

AD-A049 548

MINNESOTA UNIV MINNEAPOLIS DEPT OF THEORETICAL STATISTICS F/6 12/1
SIMULATION STUDIES ON SOME NEAREST NEIGHBOR RULES FOR STATISTIC--ETC(U)
NOV 77 S DAS GUPTA, D AARONS DAAG-29-76-G-0038

UNCLASSIFIED

UMN/DTS/TR-303

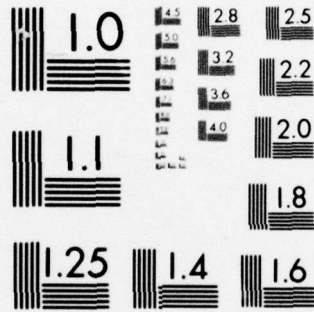
ARO-13149.4-M

NL

| OF |
AD
A049548



END
DATE
FILMED
3 -78
DDC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

ARO 13149.4-M

(Handwritten circled number 12)
AP



AD A 049548

AD No. _____
JDC FILE COPY

(Handwritten circled note)
See 1473
in book

UNIVERSITY OF MINNESOTA

SCHOOL OF
STATISTICS



DDC
RECEIVED
FEB 8 1978
REGISTERED

DISTRIBUTION STATEMENT A
Approved for public release
Distribution Unlimited

12

Simulation Studies on Some Nearest Neighbor Rules
for Statistical Classification. (1)

By

David Aarons and Somesh Das Gupta

University of Minnesota
Technical Report No. 303

November 1977

ACCESSION for	
NTIS	White Section <input checked="" type="checkbox"/>
DDC	Buff Section <input type="checkbox"/>
UNANNOUNCED	<input type="checkbox"/>
JUSTIFICATION	
BY	
DISTRIBUTION/AVAILABILITY CODES	
Dist.	AVAIL. and/or SPECIAL
A	

DDC
RECEIVED
FEB 3 1978
D

DISTRIBUTION STATEMENT A

Approved for public release;
Distribution Unlimited

Simulation Studies on Some Nearest Neighbor Rules
for Statistical Classification. ⁽¹⁾

by

David Aarons and Somesh DasGupta
University of Minnesota

(1) This research was supported by a grant from the Mathematics Division,
U.S. Army, Research Triangle Park, N.C.; Grant DAAG 29-76-G-0038.

1. Introduction. The two-population classification problem is to identify a population π_0 with one of two given populations π_1 and π_2 based on observations from these populations on a random vector X . We shall consider here X to be univariate. Let F_i be the c.d.f. of X in π_i ($i = 0, 1, 2$). Thus our problem is to test $H_1: F_0 = F_1$ vs. $H_2: F_0 = F_2$. In this paper we have considered some rules which are suggested in the literature when F_1, F_2 are not known except that they are continuous. We have studied the performances of the following three rules by simulation.

Let X_0, X_{1i} ($i = 1, \dots, n_1$), X_{2i} ($i = 1, \dots, n_2$) be random observations on X from the populations π_0, π_1, π_2 , respectively.

Rule I. 1-NN (nearest neighbor) Rule: Measure distances of X_0 from X_{1i} 's and X_{2i} 's and based on these distances classify X_0 into the population to which its nearest neighbor belongs.

Rule II. 1-RNN (rank nearest neighbor) Rule: Pool all the observations and order them.

(a) If X_0 is the largest or the smallest observation classify X_0 into the population of its nearest neighbor (based on ranks).

(b) If both the right-hand and the left-hand nearest neighbor of X_0 (denoted by U_1 and V_1) belong to the same population, classify X_0 into that population.

(c) If U_1 and V_1 belong to different populations classify X_0 into π_1 and π_2 with probabilities $1/2$ and $1/2$, respectively. (We call this case a "tie".)

Rule III. 2-RNN Rule: Apply the 1-RNN rule. If a tie occurs, delete the observations corresponding to U_1 and V_1 and apply the 1-RNN rule again on the remaining observations.

