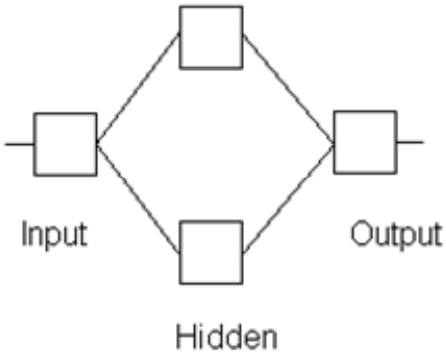
Analysis Method: nnfeat

Brian T. Luke (lukeb@ncifcrf.gov)

This procedure measures the quality of each feature using a (1,2,1) Feed-Forward Back-Propagation Artificial Neural Network as shown in the figure to the right. In contains one input node (the intensities of this feature) and two processing nodes in the hidden layer, with no input or processing bias. The single output node determines if it is a Case (1) or a Control (0), and the RMS error determines the quality of the prediction. Therefore, this procedure is only valid for a two-State analysis.



Since the optimum value of the four weights into and out of the hidden layer are optimized using a steepest descent procedure, the final results are dependent upon the order of the samples being analyzed, the number of times each sample is examined, the initial value of the weights, and the learning rate (scaling of the weight change). In the current implementation the initial weights have an initial absolute value of 0.05 or less, the learning rate is set to 0.5, and each sample is processed five times. The order of the samples is randomized and all are sequentially examined. The order is re-randomized and re-examined a total of 80 times to (hopefully) find the optimum weights. The entire process is run three times with different initial weights, which means that this is by far the slowest procedure. The smallest RMS error from these runs determines the quality of this feature.

The results examining 10,000 features representing either Feature-a or Feature-b, and comparing their scores against the minimum possible RMS error 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	0.4424	8	8	15	17	41	29	98	80	240	152	584	304	1220	556
45	0.4522	2	2	10	8	31	25	112	95	333	241	907	425	2075	946
60	0.4582	1	1	6	9	23	31	124	98	449	268	1240	683	3128	1367
90	0.4748	7	7	32	47	220	247	857	752	2454	1799	5233	3481	7937	5547
150	0.4812	2	2	47	65	366	437	1836	1743	5016	4027	8246	6796	9635	8591
300	0.4933	219	310	1803	1773	6080	4913	9050	7934	9678	9112	9802	9497	9809	9634

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 RMS error 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 RMS error 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).

For larger datasets, there is little difference in the number of features with sufficiently low RMS errors between Feature-a and Feature-b types. The same is true for the smaller datasets for small values of Za and 2Zb. As Za and 2Zb increases for the smaller datasets, more features with sufficiently small RMS errors are found for type Feature-a than Feature-b. For example in the smallest dataset with 30 Cases and 30 Controls, at least 50% of the features have an RMS error that is less than the threshold (0.4424) if Za=55 for type Feature-a. This means that the intensity range of one State is only 45% as large as the other and 27.5% of all samples have intensity in a range that is not covered by samples in the other State. For type Feature-b, the shift in intensity ranges must be at least 35% of their range (2Zb=70) so that 35% of all samples have intensity in a range that is not covered by samples in the other State.

(Last updated 4/4/07)