Analysis Method: catboot (formerly known as *fqual* [Hab-05])

Brian T. Luke (lukeb@ncifcrf.gov)

For each feature, this procedure uses a Bootstrap technique to determine the probability that a particular sample belongs to the correct state. For each Bootstrap run, M samples are removed from the group of N samples and the remaining N-M samples are used to determine the average intensity for each state, denoted I_s. The distance from the mth removed sample to the centroid of state s is $d_{m,s}$.

$$d_{m,s} = |I_m - I_s|$$

The un-normalized probability that this sample belongs to state s is give by the following.

$$P_s = \alpha / d_{m,s}$$

The scaling factor α is controlled by a user supplied fractional value f, which represents the fraction of the total range for which the probability drops to 0.5.

$$P_s = 0.5 \text{ when } d_{m,s} = f(I_{max} - I_{min})$$

The scaling factor is then

$$\alpha = f(I_{max} - I_{min}) / 2.0$$

Associated with each probability that it belongs to a given state is a probability that the classification is "undetermined", P_s^u.

$$\begin{aligned} P_s^u &= U && \text{if } P_s < (1-2U) \\ &= (1-P_s)/2 && \text{if } (1\text{-}2U) \le P_s \le 1.0 \\ &= 0 && \text{if } P_s > 1.0 \end{aligned}$$

If $U = 0.1$, for example, then P_s^u equals 0.1 if P_s is less than 0.8; it linearly drops from 0.1 to 0.0 as P_s increases from 0.8 to 1.0; and is 0.0 whenever P_s is greater than 1.0. This term is included so that if one of the removed samples is an outlier (has an intensity that is significantly far from the centroids) it will receive a classification of "undetermined".

The probability that it belongs to the correct state, P_{Sm}, is therefore

$$P_{Sm} = P_{Sm} \Big/ \sum_{s=1}^{S} (P_s + P_s^u)$$

The quality of feature l for this Bootstrap run b is therefore

$$Q_{l,b} = \sum_{m=1}^{M} P_{Sm}$$

If this is run *Nb* times, the overall quality of this feature is given by the following.

$$Q_l = \sum_{b=1}^{Nb} Q_{l,b} \Big/ (Nb \times M)$$

The features are then ranked from highest to lowest values of Q_l.

For each examination of 10,000 features, $M = 3$, so three samples are removed from each Bootstrap run, and the number of Bootstrap runs, *Nb*, is 1000. In addition, $f = 0.1$ meaning α is adjusted so that the probability of belonging to a particular state is 0.5 if the distance to its centroid is one-tenth of the total range of intensities. Finally, U = 0.1 which represents the maximum un-normalized probability that the membership in a particular state is undetermined. The results of these examinations are shown in the following table.

Each	Thresh	10a	10b	15a	15b	20a	20b	25a	25b	30a	30b	35a	35b	40a	40b
30	0.4689	3	3	4	5	18	4	49	19	95	33	202	61	531	94
45	0.4352	13	8	11	21	50	39	149	93	345	152	845	250	1799	453
60	0.4323	0	1	11	7	25	17	91	44	312	101	780	189	1789	349
90	0.4138	5	4	15	12	75	26	269	124	715	237	1898	548	3948	970
150	0.3958	6	11	71	36	256	134	963	409	2567	976	5015	1991	7809	3271
300	0.3760	115	81	720	461	2461	1509	5545	3320	8340	5716	9683	7717	9972	8919

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 value of Q_l 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. The remaining columns show the number of times in 10,000 randomly generated feature intensities that a feature has a value of Q_l 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).

For putative markers described by Feature-a, there is at least a 50% chance that a feature will be identified (Q_l greater than the maximum value expected from a feature with no information) if Za=25 and there are 300 Cases and 300 Controls in the dataset. This means that one of the States must have a range that is only 75% of the other for it to have more than a 50% chance of scoring higher than if they both had the same ranges. If the

number of Cases and Controls is only 150, a feature has to have Za=35 or larger to have at least a 50% chance of being identified, or one State can only have a range of intensities that is 65% of the other. For datasets with 90, 60, 45 and 30 cases and controls, the values of Za that produce at least 50% of the features with scores above the highest obtained from features with no information are 45, 50, 50, and 60, respectively. This means that the range of intensities for one State can only be 65, 50, 50, and 40% of the range of the other, respectively, to have a 50% or better chance of being identified.

Putative markers corresponding to Feature-b are much harder to identify. With 300 Cases and 300 Controls 2Zb must be about 30 or higher for a 50% chance of identification, while for 150 Cases and 150 Controls, having 2Zb=40 only yields about a 33% chance of identification. Since these features have equally sized ranges, and they are only shifted by an amount Zb relative to each other, (100-Zb) is the fraction of the ranges for each State that overlap. Therefore, with 300 Cases and 300 Controls, 85% of the range spanned by each State can overlap and there is at least a 50% chance that its z-score will be higher than any of the 10,000 examined features that have 100% overlapping ranges. For datasets with 30, 45, 60, 90, and 150 Cases and Controls, the values of 2Zb that produce at least 50% of the features with scores above comparable features with 100% overlap are 95, 80, 75, 60, and 50, respectively. This means that the percentage of overlap for the ranges of intensities in the two States can be at most 52.5, 60.0, 62.5, 70.0, and 75.0%, respectively.

(Last updated 4/4/07)