the test statistic value $t = 3.15$ and the critical value $t_{0.025,46} = 2.013$. Computationally, the k-RNN rule is easy to implement and needs less computational time than the k-NN rule. The best case found by us for the k-RNN rule using all 30 features produces 4% error rate, that is 96% correct classification rate. This result is comparable to the results reported in past by Bennett and Mangasarian [18] which generated 97.5% accuracy by using three best features. We like to stress here that our k-RNN rule used all 30 features for Tables 1–4, while in Ref. [18] only the three best features were used. This indicates a robust behavior of the k-RNN classification rule. To establish this robustness further, next we consider only the three selected features used in Wolberg et al. [16,17]. Recall that these three features are *mean texture*, *worst mean area*, and *worst mean smoothness*. We again use the same computational protocalls for dividing the data set into training and test sets. The results are presented in Tables 5–8 for $k = 1$–6.

From Tables 5–8, we notice that the k-RNN rule performs uniformly better than the k-NN rule. In these four tables the average error rate generated by the k-RNN rule is 0.032 with a standard deviation (s.d.) of 0.011, whereas the k-NN rule produced an average error rate of 0.103 with a s.d. of 0.021. The t-test reveals that average error rate for the

Table 9
Summary of results on X_{WDBC} for the first partition

k	k-RNN rule				k-NN rule			
	Confusion matrix		Prob. of false positive false negative	Avg. error rate	Confusion matrix		Prob. of false positive false negative	Avg. error rate
1	330	14	0.041	0.035	332	12	0.035	0.039
	3	136	0.021		7	132	0.050	
2	329	15	0.043	0.039	236	8	0.023	0.037
	4	135	0.029		10	129	0.072	
3	329	15	0.043	0.039	334	10	0.029	0.025
	4	135	0.029		2	137	0.014	
4	329	15	0.043	0.039	338	6	0.017	0.031
	4	135	0.029		9	130	0.065	
5	329	15	0.043	0.037	334	10	0.029	0.031
	3	136	0.021		5	134	0.036	
6	329	15	0.043	0.037	337	7	0.020	0.039
	3	136	0.021		12	127	0.086	

Table 10
Summary of results on X_{WDBC} for the second partition

k	k-RNN rule				k-NN rule			
	Confusion matrix		Prob. of false positive false negative	Avg. error rate	Confusion matrix		Prob. of false positive false negative	Avg. error rate
1	335	9	0.026	0.041	337	7	0.020	0.033
	11	128	0.079		9	130	0.065	
2	334	10	0.029	0.039	338	6	0.017	0.048
	9	130	0.065		17	122	0.122	
3	334	10	0.029	0.037	335	9	0.026	0.037
	8	131	0.056		9	130	0.065	
4	334	10	0.029	0.037	338	6	0.017	0.041
	8	131	0.056		14	125	0.101	
5	334	10	0.029	0.037	337	7	0.020	0.035
	8	131	0.056		10	129	0.072	
6	334	10	0.029	0.039	338	6	0.017	0.041
	9	130	0.065		14	125	0.101	

k-RNN rule is (highly) significantly smaller than that of the k-NN rule, since the test statistic value $|t| = 14.73$ and the critical value $t_{0.025,46} = 2.013$. These three selected features have widely different domains and variations. The *mean texture* $\in [9, 40]$, *worst mean smoothness* $\in [0.06, 0.223]$, while *worst mean area* $\in [184, 4259]$. Because of the high differences in feature values the k-NN classifier may not perform well. The smaller values contribute very little to the distance based function when there already exists a large valued feature with high variance in the data set. Whereas our method plays a robust role as the score function takes into account the covariance structures of both populations. Here the best error rate for the k-RNN rule is 1.9% (i.e., the best accuracy is 98.1). The best accuracy rate obtained by authors in Refs. [16–18] for this data set is 97.5%. The k-RNN classifier marginally beats the earlier best results.

3.2.2. WBC database

Ignoring the 16 points with missing features, the WBC database resulted in X_{WBC} having 683 points in $\mathbf{R}^9$ with 444 benign and 239 malignant cases. In this case too we used $|X_{Tr,B}| = |X_{Te,M}| = 100$, $X_{Tr} = X_{Tr,B} \cup X_{Tr,M}$, consequently $|X_{Tr}| = 483$. Like X_{WDBC}, we estimated $\hat{D}(.)$ using X_{Tr}, which in turn is used to design the k-RNN classifier and tested on X_{Te}. The experiment was repeated for several random partitions $\{X_{Tr}, X_{Te}\}$ and we report here only two typical results in Tables 9 and 10.

