Rigorous Non-Disjoint Discretization for Naive Bayes Supplementary Material

Huan Zhang, Liangxiao Jiang, Geoffrey I. Webb

1. Toy Example

For ease of understanding, in this section, we will give a toy example to show how RNDD works step by step. Assume in a binary classification problem, the training data size is 1156. Among them, 800 instances belong to class 0, and 356 instances belong to class 1. Part of the sorted value distribution of the quantitative attribute A_j is listed in Table 1. We assume that 900 instances in this dataset are with known values of a_j, and the remaining 256 instances are with missing values. Based on the framework in **Algorithm 1** and **Algorithm 2** in the paper, we show the detailed discretization processes and classification processes for A_j as follows.

Training phase:

Line 4: Shown in Table 1.

Line 5: $N_j = 900$

Line 6: $I_j = 3 * \sqrt{900} = 90$

Line 7: $\theta_j^{Sin} = \frac{\sqrt{900}}{3} = 10$

Line 9: $I_j^{Sin} = 2$ ($a_j = 0$ and $a_j = 5.4$)

Line 10: $N_j^{Sin} = 448 + 12 = 460$

Line 11: $CP_j^{Sin} = \{\frac{-1+0}{2}, \frac{0+1}{2}, \frac{5.2+5.4}{2}, \frac{5.4+6}{2}\}$

Line 13: $I_j^{NSin} = 90 - 2 = 88$

Line 14: $N_j^{NSin} = 900 - 460 = 440$

Line 15: $\theta_j^{NSin} = \frac{440}{88} = 5$

Line 17: $CP_j^{NSin} = \{\frac{2.5+4.5}{2}, \frac{5+5.2}{2}, \frac{5.2+5.4}{2}, \frac{11+13}{2}, \frac{15+16.6}{2}\}$

Table 1: The value distribution of a quantitative attribute in the toy example.

Value	...	-1	0	1	2.5	4.5	5	5.2	5.4	6	8	11	13	15	16.6	...
Frequency	...	2	448	4	4	2	3	5	12	2	2	3	4	4	2	...
Class 0	...	0	300	3	1	1	3	3	2	0	2	2	4	2	0	...
Class 1	...	2	148	1	3	1	0	2	10	2	0	1	0	2	2	...

Table 2: The generated atomic intervals in the toy example.

Interval	...	[..., -0.5)	[-0.5, 0.5)	[0.5, 3.5)	[3.5, 5.1)	[5.1, 5.3)	[5.3, 5.7)	[5.7, 12)	[12, 15.8)	[15.8, ...)	...
Frequency	...	...	448	8	5	5	12	7	8	...	...
Class 0	...	...	300	4	4	3	2	4	6	...	...
Class 1	...	...	148	4	1	2	10	3	2	...	...

Line 18: $\mathbf{CP_j} = \{\frac{-1+0}{2}, \frac{0+1}{2}, \frac{2.5+4.5}{2}, \frac{5+5.2}{2}, \frac{5.2+5.4}{2}, \frac{5.4+6}{2}, \frac{11+13}{2}, \frac{15+16.6}{2}\}$

Line 22: Shown in Table 2.

Line 23: $\theta^{Cla} = MAX(30, \sqrt{1156}) = 34$

Classification phase:

Assume after discretization, we generate 80 atomic intervals for A_j. For two test instances values 0 and 5.4, their detailed classification processes are listed as follows.