The first rule was suggested and studied by Fix and Hodges (1951, 1953). DasGupta and Lin (1977) proposed the RNN rules and obtained the asymptotic probabilities of misclassification as $n_1, n_2 \rightarrow \infty$. For a given rule δ , let its PMC under $F_0 = F_1$ be given by

$$\alpha(\delta) = \Pr[\delta \text{ classifies } X_0 \text{ into } \pi_2 | F_0 = F_1] .$$

Let $\alpha_1^*, \alpha_2^*, \alpha_3^*$ be the asymptotic values of α corresponding to the above rules 1, 2 and 3. Let f_i be the p.d.f. of F_i with respect to Lebesgue measure ($i = 1, 2$) and $p_i = \lim n_i / (n_1 + n_2)$ ($i = 1, 2$) as $\min(n_1, n_2) \rightarrow \infty$. It was shown by Fix and Hodges (1951) and DasGupta and Lin (1977) that

$$\alpha_1^* = \alpha_2^* = \int_{-\infty}^{\infty} p_2 f_1(x) f_2(x) dx / \{p_1 f_1(x) + p_2 f_2(x)\}$$

$$\alpha_3^* = \alpha_2^* + \int_{-\infty}^{\infty} \frac{p_1 p_2 f_1(x) f_2(x) \cdot \{p_2 f_2(x) - p_1 f_1(x)\}}{\{p_1 f_1(x) + p_2 f_2(x)\}^3} f_1(x) dx .$$

In this paper we have studied the finite-sample performances of these rules by estimating α based on samples from sets of two given populations.

2. The Experiment. Different steps of our simulation study are given below.

(i) Two known but different univariate distributions F_1 and F_2 are chosen.

(ii) Random samples of sizes n_1 and n_2 from F_1 and F_2 , respectively, are obtained; these samples are called training samples.

(iii) A random sample of size n_0 from $F_0 = F_1$ is obtained. We call this a test sample.

(iv) For each observation in the test sample a given classification rule δ (one of the above three rules) is applied and let n_{02} be the number of the observations in the test sample which are classified by δ

into F_2 . Let $\hat{\alpha}(\delta) = n_{02}/n_0$ be the proportion of test samples misclassified into F_2 .

(v) Steps (ii)-(iv) are repeated r times for new training and test samples keeping n_1 , n_2 and n_0 fixed.

(vi) The mean and the standard error of the mean based on r values of $\hat{\alpha}(\delta)$ thus obtained are recorded.

(vii) Steps (ii)-(vi) are repeated for different values of n_1 , n_2 and r .

(viii) F_2 is characterized by a parameter θ . For different values of θ steps (i)-(vii) are repeated.

Our choices are given in the following table.

F_1	F_2	Parameters	$n_1=n_2$	n_0	r
$N(0,1)$	$N(\theta,1)$	$\theta=0, \pm 1, \pm 2, 3$	25 100	100 400	20 4
$N(0,1)$	$N(0,\theta)$	$\theta=2, 3, 1/2, 1/3$	25 100	100 400	20 4
e^{-x} (density)	$\theta e^{-\theta x}$	$\theta=1, 2, 3, 4,$ $1/2, 1/3, 1/4, 1/8$	100	100	20
Cauchy (0,1)	Cauchy ($\theta,1$)	$\theta=0, \pm 1, \pm 2, \pm 3$	25 100	100 400	20 4

Samples are generated by a library subroutine available on the CDC 6400 at the University of Minnesota.

Note 1. In the following tables "Half" refers to taking one-half the number of ties to count as misclassified and "R-half" refers to resolving the ties by the use of uniform random number generator.

Note 2. In some of the following tables EPMC denotes an estimate of the asymptotic PMC ($\alpha_1^* = \alpha_2^*$) of the 1-NN and 1-RNN rules. These are derived by the method of runs as suggested in Das Gupta and Lin (1977).

