Pēteris Račinskis
pr20015

Mājas darbs 2.3: ReliefF, faktoru analīze

Uzdevums 1

a. ReliefF - parametri

Izmantojot programmatūras pakotnē Weka pieejamo ReliefF algoritma realizāciju, pieejami sekojošie parametri:

- *numNeighbours* - katram datu punktam tuvāko savai un citai klasei piederošo vektoru skaits, kas tiek lietots atribūtu rangu aprēķinā;
- *sampleSize* - algoritmā izmantotās datu kopas apakškopas izmērs. Vērtība -1 - pilnā pārlase;
- *seed* - nejaušo skaitļu ģeneratora izejas parametrs, ļauj novērtēt algoritma stabilitāti pret sākuma nosacījumiem;
- *sigma* - parametrs eksponenciālajai distances svēršanas funkcijai, ko izmanto, ja kaimiņi tiek svērti pēc distancēm. Ieteicamā vērtība - 0.1-0.2 * *numNeighbours*;
- *weightByDistance* - ieslēgt/izslēgt kaimiņu vektoru svēršanu pēc distances.

b. rezultāti ar datu kopu ionosphere, paredzot klasi = class

Mainot tikai kaimiņu skaitu (5,10,20,50), rezultāti nav īpaši stabili - atribūti, kas dominē pie maziem skaitiem zaudē nozīmi, vietā nāk citi:

Ranked attributes:	Ranked attributes:	Ranked attributes:	Ranked attributes:
0.1044 8 a08	0.1107 24 a24	0.1181 3 a03	0.1542 5 a05
0.1021 24 a24	0.1039 3 a03	0.1113 5 a05	0.124 7 a07
0.0842 27 a27	0.1022 8 a08	0.1043 8 a08	0.1201 3 a03
0.0834 5 a05	0.0922 5 a05	0.1024 24 a24	0.1128 15 a15
0.0823 34 a34	0.0846 14 a14	0.0961 7 a07	0.1079 21 a21
0.0821 3 a03	0.081 7 a07	0.0913 15 a15	0.1077 13 a13
0.0778 14 a14	0.0798 16 a16	0.0904 14 a14	0.0869 8 a08
0.0755 6 a06	0.0787 34 a34	0.0894 21 a21	0.0869 23 a23
0.0744 29 a29	0.0782 29 a29	0.088 16 a16	0.0842 9 a09
0.0735 7 a07	0.0769 9 a09	0.0837 13 a13	0.0835 19 a19
0.0725 28 a28	0.0765 12 a12	0.0813 9 a09	0.0822 17 a17
0.072 16 a16	0.0763 6 a06	0.0811 6 a06	0.0786 31 a31
0.0717 32 a32	0.0761 19 a19	0.0795 12 a12	0.0783 29 a29
0.0702 12 a12	0.0748 15 a15	0.0776 19 a19	0.0771 25 a25
0.0693 19 a19	0.0733 25 a25	0.0761 22 a22	0.0726 33 a33
0.0692 21 a21	0.0716 27 a27	0.0743 29 a29	0.0707 11 a11
0.0682 15 a15	0.0712 28 a28	0.0741 33 a33	0.0692 12 a12
0.0657 31 a31	0.0709 22 a22	0.0727 25 a25	0.068 27 a27
0.0648 22 a22	0.0709 21 a21	0.072 34 a34	0.067 14 a14
0.0641 26 a26	0.0682 13 a13	0.0698 17 a17	0.0669 6 a06
0.0638 9 a09	0.0675 33 a33	0.0676 31 a31	0.0606 4 a04
0.0631 33 a33	0.0664 32 a32	0.0652 27 a27	0.059 24 a24
0.0615 20 a20	0.0634 31 a31	0.0634 32 a32	0.0585 16 a16
0.0611 30 a30	0.0622 18 a18	0.0631 10 a10	0.0576 28 a28
0.0602 13 a13	0.0613 10 a10	0.0613 11 a11	0.0573 10 a10
0.0596 4 a04	0.0593 26 a26	0.0577 4 a04	0.0515 1 a01
0.0592 10 a10	0.0585 4 a04	0.0558 18 a18	0.0473 20 a20
0.0591 18 a18	0.0584 17 a17	0.0552 20 a20	0.0458 18 a18
0.0581 17 a17	0.0574 30 a30	0.0552 23 a23	0.0448 22 a22
0.0577 25 a25	0.0559 20 a20	0.0544 28 a28	0.0388 32 a32
0.0564 1 a01	0.0533 11 a11	0.0514 26 a26	0.0374 34 a34
0.0543 23 a23	0.0497 23 a23	0.0473 30 a30	0.0345 26 a26
0.0494 11 a11	0.035 1 a01	0.0211 1 a01	0.0288 30 a30
0 2 a02	0 2 a02	0 2 a02	0 2 a02

