Return to Contents PDF Version

Description of the Tables

Brian T. Luke (lukeb@ncifcrf.gov)

This is a description of the classification results for one of the artificial datasets with non-informative peak intensities. These text boxes contain the descriptions.

Random_Intensity_60_1a: 60 Cases, 60 Controls, 300 Peaks

The first line lists the name of the analysis and corresponds to the datasets named case_60_1a.txt and control_60_1a.txt. A description of the structure of these datasets is available. Identifying Putative Biomarkers contains links that describe each of the 10 methods currently employed in the BioMarker Development Kit (BMDK). The following table lists the "peaks" that were selected by each method. Datasets containing only these putative biomarkers are stored as case_60_1a_pb.txt and control_60_1a_pb.txt. In this example, the second column states that catboot selected Peak 285 as the best discriminator, followed by Peaks 193, 228, 260, and 56. For many of these methods, ties can occur. Under the column labeled kolsmir, Peaks 113 and 285 had the third highest score, and Peaks 30, 163, and 228 tied for fifth.

BMDK Analysis

22 peaks selected as putative biomarkers by the 10 methods within BMDK

Peak	catboot	student	dtgini	dtinfg	nnfeat	chisq	kruswal	kolsmir	extreme	yip
2									4	
5								5		
30						4				
56	5									
60		5					5			5
68						4				
79						1				
88		3	2	4	5		4	1		3
110									4	
113			3	5	2			3		
163								5		
193	2									
206						2				
216			4	3					2	
220			5	1					1	
228	3	2			3	4	2	5		2
240						2				
241									2	
252		1	1	2	1	4	1	2		1
260	4									
261						4				
285	1	4			4		3	3	4	4

The next table lists which of the selected peaks were used in each of classifiers based on a distance-dependent 6-nearest neighbor algorithm. A Euclidean distance is used to determine the separation between any given pair of samples, but is determined four different ways. The first is the standard Euclidean distance where the absolute difference (ad) in peak intensities is used, while the other three allow a peak with a smaller overall intensity to contribute as much to the inter-sample distance as a large peak. The first uses the relative difference (rd) in peak intensities, which is the difference in a pair of intensities divided by their average. The second method initially scales all peak intensities so that there is a constant range (cr) of intensities of [0,10], and the third scales all intensities by the standard deviation (sd) of this peaks intensities which effectively converts all intensities into their z-values. The column headings give the number of peaks used in the classifier and the exact form of the distance metric. For example, the fourth column, "3-ad", states that three peaks are used and the distance metric uses the absolute difference in the intensities. When all triplets are examined, the classifier with the highest quality (sensitivity + specificity - %undetermined) used Peaks 113, 216, and 260. The best two-peak classifier that scaled the intensities to either a constant range ("2-cr") or by the standard deviation ("2-sd") used Peaks 216 and 252. It should be noted that "1-cr" and "1-sd" are not included since they produce the same result as "1-ad".

Peaks used in each of the best distance-dependent 6-nearest neighbor classifiers

Peak	1-ad	2-ad	3-ad	1-rd	2-rd	3-rd	2-cr	3-cr	2-sd	3-sd
2										
5										
30										
56										
60						X				
68										
79										
88										
110										
113			X		X	X		X		X
163										
193										
206						X				
216	X		X		X		X		X	
220										
228		X								
240										
241								X		X
252							X		X	
260			X							
261		X		X						
285								X		X

The next table lists the sensitivity, specificity, %undetermined and quality for each of the best classifiers that used the peaks in the preceding table.

Sensitivity, specificity, %undetermined, and quality (sensitivity + specificity - %undetermined) for each of the best distance-dependent 6-nearest neighbor classifiers using any of the 22 putative biomarkers.