Test instance #1: $a_j = 0$:

Line 3: The atomic interval that a_j located in is [-0.5, 0.5), and its frequency $f_{jr} = 448$

Line 4: Since $448 > 34$, we directly estimate the conditional probability:

$$P(a_j = 0 \mid c = 0) = \frac{300 + \frac{1}{80}}{800 + 1} \approx 0.3745 \tag{1}$$

$$P(a_j = 0 \mid c = 1) = \frac{148 + \frac{1}{80}}{356 + 1} \approx 0.4146 \tag{2}$$

Test instance #2: $a_j = 5.4$:

Line 3: The atomic interval that a_j located in is [5.3, 5.7), and its frequency

$f_{jr} = 12$

Line 4: Since $12 < 34$, next step is Line 7

Line 7: $\theta_L^{Cla} = \theta_R^{Cla} = \frac{34-12}{2} = 11$

Line 8:

$$f_{j,0} = 2 + \underbrace{\left(3 * \frac{1}{2} * 1.0\right) + \left(4 * \frac{1}{2^2} * 1.0\right) + \left(4 * \frac{1}{2^3} * \frac{11-(5+5)}{8}\right)}_{\text{left adjacent atomic intervals}} + \underbrace{\left(4 * \frac{1}{2} * 1.0)\right) + \left(6 * \frac{1}{2^2} * \frac{11-7}{8}\right)}_{\text{right adjacent atomic intervals}} = \frac{117}{16} \tag{3}$$

$$f_{j,1} = 10 + \underbrace{\left(2 * \frac{1}{2} * 1.0\right) + \left(1 * \frac{1}{2^2} * 1.0\right) + \left(4 * \frac{1}{2^3} * \frac{11-(5+5)}{8}\right)}_{\text{left adjacent atomic intervals}} + \underbrace{\left(3 * \frac{1}{2} * 1.0)\right) + \left(2 * \frac{1}{2^2} * \frac{11-7}{8}\right)}_{\text{right adjacent atomic intervals}} = \frac{209}{16} \tag{4}$$

Line 9: Estimate the conditional probability:

$$P(a_j = 5.4 \mid c = 0) = \frac{\frac{117}{16} + \frac{1}{80}}{700 + 1} \approx 0.0104 \tag{5}$$

$$P(a_j = 5.4 \mid c = 1) = \frac{\frac{209}{16} + \frac{1}{80}}{324 + 1} \approx 0.0402 \tag{6}$$

2. Tables of the Experimental Section

Table 3: Descriptions of the 29 UCI datasets used in the experiments.

Dataset Name	Instance Number	Attribute Number	Quantitative Attributes (%)	Class Number	Missing Values
Labor-Negotiations	57	16	50.00	2	Y
Echocardiogram	74	6	83.33	2	Y
Iris	150	4	100.00	3	N
Hepatitis	155	19	31.58	2	Y
Wine-Recognition	178	13	100.00	3	N
Sonar	208	60	100.00	2	N
Glass-Identification	214	9	100.00	3	N
Heart-Disease-(Cleveland)	270	13	53.85	2	N
Liver-Disorders	345	6	100.00	2	N
Ionosphere	351	34	100.00	2	N
Horse-Colic	368	21	33.33	2	Y
Credit-Screening-(Australia)	690	15	40.00	2	Y
Breast-Cancer-Wisconsin	699	9	100.00	2	N
Pima-Indians-Diabetes	768	8	100.00	2	N
Vehicle	846	18	100.00	4	N
Anneal	898	38	15.79	6	Y
German	1000	20	35.00	2	N
Multiple-Features	2000	6	50.00	10	N
Hypothyroid	3163	25	28.00	2	Y
Satimage	6435	36	100.00	6	N
Musk	6598	166	100.00	2	N
Pioneer-1-Mobile-Robot	9150	36	80.56	57	N
Handwritten-Digits	10992	16	100.00	10	N
Australian-Sign-Language	12546	8	100.00	3	N
Letter-Recognition	20000	16	100.00	26	N
Adult	48842	14	42.86	2	Y
Ipums.la.99	88443	60	33.33	13	N
Census-Income	299285	41	19.51	2	Y
Forest-Covertype	581012	54	18.52	7	N

Table 4: NB model's error rate comparisons for RNDD versus ENDD, NDD, PKID, EFD and EWD.

