Analysis of Datasets Containing No Biological Information

Brian T. Luke (lukeb@ncifcrf.gov)

To examine the classification accuracy of various methods, and therefore the significance of each method, on datasets with no biological information, 30 artificial datasets with random peak intensities have been created. All datasets contain 300 "peaks" and have the same number of Cases and Controls (30, 42, 60, 90, 150, and 300 of each category). For each number of Cases and Controls, five independent datasets are available for download and analysis. Descriptions of the structure of the datasets as well as the procedure used in their generation are available. Below is a summary of the results for distinguishing Cases from Controls for all 30 datasets using the BioMarker Development Kit (BMDK), a decision tree (DT), and a medoid classification algorithm (MCA). Specific results for each dataset using these three classification methods are also available.

Table 1 lists the best classification results obtained by the BMDK suite of programs for each dataset. The second and third columns list the number of Cases and Controls in each dataset and the fourth column lists the number of putative biomarkers identified by the 10 methods currently employed in BMDK. The final three columns list the quality of the best classifier after an exhaustive examination of all sets of one, two, and three putative biomarkers using a distance-dependent 6-nearest neighbor algorithm, respectively. Since this classifier has the ability to return a classification of "Undetermined", the quality of this classifier is the sum of the sensitivity and specificity minus the percentage of samples that are "Undetermined". Since an "Undetermined" classification may be caused by an incomplete coverage of feature space by the available samples, which should be avoided, the results in Table 1 have the added restriction that %undetermined cannot exceed 5% for any classifier. With this restriction, the fifth dataset containing 30 Cases and 30 Controls (Set **30_5a**) could not find a single valid 3-peak classifier. These results use a standard Euclidean distance to determine the distance between samples in feature space since none of the other three available distance metrics were able to identify a valid 3-peak classifier for any of the 30 datasets.

If the minimum required accuracy of a classifier is a sensitivity and specificity of 85%, the minimum acceptable quality score is 165, assuming that %undetermined=5%. None of the results presented in Table 1 reach this minimum level of accuracy. This result is expected since the datasets are constructed to contain no information. It is also expected that the quality of the best classifier should decrease as the number of Cases and Controls increases. While this is generally true, the 1-peak classifier for the fifth dataset containing 60 Cases and 60 Controls (Set **60_5a**) produced an anomalously high quality of 151.7 (sensitivity=78.3%, specificity=73.3%, %undetermined=0.0%). An intensity plot for this peak is shown in Figure 1. The left column shows the intensities for all Cases, while the intensities for all Controls are in the left column. It is clear from this figure that the accuracy of the peak is caused by a high density of intensities for one category in a region with a low density of intensities for the other. The only way to justify this peak as a useful classifier is to assume that each category is composed of

several States and that each State is represented by a specific range of intensities for this peak. This division of the individuals in each category into multiple States would have to be biologically verified before this classifier should be used; even though its quality is still well below the minimum threshold.

The classification results using a decision tree (DT) are listed in Table 2. This table lists the quality (sum of sensitivity and specificity) of the best and 200th best decision tree across four runs for each dataset. Two of the runs convert a decision node into a terminal node if it contains at most 1% of either the Cases or Controls, while the other two runs increases this criteria to 4%. All four runs use a different seed to the random number generator. For the five datasets with 30 Cases and 30 Controls, 18 of the 20 runs identify at least 200 decision trees with an average sensitivity and specificity of 95% or higher. All runs identified 200 decision trees with an average sensitivity and specificity above 91% and one run identified at least 200 decision trees with a sensitivity and specificity of 100%.

Using and average sensitivity and specificity of 85% as a minimum required accuracy for a "useful" classifier, Table 2 shows that this level is regularly achieved if the dataset contains 60 Cases and 60 Controls or less. If there are 90 Cases and 90 Controls, an average sensitivity and specificity of 83.3% is achieved, which is slightly below the required accuracy. For the largest dataset (300 Cases and 300 Controls) the average sensitivity and specificity is below 70% but still much higher than the BMDK results in Table 1.

The classification accuracies (sum of sensitivity and specificity) using the mediod classification algorithm (MCA) are listed in Table 3. For each dataset, two runs are performed when five, six or seven of the 300 features are used in the classifier. The first examines all Cases and then all Controls while the second reverses this order. All six runs for a given dataset use a different seed to the random number generator and the number of Case-cells or Control-cells is not allowed to exceed two-thirds of the number of Cases or Controls, respectively. This would allow up to one-third of the samples to effectively be treated as a testing set while still maintaining full coverage of the fingerprint patterns.