3. Tables

Table 3.1

Proportion of test sample misclassified into π_2 .

$F_1 = N(0,1)$, $F_2 = N(\theta,1)$; $n_1 = n_2 = 25$, $n_0 = 100$, $r = 20$.

Optimal (assuming θ is known and for minimax rule) PMC is

$\phi(-|\theta|/2)$.

θ \ Rule	1NN		RNN		2-RNN		Opt. Exp't. PMC
	MEAN	s.e.	MEAN	s.e.	MEAN	s.e.	
$\theta = 0$.479	.017	Half .479 Rhalf .485	.013 .016	Half .479 Rhalf .484	.014 .015	.500
$\theta = 1$.374	.018	Half .381 Rhalf .374	.014 .016	Half .343 Rhalf .340	.021 .021	.308
$\theta = -1$.426	.020	Half .426 Rhalf .432	.014 .017	Half .421 Rhalf .425	.025 .024	.308
$\theta = 2$.195	.018	Half .194 Rhalf .196	.018 .018	Half .165 Rhalf .164	.017 .018	.159
$\theta = -2$.245	.020	Half .254 Rhalf .251	.018 .018	Half .258 Rhalf .255	.019 .018	.159
$\theta = 3$.086	.012	Half .089 Rhalf .084	.012 .011	Half .062 Rhalf .061	.010 .009	.067
$\theta = -3$.105	.013	Half .114 Rhalf .113	.012 .011	Half .119 Rhalf .118	.015 .015	.067

Table 3.2

Proportion of test sample misclassified into π_2 .

$F_1 = N(0,1), F_2 = N(\theta,1); n_1 = n_2 = 100, n_0 = 400, r = 4.$

θ \ Rule	1NN		RNN		EPMC	2-RNN		Opt. Exp't.
	MEAN	s.d.	MEAN	s.d.		MEAN	s.d.	PMC
$\theta = 0$.490	.018	Half .482 .008 Rhalf .475 .006		.48	Half .509 .014 Rhalf .501 .016		.500
$\theta = 1$.415	.010	Half .398 .014 Rhalf .404 .024		.36	Half .351 .009 Rhalf .358 .024		.308
$\theta = -1$.402	.010	Half .394 .007 Rhalf .397 .007		.38	Half .347 .025 Rhalf .344 .024		.308
$\theta = 2$.208	.010	Half .210 .010 Rhalf .208 .009		.22	Half .200 .011 Rhalf .199 .012		.159
$\theta = -2$.209	.012	Half .213 .008 Rhalf .215 .009		.22	Half .197 .013 Rhalf .200 .014		.159
$\theta = 3$.088	.011	Half .083 .009 Rhalf .082 .007		.10	Half .065 .005 Rhalf .066 .006		.007
$\theta = -3$.104	.012	Half .101 .008 Rhalf .107 .013		.09	Half .088 .012 Rhalf .094 .014		.007

Table 3.3.

Proportion of test sample misclassified into π_2 .

$F_1 = N(0,1), F_2 = N(0,\theta); n_1 = n_2 = 25, n_0 = 100. r = 20.$

θ \ Rule	1NN		RNN		2-RNN			
	MEAN	s.e.	MEAN	s.e.	MEAN	s.e.		
$\theta = 2.0$.375	.009	Half	.394	.008	Half	.353	.014
			Rhalf	.393	.010	Rhalf	.355	.015
$\theta = 3.0$.399	.014	Half	.346	.013	Half	.293	.019
			Rhalf	.337	.013	Rhalf	.295	.018
$\theta = .5$.417	.017	Half	.438	.015	Half	.461	.020
			Rhalf	.337	.018	Rhalf	.460	.021
$\theta = 1/3$.359	.022	Half	.376	.018	Half	.393	.019
			Rhalf	.380	.019	Rhalf	.391	.019

Table 3.4

Proportion of test sample misclassified into π_2 .

