Analysis Method: extreme

Brian T. Luke (lukeb@ncifcrf.gov)

The Komogorov-Smirnov test may identify a feature as important if in the middle of overlapping ranges one category has a high density of points while the other has a very low density. This may be due to a sampling problem and what is really desired are features where one category has a large number of intensities that are either above or below the intensity range of all other samples in the other categories.

The *extreme* algorithm simply orders all intensities from lowest to highest and starting from both extremes finds the maximum number of samples from a single category before a sample from another category is observed. Therefore, in contrast to t he Komogorov-Smirnov test, this procedure works with more than two categories in the dataset.

The results examining 10,000 features representing either Feature-a or Feature-b, and comparing their scores against the maximum possible score obtained from features with no information is shown in the following table.

Each	Thresh	10a	10b	15a	15b	20a	20b	25a	25b	30a	30b	35a	35b	40a	40b
30	13	1	4	12	9	60	7	272	22	761	61	1854	142	3640	278
45	14	5	4	99	9	548	43	2029	113	4557	379	7159	839	8929	1815
60	18	6	0	42	2	463	7	2139	43	5136	182	8035	538	9429	1375
90	17	147	3	1866	56	6168	400	9192	1714	9917	4186	9996	7021	10000	8966
150	14	6239	888	9819	5049	9999	9023	10000	9929	10000	9998	10000	10000	10000	10000
300	15	9992	8210	10000	9991	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000

As stated earlier, the first column represents the number of Cases and the number of Controls in each dataset. The second column represents the maximum count for samples from a single category obtained from 10,000 features where the intensities for both Cases and Controls are randomly assigned within the range of 0.0 to 100.0, and this number is reasonably insensitive to the number of Cases and Controls. The remaining columns show the number of times in 10,000 randomly generated feature intensities that a feature has a sample count that is above this threshold. The headings for these column show whether the features represent Feature-a or Feature-b, described previously, and the value of Za or 2Zb. For example, the column labeled 10a is for features that represent Feature-a with Za=10, while the column labeled 10b is for features that represent Feature-b with 2Zb=10 (Zb=5).

This procedure recognizes putative biomarkers represented by Feature-a better than those for Feature-b if the dataset contains a relatively small number of samples. For datasets with 300 cases and 300 controls, approximately 99.9% of the features with Za=10 produced a higher count than any observed feature with no information, while if 2Zb=10, 82.19% of the features had a higher count. If there are only 30 Cases and 30 Controls and the features have the form of Feature-a, there is at least a 50% chance of having a sample count from a single category greater than 13 if Za=45, meaning that the range of

intensities for one State is 55% that of the other. As the number of Cases and Controls increases from 45 to 150, the range of the smaller intensity State increases from 65% to 90% of the range of the larger intensity State. If the features have the form of Feature-b, the region of overlap increases from 62.5% to 92.5% as the number of Cases and Controls increases from 30 to 150.

(Last updated 4/29/07)