Līdzīgu rezultātu var panākt, lielākiem kaimiņu skaitiem izmantojot distances svēršanu - faktiski eksponenciāli dilstošā svaru funkcija panāk līdzīgus rezultātus, kā tālāko kaimiņu nogriešana (attēlā - knn=50, sigma={1,10,50,100}):

Ranked attributes:		Ranked attributes:		Ranked attributes:		Ranked attributes:	
0.1102	8 a08	0.1078	24 a24	0.1456	5 a05	0.152	5 a05
0.0996	24 a24	0.1046	8 a08	0.1208	3 a03	0.1224	7 a07
0.0908	29 a29	0.1032	3 a03	0.1177	7 a07	0.1204	3 a03
0.0875	27 a27	0.0957	5 a05	0.1105	15 a15	0.1124	15 a15
0.0808	34 a34	0.0864	14 a14	0.1044	21 a21	0.107	21 a21
0.079	28 a28	0.0829	16 a16	0.1038	13 a13	0.1067	13 a13
0.0788	3 a03	0.0824	7 a07	0.0912	8 a08	0.0881	8 a08
0.0719	5 a05	0.0791	34 a34	0.0836	9 a09	0.0855	23 a23
0.0712	19 a19	0.0784	6 a06	0.0829	19 a19	0.084	9 a09
0.0707	26 a26	0.0775	15 a15	0.0814	23 a23	0.0834	19 a19
0.0703	21 a21	0.0753	21 a21	0.0804	17 a17	0.0818	17 a17
0.069	6 a06	0.0753	9 a09	0.0783	29 a29	0.0783	29 a29
0.0684	1 a01	0.0751	29 a29	0.0772	31 a31	0.0782	31 a31
0.0675	32 a32	0.0746	12 a12	0.0767	25 a25	0.077	25 a25
0.0672	7 a07	0.0741	19 a19	0.0736	33 a33	0.0729	33 a33
0.064	25 a25	0.0725	27 a27	0.0724	12 a12	0.0702	11 a11
0.0628	9 a09	0.0724	22 a22	0.071	14 a14	0.0701	12 a12
0.0626	20 a20	0.0711	13 a13	0.0702	6 a06	0.068	14 a14
0.0623	10 a10	0.0707	25 a25	0.0689	11 a11	0.068	27 a27
0.062	14 a14	0.0686	33 a33	0.068	27 a27	0.0678	6 a06
0.0614	33 a33	0.0679	32 a32	0.0671	24 a24	0.0611	24 a24
0.0602	15 a15	0.0655	28 a28	0.0632	16 a16	0.0604	4 a04
0.059	30 a30	0.0644	31 a31	0.0601	10 a10	0.0597	16 a16
0.0584	16 a16	0.0614	10 a10	0.06	4 a04	0.058	10 a10
0.057	12 a12	0.0606	17 a17	0.0558	28 a28	0.0571	28 a28
0.0544	4 a04	0.0593	18 a18	0.0515	22 a22	0.0496	1 a01
0.0536	17 a17	0.0586	4 a04	0.0497	20 a20	0.0479	20 a20
0.053	31 a31	0.0578	26 a26	0.048	18 a18	0.0465	22 a22
0.0505	13 a13	0.0558	20 a20	0.0446	1 a01	0.0463	18 a18
0.0494	18 a18	0.0551	30 a30	0.0439	34 a34	0.0398	32 a32
0.0487	22 a22	0.0538	11 a11	0.0429	32 a32	0.0391	34 a34
0.0454	23 a23	0.0509	23 a23	0.0377	26 a26	0.0353	26 a26
0.0407	11 a11	0.0344	1 a01	0.0315	30 a30	0.0294	30 a30
0	2 a02	0	2 a02	0	2 a02	0	2 a02

Izmantojot datu kopas pilno pārlasi, paša algoritma nejaušajam sākuma parametram nav nekādas nozīmes - algoritms uzvedas determinēti. Šo parametru izmanto, veidojot nejaušu treniņa kopas apakškopu. Veicot krosvalidāciju šis dalījums notiek augstākā līmenī, un atklājas, ka rezultāts nav sevišķi stabils pret testa-treniņa kopu dalījumu, vismaz šai datu kopai. Pie knn=10, nesvērtām distancēm un pilnās treniņa kopas pārlases, pirmais rangs gandrīz vienmēr ir a24 taču pārējie nedaudz “klejo”.