Dataset	RNDD	ENDD	NDD	PKID	EFD	EWD
Labor-Negotiations	0.0667	0.0600	0.0600	0.0736	0.1408	0.1547
Echocardiogram	0.2622	0.2973	0.2946	0.2703	0.2892	0.3351
Iris	0.0587	0.0640	0.0640	0.0760	0.0773	0.0560
Hepatitis	0.1509	0.1587	0.1587	0.1535	0.1535	0.1510
Wine-Recognition	0.0213	0.0292	0.0292	0.0270	0.0382	0.0315
Sonar	0.2558	0.2837	0.2837	0.2644	0.2577	0.2625
Glass-Identification	0.3234	0.3215	0.3234	0.3505	0.3477	0.4047
Heart-Disease-(Cleveland)	0.1756	0.1822	0.1911	0.1726	0.1778	0.1881
Liver-Disorders	0.3705	0.3901	0.3907	0.4034	0.3924	0.3988
Ionosphere	0.1094	0.1032	0.1037	0.1054	0.1049	0.1014
Horse-Colic	0.1875	0.1929	0.1929	0.1929	0.1924	0.1957
Credit-Screening-(Australia)	0.1441	0.1455	0.1449	0.1441	0.1464	0.1554
Breast-Cancer-Wisconsin	0.0266	0.0286	0.0295	0.0263	0.0266	0.0260
Pima-Indians-Diabetes	0.2620	0.2602	0.2586	0.2651	0.2581	0.2471
Vehicle	0.3785	0.3953	0.3988	0.3939	0.3898	0.3965
Anneal	0.0394	0.0488	0.0517	0.0563	0.0372	0.0423
German	0.2578	0.2564	0.2564	0.2586	0.2578	0.2544
Multiple-Features	0.3171	0.3127	0.3127	0.3160	0.3143	0.3053
Hypothyroid	0.0176	0.0173	0.0173	0.0180	0.0249	0.0378
Satimage	0.1759	0.1767	0.1774	0.1765	0.1869	0.1886
Musk	0.0754	0.0843	0.0869	0.0835	0.1632	0.1600
Pioneer-1-Mobile-Robot	0.0212	0.0238	0.0255	0.0238	0.0983	0.0925
Handwritten-Digits	0.1170	0.1171	0.1185	0.1200	0.1253	0.1262
Australian-Sign-Language	0.3593	0.3576	0.4525	0.3593	0.3659	0.3798
Letter-Recognition	0.2579	0.2599	0.3288	0.2602	0.2668	0.2988
Adult	0.1677	0.1718	0.1722	0.1727	0.1719	0.1824
Ipums.la.99	0.1641	0.1650	0.1591	0.1670	0.1760	0.1683
Census-Income	0.2319	0.2302	0.2271	0.2333	0.2340	0.2454
Forest-Covertype	0.3171	0.3181	0.3169	0.3173	0.3291	0.3238

Table 5: NB model's Brier score comparisons for RNDD versus ENDD, NDD, PKID, EFD and EWD.

