Return to Contents PDF Version

Analysis Method: dtinfg (formerly known as *infg* [Hab-05])

Brian T. Luke (lukeb@ncifcrf.gov)

This procedure is similar to *dtgini*, in that each feature is used in a single-node decision tree to divide all samples among two or more daughter nodes. The only difference is that the Information Gain is used as the decision metric to define the optimum cut points, not $GINI_{split}$. Given a total of S states and N_s of the total N samples are in State s, the probability of being in this state is simply

$$P_s = N_s / N$$

The Information Entropy of the parent node containing all samples is then

$$IE(p) = -\sum_{s=1}^{S} P_s \ln(P_s)$$

For S states, (S-1) cut points in the intensity range are selected to produce S daughter nodes. If daughter node d contains N_d samples and $N_{s,d}$ samples from State s, the Information Entropy of this node is

$$IE(d) = -\sum_{s=1}^{S} P_{s,d} \ln(P_{s,d})$$

$$P_{s,d} = N_{s,d} / N_d$$

The overall Information Gain for feature l is the Information Entropy of the parent node minus the Information Entropy of all daughter nodes.

$$IG(l) = IE(p) - \sum_{d=1}^{s} IE(d)$$

The features are then ranked from highest to lowest Information Gain.

For each feature, the intensities are ranked from lowest to highest and the possible cut point(s) become the average intensity of non-identical sequential intensities.

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	10.25	4	6	45	15	151	18	529	56	1500	139	3170	267	5145	531
45	12.23	8	1	43	5	337	23	1317	43	3436	155	6104	371	8226	899
60	12.74	13	2	219	3	1322	49	4051	138	7190	483	9174	1317	9828	2820
90	14.52	33	0	786	15	4166	71	8064	477	9719	1735	9977	4199	9999	7016
150	14.92	1268	11	7358	322	9844	2494	9999	6621	10000	9303	10000	9946	10000	10000
300	12.1	9970	5733	10000	9896	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 Information Gain 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 an Information Gain 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 again identifies features represented by Feature-a more easily than those represented by Feature-b, and the value of Za or 2Zb needed for at least 50% of the features to have an Information Gain greater than a feature with no information decreases as the number of Cases and Controls increases. If there are 300 Cases and 300 Controls, a feature of type Feature-a with Za=10 (meaning that the range of one State is 90% that of the other) has about a 99.7% chance of having an Information Gain higher than a feature with no information. If the feature is of type Feature-b, with a 95% overlap of the ranges, there is about a 57% chance of it having a larger Information Gain. If there are only 30 Cases and 30 Controls, a Feature-a type of feature would have to have a range of intensities for one State being only 60% or less of the other before the probability of it having a sufficiently large Information Gain exceeds 50%. Conversely, a Feature-b type of feature would have to have an overlap of the range of 65% or less $(2Zb \geq 70)$ before the probability of a sufficiently large Information Gain exceeds 50%.

(Last updated 4/4/07)