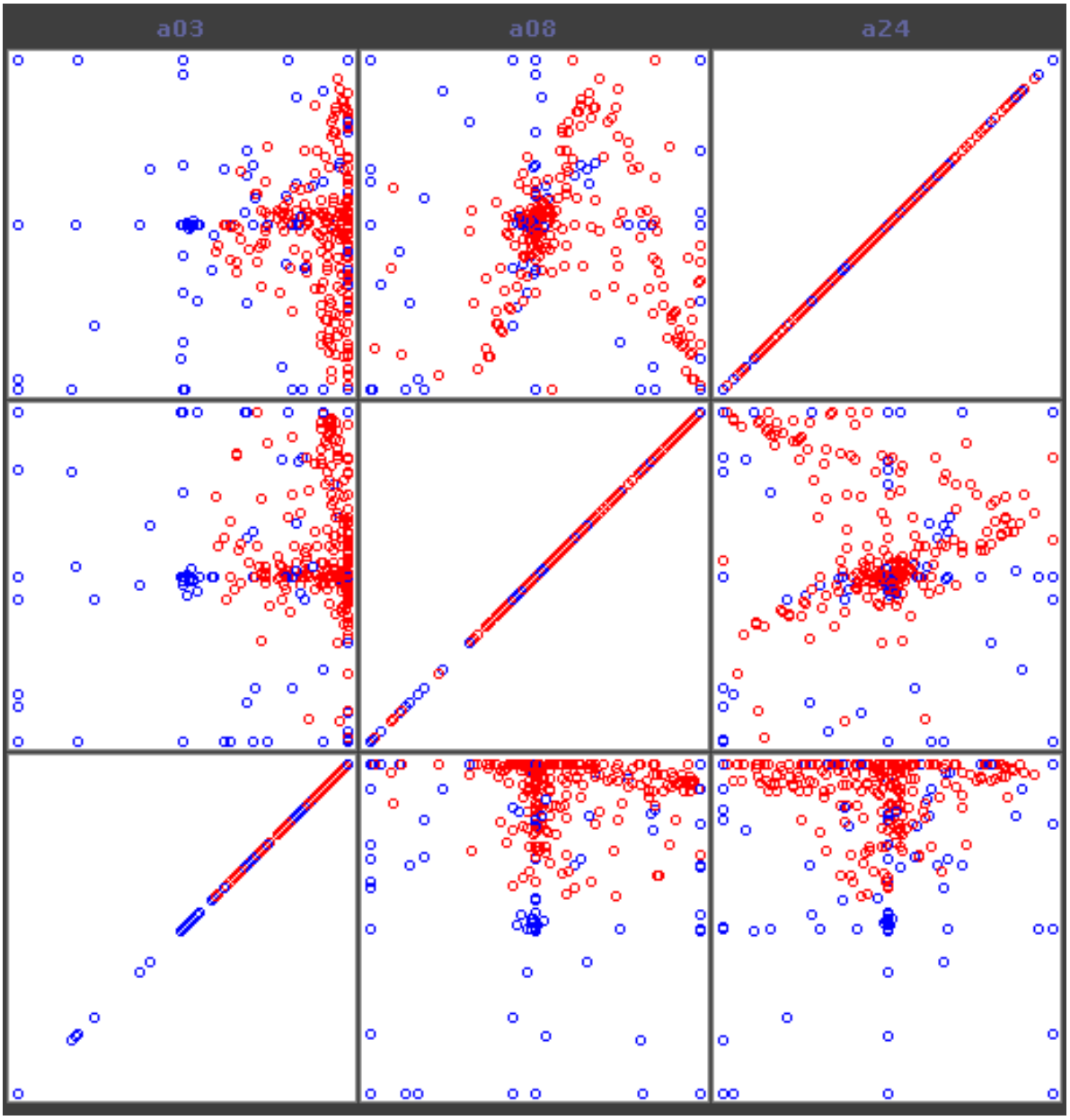
=== Attribute selection 10 fold cross-validation (stratified), seed: 10 ===