Metric	1-ad	2-ad	3-ad	1-rd	2-rd	3-rd	2-cr	3-cr	2-sd	3-sd
Sens	70.0	70.0	69.6	70.0	69.8	76.7	69.1	71.7	67.3	71.7
Spec	66.7	68.3	74.6	66.7	66.7	76.9	73.2	71.2	71.9	71.2
%Undet	0.0	0.0	4.2	0.0	13.3	31.7	7.5	12.5	6.7	12.5
Quality	136.7	138.3	140.1	136.7	123.1	122.0	134.8	130.4	132.5	130.4

The best "3-rd" classifier (three peaks using the relative difference in peak intensities) in the preceding table produced a sensitivity of 76.6% and a specificity of 76.9%, but that 31.7% of all samples were scattered too far from their nearest neighbors and produced an "Undetermined" classification. To reduce the number of samples with no near neighbors in the classifier, the search over all peak-combinations was repeated with the requirement that %undetermined cannot exceed 5%. The results for the best classifier of each type are shown in the following table, and this table shows that a search over all peak-combinations did not find a single classifier with %undetermined of 5% or less for "2-rd", "3-rd", "3-cr", or "3-sd".

Sensitivity, specificity, %undetermined, and quality (sensitivity + specificity - %undetermined) for each of the best distance-dependent 6-nearest neighbor classifiers using any of the 22 putative biomarkers with the caveat that %undetermined cannot exceed 5.0%.

Metric	1-ad	2-ad	3-ad	1-rd	2-rd	3-rd	2-cr	3-cr	2-sd	3-sd
Sens	70.0	70.0	69.6	70.0	None	None	68.3	None	68.6	None
Spec	66.7	68.3	74.6	66.7	None	None	66.1	None	66.1	None
%Undet	0.0	0.0	4.2	0.0	None	None	3.3	None	3.3	None
Quality	136.7	138.3	140.1	136.7	None	None	131.1	None	131.1	None

Fingerprint Analysis

The following table lists the sensitivity, specificity, and their sum (quality) of the best and 200th best decision tree classifiers in four different runs. Though all decision trees may contain up to seven decision nodes in a symmetric tree, the first two runs convert a decision node into a terminal node of it contains at most 1% of the Cases or Controls and the second two runs convert a decision node into a terminal node if it contains at most 4% of the Cases or Controls. All four runs use a different seed to the random number generator in the evolutionary programming search and no two trees are allowed to use the same set of seven peak intensities in the same order. It should be noted that the search for optimum cut-values for each decision node is extremely limited and these results therefore represent a lower bound to the optimum decision tree for this dataset.

Sensitivity, specificity and quality (sensitivity + specificity) for the best and 200th best decision tree constructed from any of the 300 peak intensities. The evolutionary programming search used a population size of 200 and ran for 400 generations. A decision node became a terminal node when it contained 1% (no samples) or 4% (2 samples) of a given State.

Metric	1%		1%		4%		4%	
	1st	200th	1st	200th	1st	200th	1st	200th
Sensitivity	88.3	88.3	88.3	88.3	83.3	83.3	88.3	83.3
Specificity	88.3	86.7	86.7	81.7	86.7	83.3	88.3	88.3
Quality	176.7	175.0	175.0	170.0	170.0	166.7	176.7	171.7

The next table lists the sensitivity, specificity, and their sum (quality) of the best and 200th best classifier using the medoid classification algorithm. The search for the best set of five, six, or seven features from the full set of 300 again used an evolutionary programming algorithm with the requirement that no two sets of *N* features can be the same. For a given number of features, the first run examined all Cases and then all Controls, while the second examined all Controls and then all Cases. A different seed to the random number generator was used in all six runs.

Sensitivity, specificity and quality (sensitivity + specificity) for the best and 200th best medoid classifier algorithm in each of the two runs using 5-, 6-, and 7-peak intensities from the set of 300. The evolutionary programming search used a population size of 400 and ran for 800 generations with the requirement that there are at most 40 Case-cells and 40 Control-cells.

Metric	5-Features		5-Features		6-Features		6-Features		7-Features		7-Features	
	1st	200th	1st	200th	1st	200th	1st	200th	1st	200th	1st	200th
Sens	100.0	100.0	88.3	83.3	100.0	100.0	93.3	86.7	100.0	100.0	95.0	90.0
Spec	91.7	83.3	100.0	100.0	91.7	85.0	100.0	100.0	95.0	90.0	100.0	100.0
Quality	191.7	183.3	188.3	183.3	191.7	185.0	193.3	186.7	195.0	190.0	195.0	190.0