$F_1 = N(0,1), F_2 = N(0,\theta); n_1 = n_2 = 100, n_0 = 400, r = 4.$

θ \ Rule	1NN		RNN		EPMC	2-RNN			
	MEAN	s.e.	MEAN	s.e.		MEAN	s.e.		
$\theta = 2.0$.435	.022	Half	.424	.022	.36	Half	.395	.027
			Rhalf	.426	.025		Rhalf	.396	.028
$\theta = 3.0$.333	.012	Half	.338	.010	.32	Half	.295	.012
			Rhalf	.336	.011		Rhalf	.296	.011
$\theta = .5$.397	.062	Half	.405	.011	.38	Half	.409	.006
			Rhalf	.407	.013		Rhalf	.408	.005
$\theta = 1/3$.339	.021	Half	.352	.020	.35	Half	.360	.029
			Rhalf	.354	.021		Rhalf	.361	.030

Table 3.5

Proportion of test sample misclassified into π_2 .

$$f_1(x) = e^{-x}, f_2(x) = \theta e^{-\theta x}; n_1 = n_2 = n_0 = 100, r = 4.$$

θ \ Rule	1NN		RNN		EPMC	2-RNN	
	MEAN	s.e.	MEAN	s.e.		MEAN	s.e.
$\theta = 1$.508	.016	Half .509 Rhalf .523	.013 .016	.47	Half .503 Rhalf .517	.013 .015
$\theta = 2$.442	.015	Half .434 Rhalf .438	.014 .017	.38	Half .442 Rhalf .444	.016 .016
$\theta = 3$.402	.014	Half .388 Rhalf .387	.011 .011	.36	Half .394 Rhalf .387	.013 .014
$\theta = 4$.335	.009	Half .330 Rhalf .336	.007 .008	.32	Half .327 Rhalf .330	.009 .009
$\theta = .5$.453	.010	Half .453 Rhalf .458	.009 .013	.38	Half .430 Rhalf .430	.010 .014
$\theta = 1/3$.410	.011	Half .395 Rhalf .386	.008 .009	.36	Half .346 Rhalf .335	.010 .010
$\theta = 1/4$.354	.015	Half .364 Rhalf .372	.012 .013	.32	Half .290 Rhalf .292	.013 .013
$\theta = 1/8$.247	.014	Half .248 Rhalf .259	.012 .014	.22	Half .181 Rhalf .185	.011 .010

Table 3.6

Proportion of test sample misclassified into π_2 .

$F_1 = \text{Cauchy}(0,1)$, $F_2 = \text{Cauchy}(\theta,1)$; $n_1 = n_2 = 25$, $n_0 = 100$, $r = 20$.

θ / Rule	1NN		RNN		2-RNN			
	MEAN	s.e.	MEAN	s.e.	MEAN	s.e.		
$\theta = 0$.473	.018	Half	.430	.015	Half	.488	.027
			Rhalf	.493	.018	Rhalf	.505	.029
$\theta = 1$.406	.022	Half	.418	.022	Half	.397	.031
			Rhalf	.408	.025	Rhalf	.395	.033
$\theta = -1$.398	.016	Half	.410	.012	Half	.389	.021
			Rhalf	.410	.013	Rhalf	.385	.022
$\theta = 2$.288	.021	Half	.297	.021	Half	.248	.027
			Rhalf	.288	.021	Rhalf	.238	.028
$\theta = -2$.247	.012	Half	.264	.012	Half	.248	.017
			Rhalf	.276	.015	Rhalf	.252	.019
$\theta = 3$.161	.020	Half	.168	.017	Half	.103	.017
			Rhalf	.161	.018	Rhalf	.099	.017
$\theta = -3$.153	.015	Half	.156	.013	Half	.130	.014
			Rhalf	.154	.013	Rhalf	.125	.014

Table 3.7

Proportion of test sample misclassified into π_2 .