average merit	average rank	attribute
0.112 +- 0.006	1.5 +- 0.67	24 a24
0.106 +- 0.01	2.3 +- 1.27	3 a03
0.103 +- 0.007	2.9 +- 0.94	8 a08
0.095 +- 0.004	3.5 +- 0.81	5 a05
0.087 +- 0.004	6.1 +- 2.43	14 a14
0.084 +- 0.006	7.6 +- 3.93	16 a16
0.082 +- 0.006	8.1 +- 2.51	7 a07
0.079 +- 0.005	10.3 +- 4.52	34 a34
0.078 +- 0.006	11.3 +- 3.9	9 a09
0.078 +- 0.007	11.4 +- 4.1	6 a06
0.077 +- 0.004	11.9 +- 2.74	15 a15
0.076 +- 0.004	12.3 +- 4.05	12 a12
0.076 +- 0.007	12.5 +- 5.12	29 a29
0.075 +- 0.006	13.3 +- 3.95	19 a19
0.073 +- 0.006	15.2 +- 4.58	25 a25
0.073 +- 0.003	15.8 +- 3.03	21 a21
0.072 +- 0.003	16.3 +- 3.2	22 a22
0.071 +- 0.004	17 +- 2.83	13 a13
0.07 +- 0.003	18 +- 3.41	27 a27
0.068 +- 0.005	19.2 +- 3.82	28 a28
0.068 +- 0.005	19.3 +- 3.52	32 a32
0.068 +- 0.003	19.9 +- 2.26	33 a33
0.064 +- 0.003	22.9 +- 2.02	31 a31
0.061 +- 0.006	24.3 +- 3.03	10 a10
0.061 +- 0.003	24.9 +- 2.17	18 a18
0.059 +- 0.004	26.8 +- 1.99	17 a17
0.058 +- 0.003	26.9 +- 2.26	26 a26
0.057 +- 0.005	27.7 +- 3.13	4 a04
0.056 +- 0.003	28.7 +- 1.55	30 a30
0.056 +- 0.004	28.7 +- 1.9	20 a20
0.054 +- 0.003	29.8 +- 1.47	11 a11
0.049 +- 0.003	31.6 +- 0.66	23 a23
0.031 +- 0.005	33 +- 0	1 a01
0 +- 0	34 +- 0	2 a02

=== Attribute selection 10 fold cross-validation (stratified), seed: 15 ===

average merit	average rank	attribute
0.111 +- 0.006	1.5 +- 0.81	24 a24
0.103 +- 0.007	2.6 +- 1.28	8 a08
0.106 +- 0.005	2.6 +- 1.02	3 a03
0.095 +- 0.006	3.7 +- 1.1	5 a05
0.087 +- 0.005	5.4 +- 0.8	14 a14
0.084 +- 0.01	7.5 +- 5.59	16 a16
0.082 +- 0.004	8.1 +- 1.92	7 a07
0.079 +- 0.009	11.2 +- 5.88	34 a34
0.078 +- 0.005	11.2 +- 3.99	9 a09
0.077 +- 0.005	11.6 +- 4.65	6 a06
0.076 +- 0.003	12.5 +- 2.69	12 a12
0.077 +- 0.004	12.5 +- 2.91	15 a15
0.076 +- 0.006	12.9 +- 4.48	29 a29
0.076 +- 0.005	13.1 +- 4.44	19 a19
0.074 +- 0.005	15.1 +- 3.36	21 a21
0.073 +- 0.005	15.1 +- 4.06	25 a25
0.072 +- 0.005	15.8 +- 3.19	22 a22
0.071 +- 0.005	17.3 +- 3.72	13 a13
0.07 +- 0.002	17.7 +- 1.85	27 a27
0.068 +- 0.001	18.9 +- 1.76	28 a28
0.068 +- 0.003	19.3 +- 2.53	32 a32
0.069 +- 0.005	19.9 +- 3.99	33 a33
0.063 +- 0.002	22.8 +- 1.54	31 a31
0.061 +- 0.007	24.3 +- 2.87	10 a10
0.061 +- 0.004	24.6 +- 2.33	18 a18
0.058 +- 0.003	27 +- 1.79	17 a17
0.058 +- 0.003	27 +- 2.37	26 a26
0.057 +- 0.006	27.4 +- 2.2	4 a04
0.056 +- 0.003	28.4 +- 1.2	30 a30
0.055 +- 0.004	29.3 +- 1.9	20 a20
0.054 +- 0.003	30.2 +- 1.54	11 a11
0.049 +- 0.003	31.5 +- 0.81	23 a23
0.031 +- 0.004	33 +- 0	1 a01
0 +- 0	34 +- 0	2 a02

Pēc visām parametru permutācijām ir skaidrs, ka starp svarīgākajiem parametriem varētu būt a24, a03 un a08.



Redzams, ka projekcijās a03 X a08 un a03 X a24 zilā klase veido nelielu klasteri grafika vidū, bet sarkanā sagrupējusies pret vienu no malām. a08 X a24, nekādu klasifikācijai noderīgu sakarību nav.