Dataset	RNDD	ENDD	NDD	PKID	EFD	EWD
Labor-Negotiations	0.1027	0.0781	0.0775	0.1225	0.2186	0.2408
Echocardiogram	0.3981	0.5008	0.5006	0.4005	0.4972	0.5659
Iris	0.0894	0.0985	0.0985	0.1194	0.1317	0.0858
Hepatitis	0.2478	0.2686	0.2687	0.2543	0.2656	0.2542
Wine-Recognition	0.0333	0.0472	0.0472	0.0458	0.0650	0.0506
Sonar	0.4531	0.5283	0.5283	0.4733	0.4654	0.4722
Glass-Identification	0.4895	0.5353	0.5353	0.5434	0.5579	0.5854
Heart-Disease-(Cleveland)	0.2696	0.2884	0.2937	0.2707	0.2773	0.2860
Liver-Disorders	0.4685	0.4968	0.4968	0.5164	0.5073	0.5089
Ionosphere	0.2076	0.1972	0.1973	0.1986	0.2015	0.1867
Horse-Colic	0.3243	0.3322	0.3321	0.3343	0.3292	0.3287
Credit-Screening-(Australia)	0.2301	0.2305	0.2293	0.2345	0.2350	0.2329
Breast-Cancer-Wisconsin	0.0516	0.0543	0.0532	0.0506	0.0513	0.0504
Pima-Indians-Diabetes	0.3777	0.3715	0.3705	0.3836	0.3714	0.3479
Vehicle	0.6442	0.6606	0.6627	0.6716	0.6533	0.6515
Anneal	0.0616	0.0803	0.0827	0.0915	0.0533	0.0634
German	0.3623	0.3589	0.3565	0.3618	0.3609	0.3596
Multiple-Features	0.4666	0.4626	0.4626	0.4609	0.4381	0.4374
Hypothyroid	0.0298	0.0293	0.0293	0.0303	0.0405	0.0558
Satimage	0.3397	0.3390	0.3403	0.3405	0.3577	0.3609
Musk	0.1478	0.1640	0.1683	0.1629	0.3133	0.3075
Pioneer-1-Mobile-Robot	0.0363	0.0409	0.0433	0.0403	0.1619	0.1466
Handwritten-Digits	0.2050	0.2038	0.2050	0.2088	0.2139	0.2155
Australian-Sign-Language	0.4750	0.4737	0.5333	0.4761	0.4873	0.5025
Letter-Recognition	0.3771	0.3790	0.4440	0.3796	0.3878	0.4244
Adult	0.2453	0.2516	0.2513	0.2527	0.2503	0.2656
Ipums.la.99	0.2780	0.2790	0.2663	0.2820	0.2933	0.2752
Census-Income	0.4163	0.4130	0.4061	0.4187	0.4199	0.4406
Forest-Covertype	0.4480	0.4506	0.4466	0.4481	0.4596	0.4562

Table 6: NB's model error rate comparisons for NDD versus RNDD and its some variants. RNDD-W is the method of RNDD minus its weighting strategy. RNDD-D is the method of RNDD minus its division strategy. RNDD ($k = 5$) and RNDD ($k = 7$) are the variants of RNDD under different k. RNDD ($w = \frac{1}{p}$) and RNDD ($w = \frac{1}{3^p}$) are the variants of RNDD under different weighting strategy.