If the dataset contains only 30 Cases and 30 Controls, all runs found at least one classifier that produced a sensitivity and specificity of 100% without requiring more than 20 Cases and 20 Controls for complete coverage. Two of the 6-feature runs and four of the 7-feature runs produced a final population where at least 200 unique classifiers produced a sensitivity and specificity of 100%. For the largest dataset (300 Cases and 300 Controls) one 5-feature classifier was identified that had an average sensitivity and specificity of 85%, the other nine runs found a classifier with an average sensitivity and specificity of 83.5% or higher. All 6-feature and 7-feature runs produced a final population that contains at least 200 unique classifiers with an average sensitivity and specificity above 85%; one 7-feature run found a classifier with an average sensitivity and specificity above 90.1%.

Conclusions

While it has been argued [Pet-03] that accurate classification of a testing set must imply some underlying biological principle, the results presented here clearly shows that this is not true for fingerprint-based classifiers; especially MCA classifiers. Hundreds of 7-feature MCA classifiers produce an average sensitivity and specificity above 90% for datasets with 150 Cases, 150 Controls, and only 300 features (Table 3) even though the datasets are constructed to contain no biological information. A chance [Ran-05a, Ran-05b] fitting of the data is highly possible since fingerprint-based classifiers appear to be overly flexible. Increasing the number of samples in the dataset generally decreases the quality of the best classifier if it is a chance fit, but increasing the number of features in the dataset will increase this quality. For fingerprint-based classifiers, the quality should improve if more features are used in the classifier; while the results in Table 1 suggest that increasing the number of features from two to three does not make much of an improvement, and may actually produce poorer results.

The bottom line is that fingerprint-based classifiers should not be used to analyze a dataset. Problems of chance, coverage, uniqueness, and significance result in concluding that fingerprint-based classifiers are not generalizable to the underlying population.

Figure 1: Intensities for the 60 cases (left column) and 60 controls (right column) for the peak that yielded a quality score of 151.7 (sensitivity=78.3%, specificity=73.3%, undetermined=0.0%) in the dataset of random peak intensities.

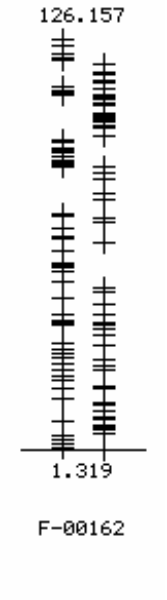

Table 1: Classification accuracy (sum of the sensitivity and specificity minus the percent undetermined) using between one and three peaks from the list of *NPB* putative biomarkers identified by BMDK in a distance-dependent 6-nearest neighbor classifier using absolute differences in the peak intensities and requiring that the %Undetermined is no more than 5%.

Set	Cases	Controls	NPB	1 Peak	2 Peaks	3 Peaks
30_1a	**30**	**30**	21	146.7	151.1	157.8
30_2a	**30**	**30**	22	137.3	153.3	155.3
30_3a	**30**	**30**	22	147.4	137.3	145.7
30_4a	**30**	**30**	21	143.3	144.9	147.6
30_5a	**30**	**30**	16	144.2	146.7	None
42_1a	**42**	**42**	27	142.9	135.7	113.0
42_2a	**42**	**42**	16	121.7	136.4	130.2
42_3a	**42**	**42**	22	131.4	136.2	137.3
42_4a	**42**	**42**	26	135.7	135.7	135.7
42_5a	**42**	**42**	21	131.0	134.7	129.8
60_1a	**60**	**60**	22	136.7	138.3	140.1
60_2a	**60**	**60**	29	121.7	129.7	128.9
60_3a	**60**	**60**	26	133.3	140.0	133.6
60_4a	**60**	**60**	22	130.0	139.3	133.7
60_5a	**60**	**60**	27	151.7	137.0	136.2
90_1a	**90**	**90**	24	121.1	133.9	125.3
90_2a	**90**	**90**	23	121.1	130.4	125.8
90_3a	**90**	**90**	27	123.3	126.7	123.3
90_4a	**90**	**90**	28	121.1	137.9	137.3
90_5a	**90**	**90**	27	118.9	131.1	135.6
150_1a	**150**	**150**	22	114.0	127.3	106.6
150_2a	**150**	**150**	25	116.7	125.3	125.3
150_3a	**150**	**150**	25	115.3	120.9	123.4
150_4a	**150**	**150**	22	118.0	123.3	125.3
150_5a	**150**	**150**	23	116.0	123.6	121.7
300_1a	**300**	**300**	26	115.7	117.7	120.2
300_2a	**300**	**300**	31	110.7	121.1	122.1
300_3a	**300**	**300**	29	111.3	121.7	119.1
300_4a	**300**	**300**	26	112.7	118.3	117.7
300_5a	**300**	**300**	26	113.7	115.0	119.5