c. rezultāti ar datu kopu unbalanced

Neatkarīgi no kaimiņu skaita (knn={10,50,100}), šķiet, dominē 31. parametrs “BBB”, taču pārējie ievērojami “klejo”:

```
Ranked attributes:
 0.3617   31 BBB
 0.2061   19 WBN_LP_L_0.50
 0.1989   21 WBN_LP_L_0.75
 0.197    12 WBN_EN_H_0.50
 0.189     7 WBN_GC_L_1.00
 0.1882   23 WBN_LP_L_1.00
 0.1876   17 WBN_LP_L_0.25
 0.1835    1 WBN_GC_L_0.25
 0.1812   14 WBN_EN_H_0.75
 0.1779   10 WBN_EN_H_0.25
 0.1682    3 WBN_GC_L_0.50
 0.1514   16 WBN_EN_H_1.00
 0.1414    5 WBN_GC_L_0.75
 0.1341    2 WBN_GC_H_0.25
 0.111    18 WBN_LP_H_0.25
 0.1092   32 BadGroup
 0.109    15 WBN_EN_L_1.00
 0.1086    8 WBN_GC_H_1.00
 0.105    13 WBN_EN_L_0.75
 0.1004    4 WBN_GC_H_0.50
 0.0949   11 WBN_EN_L_0.50
 0.0938   24 WBN_LP_H_1.00
 0.093     6 WBN_GC_H_0.75
 0.0914   26 PSA
 0.0901   22 WBN_LP_H_0.75
 0.0827   20 WBN_LP_H_0.50
 0.0708   29 NumHBD
 0.0699    9 WBN_EN_L_0.25
 0.0683   28 NumHBA
 0.0613   30 MW
 0.0569   25 XLogP
 0.0428   27 NumRot
```

```
Ranked attributes:
 0.3592   31 BBB
 0.2032   12 WBN_EN_H_0.50
 0.1987    7 WBN_GC_L_1.00
 0.1967   21 WBN_LP_L_0.75
 0.193    19 WBN_LP_L_0.50
 0.1906   23 WBN_LP_L_1.00
 0.1895   10 WBN_EN_H_0.25
 0.1839   14 WBN_EN_H_0.75
 0.1572   16 WBN_EN_H_1.00
 0.1561   17 WBN_LP_L_0.25
 0.1391    1 WBN_GC_L_0.25
 0.133     3 WBN_GC_L_0.50
 0.132     5 WBN_GC_L_0.75
 0.114     2 WBN_GC_H_0.25
 0.1133    8 WBN_GC_H_1.00
 0.1057   15 WBN_EN_L_1.00
 0.0984   13 WBN_EN_L_0.75
 0.093     6 WBN_GC_H_0.75
 0.0915    4 WBN_GC_H_0.50
 0.0885   32 BadGroup
 0.0878   24 WBN_LP_H_1.00
 0.0838   11 WBN_EN_L_0.50
 0.0806   22 WBN_LP_H_0.75
 0.0796   18 WBN_LP_H_0.25
 0.0758   26 PSA
 0.0659   20 WBN_LP_H_0.50
 0.0581    9 WBN_EN_L_0.25
 0.0485   28 NumHBA
 0.0483   29 NumHBD
 0.0404   25 XLogP
 0.0383   30 MW
 0.0284   27 NumRot
```

```
Ranked attributes:
 0.3211   31 BBB
 0.187     7 WBN_GC_L_1.00
 0.1813   21 WBN_LP_L_0.75
 0.1774   23 WBN_LP_L_1.00
 0.1746   19 WBN_LP_L_0.50
 0.1742   12 WBN_EN_H_0.50
 0.1647   14 WBN_EN_H_0.75
 0.1536   10 WBN_EN_H_0.25
 0.1423   16 WBN_EN_H_1.00
 0.1372   17 WBN_LP_L_0.25
 0.1211    5 WBN_GC_L_0.75
 0.1121    1 WBN_GC_L_0.25
 0.1105    3 WBN_GC_L_0.50
 0.0993    8 WBN_GC_H_1.00
 0.096    15 WBN_EN_L_1.00
 0.0942    2 WBN_GC_H_0.25
 0.089    13 WBN_EN_L_0.75
 0.0798    6 WBN_GC_H_0.75
 0.0757    4 WBN_GC_H_0.50
 0.0753   11 WBN_EN_L_0.50
 0.0747   24 WBN_LP_H_1.00
 0.0671   22 WBN_LP_H_0.75
 0.0665   32 BadGroup
 0.063    26 PSA
 0.0577   18 WBN_LP_H_0.25
 0.0514   20 WBN_LP_H_0.50
 0.0512    9 WBN_EN_L_0.25
 0.0369   29 NumHBD
 0.0366   28 NumHBA
 0.0321   25 XLogP
 0.0249   30 MW
 0.0179   27 NumRot
```

Veicot krosvalidāciju pie knn=10, augstāk novērtētie diezgan stabili šķiet 31,7,21:

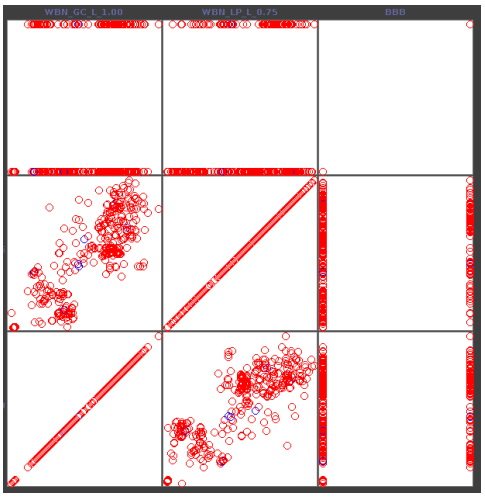
```
=== Attribute selection 10 fold cross-validation (stratified), seed: 1 ===

average merit      average rank  attribute
 0.313 +- 0.019      1   +- 0       31 BBB
 0.185 +- 0.004      2.4 +- 0.66     7 WBN_GC_L_1.00
 0.179 +- 0.007      2.9 +- 0.7     21 WBN_LP_L_0.75
 0.176 +- 0.006      4.2 +- 0.6     23 WBN_LP_L_1.00
 0.17  +- 0.008      5.1 +- 1.37    12 WBN_EN_H_0.50
 0.172 +- 0.006      5.6 +- 0.49    19 WBN_LP_L_0.50
 0.162 +- 0.003      6.8 +- 0.4     14 WBN_EN_H_0.75
 0.148 +- 0.016      8.5 +- 1.02    10 WBN_EN_H_0.25
 0.141 +- 0.002      9   +- 0.45    16 WBN_EN_H_1.00
 0.135 +- 0.005      9.7 +- 0.78    17 WBN_LP_L_0.25
 0.12  +- 0.01      11   +- 0.77     5 WBN_GC_L_0.75
 0.109 +- 0.01      13.2 +- 1.47     1 WBN_GC_L_0.25
 0.108 +- 0.011     13.3 +- 1.49     3 WBN_GC_L_0.50
 0.096 +- 0.006     13.9 +- 1.58     8 WBN_GC_H_1.00
 0.096 +- 0.009     14.5 +- 1.63    15 WBN_EN_L_1.00
 0.091 +- 0.007     15.8 +- 0.87     2 WBN_GC_H_0.25
 0.088 +- 0.008     16.6 +- 1.43    13 WBN_EN_L_0.75
 0.077 +- 0.006     18.4 +- 1.43     6 WBN_GC_H_0.75
 0.074 +- 0.006     19.7 +- 1.42    11 WBN_EN_L_0.50
 0.073 +- 0.003     19.8 +- 0.87    24 WBN_LP_H_1.00
 0.073 +- 0.006     20.4 +- 1.85     4 WBN_GC_H_0.50
 0.065 +- 0.002     22.5 +- 0.67    22 WBN_LP_H_0.75
 0.062 +- 0.01      23   +- 2.28    32 BadGroup
 0.062 +- 0.005     23.3 +- 1.27    26 PSA
 0.055 +- 0.009     25.3 +- 1.62    18 WBN_LP_H_0.25
 0.051 +- 0.004     26.2 +- 0.87     9 WBN_EN_L_0.25
 0.049 +- 0.002     26.5 +- 0.5     20 WBN_LP_H_0.50
 0.036 +- 0.007     28.5 +- 1.28    29 NumHBD
 0.036 +- 0.006     28.7 +- 0.9     28 NumHBA
 0.031 +- 0.003     29.5 +- 0.81    25 XLogP
 0.024 +- 0.004     30.8 +- 0.75    30 MW
 0.017 +- 0.004     31.9 +- 0.3     27 NumRot
```