Dataset	RNDD	NDD+D	NDD+W	RNDD (k=5)	RNDD (k=7)	RNDD ($w = \frac{1}{p}$)	RNDD ($w = \frac{1}{3^p}$)	NDD
Labor-Negotiations	0.0667	0.0600	0.0736	0.0667	0.0667	0.1018	0.0740	0.0600
Echocardiogram	0.2622	0.2649	0.2568	0.2784	0.2784	0.2541	0.2811	0.2946
Iris	0.0587	0.0547	0.0693	0.0573	0.0573	0.0560	0.0600	0.0640
Hepatitis	0.1509	0.1549	0.1445	0.1406	0.1407	0.1523	0.1445	0.1587
Wine-Recognition	0.0213	0.0292	0.0213	0.0191	0.0191	0.0169	0.0213	0.0292
Sonar	0.2558	0.2606	0.2673	0.2654	0.2750	0.2394	0.2673	0.2837
Glass-Identification	0.3234	0.3234	0.3103	0.3206	0.3140	0.3252	0.3252	0.3234
Heart-Disease-(Cleveland)	0.1756	0.1896	0.1704	0.1815	0.1756	0.1667	0.1859	0.1911
Liver-Disorders	0.3705	0.3925	0.3838	0.3733	0.3762	0.3774	0.3699	0.3907
Ionosphere	0.1094	0.1032	0.1151	0.1100	0.1106	0.1077	0.1134	0.1037
Horse-Colic	0.1875	0.1946	0.1880	0.1886	0.1864	0.1913	0.1886	0.1929
Credit-Screening-(Australia)	0.1441	0.1441	0.1429	0.1441	0.1449	0.1420	0.1446	0.1449
Breast-Cancer-Wisconsin	0.0266	0.0286	0.0263	0.0266	0.0266	0.0266	0.0263	0.0295
Pima-Indians-Diabetes	0.2620	0.2544	0.2620	0.2643	0.2609	0.2599	0.2648	0.2586
Vehicle	0.3785	0.3981	0.3842	0.3818	0.3801	0.3868	0.3778	0.3988
Anneal	0.0394	0.0365	0.0474	0.0399	0.0399	0.0412	0.0365	0.0517
German	0.2578	0.2602	0.2602	0.2600	0.2614	0.2572	0.2610	0.2564
Multiple-Features	0.3171	0.3127	0.3170	0.3182	0.3220	0.3129	0.3192	0.3127
Hypothyroid	0.0176	0.0171	0.0178	0.0175	0.0174	0.0175	0.0177	0.0173
Satimage	0.1759	0.1757	0.1763	0.1759	0.1759	0.1758	0.1761	0.1774
Musk	0.0754	0.0717	0.0783	0.0714	0.0689	0.0769	0.0737	0.0869
Pioneer-1-Mobile-Robot	0.0212	0.0293	0.0209	0.0195	0.0177	0.0244	0.0199	0.0255
Handwritten-Digits	0.1170	0.1178	0.1180	0.1168	0.1168	0.1167	0.1180	0.1185
Australian-Sign-Language	0.3593	0.4462	0.3595	0.3595	0.3596	0.3587	0.3596	0.4525
Letter-Recognition	0.2579	0.3263	0.2583	0.2579	0.2579	0.2581	0.2579	0.3288
Adult	0.1677	0.1664	0.1700	0.1672	0.1668	0.1701	0.1659	0.1722
Ipums.la.99	0.1641	0.1560	0.1662	0.1631	0.1619	0.1650	0.1631	0.1591
Census-Income	0.2319	0.2221	0.2332	0.2313	0.2312	0.2328	0.2309	0.2271
Forest-Covertype	0.3171	0.3167	0.3171	0.3171	0.3173	0.3168	0.3172	0.3169

Table 7: Selective NB (SNB) model's error rate comparisons for RNDD versus ENDD, NDD, PKID, EFD and EWD.

Dataset	RNDD	ENDD	NDD	PKID	EFD	EWD
Labor-Negotiations	0.0740	0.1119	0.0600	0.1121	0.1514	0.1687
Echocardiogram	0.3081	0.3541	0.3243	0.2784	0.3054	0.3486
Iris	0.0387	0.0667	0.0520	0.0640	0.0573	0.0347
Hepatitis	0.1677	0.1742	0.1742	0.1781	0.1936	0.1910
Wine-Recognition	0.0258	0.0348	0.0348	0.0326	0.0404	0.0292
Sonar	0.2779	0.3000	0.2981	0.2808	0.2817	0.2750
Glass-Identification	0.3327	0.3364	0.3299	0.3664	0.3729	0.4065
Heart-Disease-(Cleveland)	0.1837	0.1948	0.2052	0.1859	0.1748	0.1756
Liver-Disorders	0.3768	0.4329	0.4052	0.4272	0.4110	0.4290
Ionosphere	0.0997	0.1083	0.1049	0.1032	0.1043	0.0940
Horse-Colic	0.1810	0.1848	0.1935	0.1935	0.1745	0.1777
Credit-Screening-(Australia)	0.1516	0.1559	0.1606	0.1522	0.1586	0.1516
Breast-Cancer-Wisconsin	0.0289	0.0329	0.0323	0.0292	0.0283	0.0286
Pima-Indians-Diabetes	0.2555	0.2690	0.2596	0.2622	0.2534	0.2349
Vehicle	0.3778	0.4009	0.3998	0.3986	0.3962	0.3993
Anneal	0.0301	0.0528	0.0510	0.0568	0.0372	0.0423
German	0.2610	0.2638	0.2618	0.2588	0.2660	0.2684
Multiple-Features	0.3171	0.3127	0.3127	0.3160	0.3166	0.3074
Hypothyroid	0.0164	0.0184	0.0173	0.0181	0.0253	0.0358
Satimage	0.1814	0.1815	0.1854	0.1795	0.1906	0.1892
Pioneer-1-Mobile-Robot	0.0186	0.0215	0.0220	0.0212	0.0961	0.0920
Handwritten-Digits	0.1172	0.1172	0.1186	0.1201	0.1253	0.1262
Australian-Sign-Language	0.3592	0.3572	0.4447	0.3592	0.3659	0.3798
Letter-Recognition	0.2533	0.2543	0.3257	0.2538	0.2603	0.2943
Adult	0.1601	0.1657	0.1659	0.1698	0.1700	0.1780
Ipums.la.99	0.0671	0.0671	0.0671	0.0671	0.0671	0.0671
Census-Income	0.0649	0.0649	0.0649	0.0649	0.0644	0.0656
Forest-Covertype	0.3180	0.3200	0.3179	0.3197	0.3285	0.3233