Tables 9 and 10 show that the k-RNN rule performs as good as the conventional k-NN rule or even in some cases

performed better. The best accuracy rate produced by the k-RNN rule for this data set is approximately 97% which is much better than any accuracy rate reported in the past on this data set.

4. Concluding remarks

The k-RNN rule is a nonparametric distribution free classification rule. Through empirical study we have noted that the k-RNN rule performs as well as the k-NN rule, or even better in certain situations. The k-RNN rule is a rank based rule, so it has better robustness property. Since the k-RNN rule is rank based, it is expected to perform better whenever there are too much variations between features. The computational complexity in the k-RNN rule is much less than that of the k-NN rule. Thus the k-RNN rule lessens the burden of computing the distances of all observations from the object to be classified as they are needed for the conventional distance based k-NN rule. Users of the k-NN rule may consider the k-RNN rule as a computationally simpler alternative to the conventional k-NN implementation. Since it is a nonparametric classifier, it can applied to any data set. However, the multivariate k-RNN ranking procedure depends on the distributions' mean vectors and covariance matrices. If the class distribution is characterised by mean and co-variance structure, in particular, if the class distributions are of Gaussian nature, the performance of the k-RNN rule is expected be quite good. Since the family of distributions characterised by mean and co-variances includes many, k-RNN can be applied successfully in many cases.

Acknowledgements

The authors wish to thank the referees for constructive suggestions on the earlier version of the article.

References

[1] J. Dengler, B. Sabine, J.F. Desaga, Segmentation of microcalcifications in mamograms, IEEE Trans. Med. Imaging 12 (4) (1993) 634–642.
[2] L. Shen, R.M. Rangayyan, J.E. Desautels, Application of shape analysis to mammographic calcifications, IEEE Trans. Med. Imaging 13 (2) (1994) 263–274.
[3] N. Karssemeijer, M. Thijssen, J. Hendriks, L.V. Erning, Digital Mamography, Nijmegen, Kluwer Academic Publishers, Boston, 1998.
[4] E. Fix, J.L. Hodges, Nonparametric discrimination: consistency properties: US Air Force School of Aviation Medicine, Report No. 4, Randolph Field, TX, 1951.
[5] B.W. Silverman, M.C. Jones, E. Fix and Hodges (1951): an important contribution to nonparametric discriminant analysis and density estimation. Int. Stat. Rev. 57 (1989) 233–247.
[6] T.M. Cover, P.E. Hart, Nearest neighbor pattern classification, IEEE Trans. Inf. Theory 13 (1967) 21–26.
[7] L. Devroye, On the asymptotic probability in nonparametric discrimination, Ann. Stat. 9 (1981) 1320–1327.
[8] T.W. Anderson, Some nonparametric multivariate procedures based on statistical equivalent blocks, in: P.R. Krishnaiah (Ed.), Proceedings of the First International Symposium Analysis, Academic Press, New York, 1966.
[9] S. Dasgupta, H.E. Lin, Nearest neighbor rules for statistical classifications based on ranks, Sankhya A 42 (1980) 219–230.
[10] S.C. Bagui, B. Vaughn, Statistical classification based on k-rank nearest neighbor rule, Stat. Decisions 16 (1998) 181–189.
[11] S.C. Bagui, N. Pal, A multistage generalization of the rank nearest neighbor classification rule, Pattern Recognition Lett. 16 (1995) 601–614.
[12] R.A. Johnson, D.W. Wichern, Applied Multivariate Statistical Analysis, 4th Edition, Prentice-Hall, Englewood Cliffs, NJ, 1998.
[13] H. Cramer, Mathematical Methods of Statistics, Princeton University, Princeton, 1946.
[14] W.N. Street, W.H. Wolberg, O.L. Mangasarian, Nuclear feature extraction for breast tumor diagnosis, IS & T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology, Vol. 1905, San Jose, CA, 1993, pp. 861–870.
[15] O.L. Mangasarian, W.M. Street, W.H. Wolberg, Breast cancer diagnosis and prognosis via linear programming, Oper. Res. 43 (4) (1995) 570–577.
[16] W.H. Wolberg, W.N. Street, O.L. Mangasarian, Machine learning to diagnose breast cancer from fine-needle aspirates, Cancer Lett. 77 (1994) 163–171.
[17] W.H. Wolberg, W.N. Street, D.M. Heisey, O.L. Mangasarian, Computer-derived nuclear features distinguish malignant from benign breast cytology, Hum. Pathol. 26 (1995) 792–796.
[18] K.P. Bennett, O.L. Mangasarian, Robust linear programming discrimination of two linearly inseparable sets, Optim. Methods Software 1 (1992) 23–34.
[19] O.L. Mangasarian, W.H. Wolberg, Cancer diagnosis via linear programming, SIAM News 23 (5) (1990) 1–18.
[20] W.H. Wolberg, O.L. Mangasarian, Multi-surface method of pattern separation for medical diagnosis applied to breast cytology, Proc. Nat. Acad. Sci. 87 (1990) 9193–9196.
[21] J. Zhang, Selecting typical instances in instance-based learning, Proceedings of the Ninth International Machine Learning Conference, Aberdeen, Morgan Kaufman, Scotland, 1992, pp. 470–479.
[22] R.H. Randles, J.D. Broffitt, J.S. Ramberg, R.V. Hogg, Discriminant analysis based on ranks, J. Am. Stat. Assoc. 73 (362) (1978) 379–384.