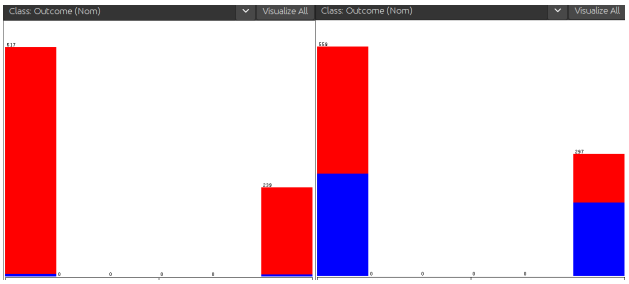
```
=== Attribute selection 10 fold cross-validation (stratified), seed: 3262 ===

average merit      average rank  attribute
 0.313 +- 0.027      1   +- 0       31 BBB
 0.185 +- 0.005      2.6 +- 0.8      7 WBN_GC_L_1.00
 0.18  +- 0.006      2.7 +- 0.64    21 WBN_LP_L_0.75
 0.177 +- 0.006      4.1 +- 0.54    23 WBN_LP_L_1.00
 0.172 +- 0.006      5.3 +- 1       19 WBN_LP_L_0.50
 0.169 +- 0.009      5.6 +- 1.11    12 WBN_EN_H_0.50
 0.162 +- 0.004      6.8 +- 0.6     14 WBN_EN_H_0.75
 0.147 +- 0.015      8.5 +- 1.12    10 WBN_EN_H_0.25
 0.141 +- 0.006      8.9 +- 0.3     16 WBN_EN_H_1.00
 0.135 +- 0.006      9.6 +- 0.8     17 WBN_LP_L_0.25
 0.119 +- 0.008     11   +- 0.45     5 WBN_GC_L_0.75
 0.109 +- 0.009     12.8 +- 1.47     1 WBN_GC_L_0.25
 0.108 +- 0.01      13.4 +- 1.28     3 WBN_GC_L_0.50
 0.096 +- 0.007     14.2 +- 1.47     8 WBN_GC_H_1.00
 0.096 +- 0.004     14.5 +- 1.28    15 WBN_EN_L_1.00
 0.091 +- 0.007     15.7 +- 0.64     2 WBN_GC_H_0.25
 0.088 +- 0.005     16.3 +- 1.1     13 WBN_EN_L_0.75
 0.077 +- 0.006     18.9 +- 1.3      6 WBN_GC_H_0.75
 0.073 +- 0.003     19.7 +- 1.1     24 WBN_LP_H_1.00
 0.074 +- 0.005     19.8 +- 1.4     11 WBN_EN_L_0.50
 0.073 +- 0.007     20.3 +- 1.55     4 WBN_GC_H_0.50
 0.065 +- 0.002     22.3 +- 0.9     22 WBN_LP_H_0.75
 0.062 +- 0.008     23.4 +- 1.69    32 BadGroup
 0.062 +- 0.005     23.4 +- 1.56    26 PSA
 0.055 +- 0.009     25.1 +- 1.51    18 WBN_LP_H_0.25
 0.051 +- 0.003     26.1 +- 0.94     9 WBN_EN_L_0.25
 0.049 +- 0.002     26.4 +- 0.66    20 WBN_LP_H_0.50
 0.036 +- 0.008     28.6 +- 1.5     29 NumHBD
 0.036 +- 0.007     28.7 +- 0.9     28 NumHBA
 0.031 +- 0.005     29.5 +- 0.67    25 XLogP
 0.023 +- 0.004     30.9 +- 0.3     30 MW
 0.017 +- 0.005     31.9 +- 0.3     27 NumRot
```

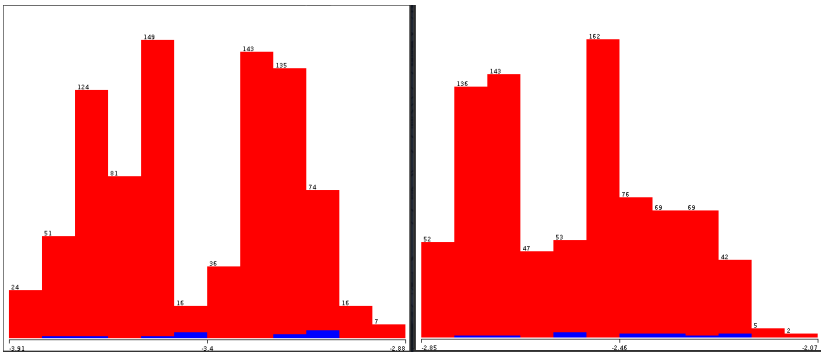
Veicot projekcijas un apskatot histogrammas, var secināt sekojošo:



- Projekcijās vizuāli redzēt nekādas izteiktas sakarības nevar;
- "BBB" gadījumā var redzēt, ka sarkano instanču skaits ar vērtību 0 ir ievērojami lielāks, nekā sarkano instanču skaits ar vērtību 1. Zilo punktu ir ļoti maz, tāpēc grūti spriest par to sadalījumu, taču sverot punktu pēc klašu izmēra redzams, ka vismaz šajā kopā tie ir samērā vienādi sadalīti. Liela nozīme atribūtam varētu būt skaidrojama ar nesamērīgajiem klašu izmēriem:



- Abiem pārējiem parametriem arī ir reģioni, kur sarkano punktu ir daudz vairāk nekā zilo (vai zilo nav vispār), bet zilie šķiet sagrupēti "klasteros" (kas, ļoti iespējams, ir iluzori - datu kopa ir niecīga):



Attiecīgi šo atribūtu kārtojumu ir grūti utzvert kā sevišķi ticamu.