$F_1 = \text{Cauchy}(0,1)$, $F_2 = \text{Cauchy}(\theta,1)$; $n_1 = n_2 = 100$, $n_0 = 400$, $r = 4$.

θ / Rule	1NN		RNN		2-RNN	
	MEAN	s.e.	MEAN	s.e.	MEAN	s.e.
$\theta = 0$.494	.015	Half .514 Rhalf .529	.013 .014	Half .506 Rhalf .512	.017 .021
$\theta = 1$.411	.010	Half .426 Rhalf .446	.009 .018	Half .381 Rhalf .390	.018 .017
$\theta = -1$.457	.029	Half .446 Rhalf .454	.033 .025	Half .394 Rhalf .393	.028 .025
$\theta = 2$.284	.007	Half .278 Rhalf .283	.008 .009	Half .217 Rhalf .219	.033 .024
$\theta = -2$.152	.016	Half .318 Rhalf .321	.022 .015	Half .254 Rhalf .257	.014 .010
$\theta = 3$.152	.016	Half .154 Rhalf .417	.015 .012	Half .088 Rhalf .087	.018 .014
$\theta = -3$.204	.034	Half .199 Rhalf .198	.032 .034	Half .105 Rhalf .103	.011 .012

4. Concluding Remarks. For all the three rules considered, it seems that $\hat{\alpha}_1$ has a definite tendency to decrease as θ moves away (in either direction) from its value under F_1 .

For small $n_1 = n_2$ there is not any marked difference in performances of these three rules although the 2-RNN rule may be a bit better. However, for large $n_1 = n_2$ the 2-RNN rule seem to have markedly better performance except for the cases $N(0,1)$ vs. $N(0,\theta)$, $\theta < 1$. This report is the first empirical study on the performances of 1NN and RNN rules, although a more detailed study especially on multi-stage RNN rules is called for.

References

1. Das Gupta, S. and Lin, H. E. (1977). Nearest neighbor rules for statistical classification based on ranks. Tech. Rep. 285, School of Statistics, University of Minnesota, Minneapolis, Minnesota.
2. Fix, E. and Hodges, J. L. (1951). Nonparametric discrimination: Consistency properties. U.S. Air Force School of Aviation Medicine. Report No. 4. Randolph Field, Texas.
3. Fix, E. and Hodges, J. L. (1953). Nonparametric discrimination. Small sample properties, Ibid., Report No. 11.

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM	
1. REPORT NUMBER (14) UMN/DTS/TR303 Technical Report No. 303	2. JOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle) (6) Simulation Studies on Some Nearest Neighbor Rules for Statistical Classification,	5. TYPE OF REPORT & PERIOD COVERED (9) Technical Report		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) (10) Somesh/Das Gupta and David/Aarons	8. CONTRACT OR GRANT NUMBER(s) (15) DAAG-29-76-G-1038 ✓		
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Theoretical Statistics University of Minnesota Minneapolis, MN 55455 ✓	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS		
11. CONTROLLING OFFICE NAME AND ADDRESS U.S. Army Research Office P.O. Box 12211 Research Triangle Park, NC 27709 ✓	12. REPORT DATE (11) November 1977	13. NUMBER OF PAGES 11 (12) 15 p.	15. SECURITY CLASS. (of this report) Unclassified
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) (18) ARO (19) 13149.4-M	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE NA		
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited.			
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report) NA			
18. SUPPLEMENTARY NOTES The findings in this report are not to be construed as an official Department of Army position, unless so designated by other authorized documents.			
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Classification, nearest neighbor rule, rank nearest nearest neighbor rule, simulation, probability of misclassification.			
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) Abstract: The performances of the distance nearest neighbor rule, one-stage rank nearest neighbor rule and two-stage rank nearest neighbor rule for the two-population statistical classification problem are studied through simulation processes. Normal distributions with different mean, normal distributions with different variances, exponential distributions and Cauchy distributions are considered.			