About the Author—SUBHASH C. BAGUI received his B.Sc. in Statistics from University of Calcutta in 1979, his M. Stat. in Statistics from Indian Statistical Institute in 1982, and his Ph.D. in Statistics from University of Alberta, Edmonton, Alberta, in 1989. He is currently a Professor in the Department of Mathematics and Statistics at The University of West Florida, Pensacola, Florida. He has authored a book

titled, *Handbook of Percentiles of Noncentral t-Distributions*, published by CRC Press. His research interests include statistical classification and pattern recognition, bio-statistics, construction of designs, tolerance regions and reliability. Dr. Subhash Bagui is an Associate Editor of the *Journal of Applied Statistical Science*.

About the Author—SIKHA S. BAGUI received her B.S. from Cuttington University, Monrovia, Liberia, in 1984, MBA in Information Systems from University of Toledo in 1986, and Ed.D. in Computer Science from University of West Florida, Pensacola, Florida, in 2000. Dr. Sikha Bagui is currently a lecturer in the Department of Computer Science at the University of West Florida. She has co-authored a book titled, "*Learning SQL: A Step-By-Step Guide using Oracle*", Addition Wesly (2002). Her research interests include database, data mining, pattern recognition, image processing, multimedia, and computer education.

About the Author—KUHU PAL obtained B.Sc. degree with honors in Physics in 1984 from the University of Burdwan and M.Sc. and Ph.D. degrees in Physics from Banaras Hindu University in 1987 and 1993, respectively. After that she worked as a Research Associate first in the Physics Department of Banaras Hindu University and then from September 1995 in the Machine Intelligence Unit of Indian Statistical Institute, Calcutta. In September 1999 she joined the MCKV Institute of Engineering as a lecturer and left that for visiting the Computer Science Department of the University of West Florida, USA from January 2000 for a period of six months. Her research interest includes pattern recognition, fuzzy sets theory, fuzzy logic controllers, neural networks, and computational material science.

About the Author—NIKHIL R. PAL obtained B.Sc. with honors in Physics and Master of Business Management from the University of Calcutta in 1979 and 1982, respectively. He obtained M. Tech. (Computer Science) and Ph.D. (Computer Science) from the Indian Statistical Institute in 1984 and 1991, respectively. Currently he is a Professor in the Electronics and Communication Sciences Unit of the Indian Statistical Institute, Calcutta and he is the present Professor-in-charge of the Computer and Communication Sciences Division. From September 1991 to February 1993, July 1994 to December 1994, October 1996 to December 1996, January 2000 to July 2000 he visited the Computer Science Department of the University of West Florida. He was a guest faculty of the University of Calcutta also. His research interest includes image processing, pattern recognition, fuzzy sets theory, measures of uncertainty, neural networks, genetic algorithms, and fuzzy logic controllers. He has co-authored a book titled *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*, Kluwer Academic Publishers, 1999, co-edited a volume *Advances in Pattern Recognition and Digital Techniques*, *ICAPRDT99* and edited a book titled *Pattern Recognition in Soft Computing Paradigm*, World Scientific, 2001. He is an associate editor of the *International Journal of Fuzzy Systems*, *International Journal of Approximate Reasoning*, *IEEE Transactions on Fuzzy Systems*, *IEEE Transactions on Systems Man and Cybernetics-B* (*Electronic version*) and an area editor of *Fuzzy Sets and Systems*.