Table 2: Classification accuracy (sum of the sensitivity and specificity) using the decision tree algorithm for the 1st and 200th best classifier as a function of the number of Cases and Controls.[(a)]

Set	Cases	Controls	1%		1%		4%		4%	
			1st	200th	1st	200th	1st	200th	1st	200th
30_1a	**30**	**30**	200.0	196.7	196.7	196.7	190.0	190.0	196.7	190.0
30_2a	**30**	**30**	196.7	196.7	196.7	193.3	196.7	193.3	200.0	196.7
30_3a	**30**	**30**	190.0	190.0	193.3	190.0	196.7	193.3	196.7	193.3
30_4a	**30**	**30**	196.7	196.7	196.7	193.3	193.3	18677	193.3	193.3
30_5a	**30**	**30**	196.7	193.3	200.0	200.0	190.0	183.3	196.7	193.3
42_1a	**42**	**42**	188.1	185.7	185.8	178.5	183.3	180.9	183.3	180.9
42_2a	**42**	**42**	188.1	185.7	185.8	178.5	183.3	180.9	183.3	180.0
42_3a	**42**	**42**	183.3	183.3	188.1	185.7	185.7	185.7	183.3	180.9
42_4a	**42**	**42**	183.3	180.9	178.5	178.6	178.5	176.2	180.9	178.5
42_5a	**42**	**42**	190.5	185.7	185.7	180.9	188.1	185.7	180.9	176.2
60_1a	**60**	**60**	176.6	175.0	175.0	170.0	175.0	166.6	176.6	171.6
60_2a	**60**	**60**	176.6	171.6	175.0	171.6	171.6	168.3	173.3	168.3
60_3a	**60**	**60**	171.6	171.6	178.3	175.0	176.6	173.4	163.3	171.7
60_4a	**60**	**60**	176.6	173.3	176.6	171.6	176.7	173.3	171.6	168.3
60_5a	**60**	**60**	171.6	170.0	170.0	168.3	171.6	168.3	170.0	165.0
90_1a	**90**	**90**	156.7	155.6	160.0	158.9	160.0	158.9	162.3	158.9
90_2a	**90**	**90**	166.7	165.6	163.4	160.0	160.0	156.7	158.9	155.6
90_3a	**90**	**90**	158.9	157.8	161.1	158.9	158.9	156.7	164.5	152.2
90_4a	**90**	**90**	160.0	158.9	162.2	160.0	160.0	156.7	160.0	157.8
90_5a	**90**	**90**	166.7	164.4	162.2	158.9	166.7	165.6	161.1	158.9
150_1a	**150**	**150**	152.0	150.0	152.0	150.0	150.0	148.0	152.7	150.0
150_2a	**150**	**150**	149.3	146.0	150.7	149.3	150.7	149.3	151.3	150.0
150_3a	**150**	**150**	151.3	148.7	150.7	148.7	149.3	147.3	150.7	149.3
150_4a	**150**	**150**	152.0	150.0	149.3	148.0	155.3	154.0	153.3	151.3
150_5a	**150**	**150**	152.7	151.3	154.0	152.7	149.3	146.7	148.7	146.7
300_1a	**300**	**300**	138.3	137.3	136.3	135.3	137.3	136.0	135.7	134.7
300_2a	**300**	**300**	137.0	136.0	136.3	135.3	137.0	135.7	137.7	136.3
300_3a	**300**	**300**	136.7	136.0	137.0	136.0	134.0	132.0	136.3	135.0
300_4a	**300**	**300**	136.0	135.0	136.7	135.3	136.3	133.7	136.0	134.7
300_5a	**300**	**300**	135.0	133.7	136.3	135.0	135.7	134.7	135.7	135.0

[(a)]A decision node was converted to a terminal node if it contained at most 1% of the samples from either category (1% runs) or if it contained at most 4% of the samples from either category (4% runs). All four runs used different seeds to the random number generator that controlled the Evolutionary Programming search.

Supplemental Table 6: Classification accuracy (sum of the sensitivity and specificity) using the medoid classification algorithm for the 1st and 200th best classifier as a function of the number of cases and controls from two runs using five, six, and seven peaks with random intensities.[(a)]