Uzdevums 2

a. Faktoru analīze praksē

Atrast kādas statistiskas metodes "veiksmes stāstus" vienkārši veicot meklēšanu tīmeklī ir ievērojami grūtāk, nekā atrast pamācības, kā šo metodi izmantot kādā no populārākajām aprēķinu veikšanas izpildvidēm - R, Python, mathlab, u.t.t. - jo meklēšanas rīki mūsdienās tiek agresīvi un sistemātiski ekspluatēti mārketinga nolūkos, un mācīties gribētāju maciņu tukšošana ar apšaubāmas kvalitātes tiešsaistes kursiem ir daudzus miljardus vērta industrija. Pat meklējot publikācijas ar atslēgvārdiem "factor analysis" rīkā scholar.google.com, visa pirmā rezultātu lapa sastāv no grāmatām par šo tematiku un rakstiem, kas ievada līmenī apraksta tematu. Sapratis, ka tik viegli atrast reālus pētījumus, kur šī metode tiek pielietota, nebūs, mājas darba autors ir izlēmis aprakstīt kādu saprotamāku piemēru, kur faktoru analīze veikta ar datu kopu.

"A Beginner's Guide to Factor Analysis: Focusing on Exploratory Factor Analysis" ir raksts, kā lielākā daļa veltīta faktoru analīzes metodes aprakstam vispārīgi (aptuveni tāpat, kā lekcijās), bet papildus doti arī praktiski padomi metodes izmantošanai, un beigās dots paraugs, kur metode IBM SPSS platformā pielietota reālu datu skaidrošanai ar faktoriem.

Datu kopa veidota veicot aptauju Kanādā, kur dalībniekiem prasīts novērtēt dažādu ar pārtiku saistītu risku nozīmīgumu:

In the second section, I am going to read you a list of items related to food safety. I would like to get your opinion about the potential risk it can represent for the Canadian public. Please respond to the following questions using the same 5-point scale used previously. **The question is what level of risk to Canadians would you say there is related to the following:**

	Not at all	A little	Moderately	Very Much	Extremely	Don't Know/ No Opinion	No Response
Bacteria in food (e.g., E. coli, Salmonella)							
Pesticides							
Imported food							
Tap water							
Food irradiation (to preserve food)							
Use of antibiotics in livestock							
Mad cow disease							
Disease in wild game							
Foot and Mouth disease							
Food additives (def: chemicals used to preserve or color food or improve its taste)							
Bottled water							
Genetically modified foods							
Improper food labeling							
Mercury in fish							
Growth hormones							
Artificial sweeteners (aspartame, saccharin)							
Food packaging materials (food wrapped in plastics)							
Agroterrorism (def: deliberate introduction of harmful agents into the food chain)							

Mērķis ir skaidrot riska novērtējumu kopu ar dažiem faktoriem, kas raksturo katra dalībnieka raizes par kādu potenciālu risku grupu. Lai gan rakstīts ir diezgan daudz par lietām, kas galā netiek izmantotas, vai ir specifiskas izmantotajai programmatūras pakotnei, procesu var kopumā aprakstīt sekojoši:

- Datu kopai tiek aprēķināta faktoru matrica ar daudziem faktoriem;
- Izmantojot elkoņa likumu faktoru īpašvērtībām, tiek izlemts rezultātā atstāt tikai 3 faktorus;
- Izmantojot Kaisera "Varimax" metodi tiek veikta faktoru rotācija;
- Aplūkojot iegūto rezultātu tabulu, tiek izvirzīti potenciāli nosaukumi - semantiskas nozīmes - iegūtajiem faktoriem.

Rotated Factor Matrix[a]

	Factor		
	1	2	3
Growth_hormones	.802		
GMO	.614		.380
Antibiotics_food	.598		.344
Mercury_fish	.519	.399	
Pesticides	.514	.344	.359
Food_additives	.505		.497
Improper_label	.490	.385	
Foot_mouth		.833	
Mad_cow		.730	
Wild_game		.707	
Agroterrorism		.471	
Bacteria		.419	
Food_packaging	.341		.578
Food_irradiation	.349		.547
Bottled_water			.534
Artificial_sweet	.406		.525
Tap_water			.486
Imported_food			.377

Extraction Method: Principal Axis Factoring.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 5 iterations.

Pēc katra faktora izskaidrotajiem lielumiem izlemts faktorus nosaukt attiecīgi par:

1. Rūpnieciskas apstrādes riskiem;
2. Dzīvnieku izcelsmes pārtikas riskiem;
3. Iepakošanas procesu riskiem.

Subjektīvi novērtējot iegūto dalījumu kategorijās, pirmie divi faktori tik tiešām šķiet ticami (ja baktērija un agroterorists ir uztverami par sava dzīvniekiem), taču trešais jau sāk šķist visnotaļ šaubīgs. Protams, jāatceras, ka tiek vērtēts nevis objektīvs risku dalījums kategorijās, bet gan to klasifikācija Kanādas iedzīvotāju prātos.