Table 8: Selective NB (SNB) model's Brier score comparisons for RNDD versus ENDD, NDD, PKID, EFD and EWD.

Dataset	RNDD	ENDD	NDD	PKID	EFD	EWD
Labor-Negotiations	0.1169	0.1772	0.0879	0.1708	0.2477	0.2662
Echocardiogram	0.4232	0.5288	0.4944	0.3889	0.5040	0.5204
Iris	0.0608	0.1006	0.0811	0.1002	0.1044	0.0601
Hepatitis	0.2647	0.2835	0.2892	0.2884	0.2992	0.2911
Wine-Recognition	0.0434	0.0525	0.0569	0.0542	0.0674	0.0507
Sonar	0.4383	0.5302	0.5080	0.4917	0.4854	0.4559
Glass-Identification	0.4980	0.5499	0.5395	0.5585	0.5807	0.5756
Heart-Disease-(Cleveland)	0.2830	0.2974	0.3079	0.2872	0.2811	0.2782
Liver-Disorders	0.4838	0.5168	0.5104	0.5342	0.5086	0.5195
Ionosphere	0.1748	0.1930	0.1840	0.1772	0.1778	0.1718
Horse-Colic	0.2817	0.2920	0.2997	0.3280	0.2916	0.2972
Credit-Screening-(Australia)	0.2205	0.2251	0.2323	0.2349	0.2399	0.2293
Breast-Cancer-Wisconsin	0.0547	0.0579	0.0563	0.0549	0.0544	0.0538
Pima-Indians-Diabetes	0.3468	0.3687	0.3583	0.3579	0.3472	0.3234
Vehicle	0.6452	0.6689	0.6641	0.6727	0.6457	0.6460
Anneal	0.0501	0.0800	0.0815	0.0921	0.0534	0.0635
German	0.3641	0.3650	0.3606	0.3601	0.3572	0.3607
Multiple-Features	0.4666	0.4626	0.4626	0.4609	0.4336	0.4332
Hypothyroid	0.0265	0.0299	0.0286	0.0306	0.0403	0.0544
Satimage	0.3266	0.3267	0.3267	0.3285	0.3538	0.3545
Pioneer-1-Mobile-Robot	0.0311	0.0357	0.0369	0.0352	0.1533	0.1433
Handwritten-Digits	0.2054	0.2039	0.2051	0.2089	0.2139	0.2155
Australian-Sign-Language	0.4751	0.4738	0.5156	0.4764	0.4875	0.5025
Letter-Recognition	0.3633	0.3652	0.4339	0.3651	0.3744	0.4099
Adult	0.2302	0.2379	0.2372	0.2429	0.2441	0.2565
Ipums.la.99	0.1147	0.1148	0.1147	0.1147	0.1147	0.1147
Census-Income	0.0953	0.0953	0.0953	0.0953	0.0953	0.0955
Forest-Covertype	0.4418	0.4439	0.4389	0.4370	0.4459	0.4523