Set	Cases	Controls	Run	5 Peaks		6 Peaks		7 Peaks	
				1st	200th	1st	200th	1st	200th
30_1a	**30**	**30**	**1**	200.0	193.3	200.0	196.7	200.0	196.7
			2	200.0	193.3	200.0	196.7	200.0	200.0
30_2a	**30**	**30**	**1**	200.0	193.3	200.0	196.7	200.0	196.7
			2	200.0	193.3	200.0	196.7	200.0	196.7
30_3a	**30**	**30**	**1**	200.0	190.0	200.0	196.7	200.0	200.0
			2	200.0	193.3	200.0	200.0	200.0	196.7
30_4a	**30**	**30**	**1**	200.0	193.3	200.0	196.7	200.0	196.7
			2	200.0	193.3	200.0	193.3	200.0	200.0
30_5a	**30**	**30**	**1**	200.0	193.3	200.0	200.0	200.0	196.7
			2	200.0	193.3	200.0	196.7	200.0	200.0
42_1a	**42**	**42**	**1**	195.2	188.1	197.6	190.5	197.6	192.9
			2	195.2	188.1	197.6	190.5	197.6	192.9
42_2a	**42**	**42**	**1**	197.6	188.1	197.6	192.9	197.6	192.9
			2	192.9	188.1	197.6	192.9	197.6	195.2
42_3a	**42**	**42**	**1**	192.9	188.1	197.6	190.5	197.6	192.9
			2	195.2	188.1	195.2	190.5	197.6	192.9
42_4a	**42**	**42**	**1**	195.2	188.1	197.6	190.5	197.6	192.9
			2	195.2	188.1	195.2	190.5	197.6	195.2
42_5a	**42**	**42**	**1**	192.9	185.7	195.2	188.1	197.6	192.9
			2	195.2	188.1	197.6	192.9	197.6	195.2
60_1a	**60**	**60**	**1**	191.7	183.3	191.7	185.0	195.0	190.0
			2	188.3	181.7	193.3	186.7	195.0	190.0
60_2a	**60**	**60**	**1**	188.3	181.7	193.3	185.0	195.0	186.7
			2	193.3	183.3	193.3	185.0	193.3	186.7
60_3a	**60**	**60**	**1**	190.0	183.3	192.7	185.0	193.3	186.7
			2	193.3	183.3	193.3	186.7	193.3	188.3
60_4a	**60**	**60**	**1**	193.3	183.3	193.3	185.0	195.0	190.0
			2	190.0	183.3	193.3	185.0	191.7	186.7
60_5a	**60**	**60**	**1**	193.3	181.7	193.3	186.7	195.0	190.0
			2	190.0	183.3	191.7	186.7	193.3	190.0
90_1a	**90**	**90**	**1**	184.4	178.9	188.9	181.1	188.9	183.3
			2	184.4	177.8	185.6	180.0	188.9	182.2
90_2a	**90**	**90**	**1**	185.6	177.8	188.9	181.1	190.0	184.4
			2	184.4	178.9	186.7	180.0	190.0	183.3
90_3a	**90**	**90**	**1**	184.4	177.8	186.7	180.0	191.1	183.3
			2	187.8	178.9	188.9	182.2	188.9	184.4
90_4a	**90**	**90**	**1**	183.3	178.9	186.7	181.1	188.9	182.2
			2	185.6	177.8	187.8	181.1	190.0	182.2

90_5a	**90**	**90**	**1**	185.6	177.8	185.6	180.0	188.9	182.2
			2	186.7	181.1	187.8	182.2	190.0	185.6
150_1a	**150**	**150**	**1**	182.0	176.0	183.3	179.3	186.7	181.3
			2	180.0	174.0	182.0	178.0	184.7	179.3
150_2a	**150**	**150**	**1**	183.3	174.7	182.7	178.7	184.7	180.7
			2	181.3	174.7	182.7	178.0	187.3	180.0
150_3a	**150**	**150**	**1**	180.0	174.7	183.3	178.7	184.0	180.7
			2	180.0	175.3	184.0	179.3	185.3	180.7
150_4a	**150**	**150**	**1**	180.7	175.3	182.0	178.7	184.0	180.7
			2	180.0	174.7	185.3	178.7	185.3	181.3
150_5a	**150**	**150**	**1**	182.0	175.3	182.7	178.7	185.3	180.7
			2	180.0	174.0	182.7	178.0	184.0	180.0
300_1a	**300**	**300**	**1**	167.3	163.3	176.3	171.7	179.7	175.0
			2	170.3	163.3	175.7	172.3	178.7	175.0
300_2a	**300**	**300**	**1**	169.7	163.3	175.3	171.3	178.7	174.0
			2	169.3	163.3	179.0	172.7	180.3	175.7
300_3a	**300**	**300**	**1**	168.0	163.3	175.3	172.3	179.3	175.3
			2	168.7	163.3	178.3	172.3	178.7	175.3
300_4a	**300**	**300**	**1**	168.0	163.3	176.7	172.3	178.3	175.3
			2	167.0	163.3	176.7	171.3	178.7	175.0
300_5a	**300**	**300**	**1**	168.3	163.3	176.7	171.7	178.7	174.3
			2	168.3	163.7	176.3	172.7	178.7	175.0

[(a)]Run 1 examined all cases and then controls, while Run 2 examined the controls and then the cases.

(Last updated 5/2/07)