Analysis Method: dtgini (formerly known as *gini* [Hab-05])

Brian T. Luke (lukeb@ncifcrf.gov)

This procedure uses each feature to perform a one-node split of all *N* samples in the parent node into *D* daughter nodes using *D*-1 cut points, as in a decision tree. The quality of the split is determined by the value of $GINI_{split}$. In the d^{th} daughter node, the probability of being in State *s*, $P_{s,d}$, is just the number of samples from this state in this node divided by the total number of samples in this node. The Gini value for this node is

$$GINI(d) = 1.0 - \sum_{s=1}^{S} P_{s,d}^2$$

If each daughter node contains N_d samples, *GINIsplit* for this feature is then

$$GINI_{split}(l) = \sum_{d=1}^{D} \left(\frac{N_d}{N} \right) GINI(d)$$

The features are then ordered from lowest to highest $GINI_{split}(l)$.

For this procedure, *D*=*S*, so there is one daughter node for each State. This means that if three States are present, two cut points will be used to produce three daughter nodes. The feature intensities are ordered in ascending (or descending) order and the possible cut points are the midpoints between subsequent intensities. $GINI_{split}$ is determined for each cut point (or combinations of cut points) and the cut point(s) with the lowest $GINI_{split}$ are used.

The results examining 10,000 features representing either Feature-a or Feature-b, and comparing their scores against the minimum 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	0.333	1	1	8	1	28	12	76	12	275	42	728	65	1798	96
45	0.39	12	5	38	9	226	30	886	66	2488	166	5053	357	7484	738
60	0.41	8	2	81	6	557	42	2327	111	5308	301	8080	727	9517	1601
90	0.431	13	0	297	10	2610	48	6636	244	9253	925	9915	2434	9996	4963
150	0.462	952	34	6778	318	9744	2156	9995	6031	10000	9070	10000	9872	10000	7294
300	0.483	9892	3746	10000	9634	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 minimum value of $GINI_{split}$ 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 $GINI_{split}$ that is below 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 again identifies weak features more easily if they resemble Feature-a than Feature-b. If a Feature-a type of feature has Za=10, meaning that the range of one State spans 90% of the range of the other, this feature has a 98.9% chance of having a lower $GINI_{split}$ value than any observed feature with Za=0 (i.e. no information). Conversely, if the feature resembles Feature-b with 2Zb=10, meaning that there is a 95% overlap in the ranges, there is only about a 37.5% chance that it will have a lower $GINI_{split}$ value that a non-informative feature. As the number of subjects in the dataset decreases, it becomes harder to distinguish a feature with different ranges from one without. For example, if there are only 30 Cases and 30 Controls, Za must be at least 50 for a Feature-a type of feature (meaning one State has an intensity range that is only 50% of the other) of 2Zb must be at least 85 for a Feature-b type of feature (meaning that the ranges only have a 57.5% overlap) before the feature has at least a 50% chance of having a *GINIsplit* value that is less than one observed for a non-informative feature.

(Last updated 4/4/07)