Return to Contents PDF Version

Random Intensity Datasets: 150 Cases, 150 Controls

Brian T. Luke (lukeb@ncifcrf.gov)

These five pairs of datasets contain 300 features with 150 Cases and 150 Controls. These datasets are constructed with random peak intensities so that they contain no biological information. Structure of the Datasets contains a general description of datasets that can be used by programs within the BioMarker Development Kit (BMDK). Since the Cases and Controls are stored in different files, the class indices are not included in the data. Each feature has a single label, but they are simply "F-00001" through "F-00300". Each dataset has an associated document that describes the results of an analysis using the BioMarker Development Kit (BMDK), and classifiers based on a decision tree (DT) and a medoid classification algorithm (MCA). To reduce the amount of repeated information in these tables of results, Description of the Tables gives details about each table.

Analysis	#Cases #Controls	#Features	Case Dataset	Control Dataset	Analysis
Random_Intensity_150_1a	150	300	case_150_1a.txt	control_150_1a.txt	Tables
Random_Intensity_150_2a	150	300	case_150_2a.txt	control_150_2a.txt	Tables
Random_Intensity_150_3a	150	300	case_150_3a.txt	control_150_3a.txt	Tables
Random_Intensity_150_4a	150	300	case_150_4a.txt	control_150_4a.txt	Tables
Random_Intensity_150_5a	150	300	case_150_5a.txt	control_150_5a.txt	Tables

The following table lists the best classification observed results for each dataset-pair.

Set	NPB	BMDK-1	BMDK-2	BMDK-3	DT	MCA-5	MCA-6	MCA-7
150_1a	22	114.0	127.3	106.6	152.7	182.0	183.3	186.7
150_2a	25	116.7	125.3	125.3	151.3	183.3	182.7	187.3
150_3a	25	115.3	120.9	123.4	151.3	180.0	184.0	185.3
150_4a	22	118.0	123.3	125.3	155.3	180.7	185.3	185.3
150_5a	23	116.0	123.6	121.7	154.0	182.0	182.7	185.3

For each set of Cases and Controls, BMDK uses 10 different methods to search for putative biomarkers, and the number of putative biomarkers (NPB) identified for each set is listed in the second column (the Tables shown in the links above give details on which procedures selected which features). BMDK only uses these putative biomarkers to construct the final classifier based on a distance-dependent K-nearest neighbor algorithm. This classifier allows for an "undetermined" classification, so the quality metrics shown above are the sum of the overall sensitivity and specificity minus the percent "undetermined" from a leave-one-out cross-validation analysis, with the constraint that no more than 5% of the samples can be "undetermined". The third, fourth and fifth columns list the best result using between one and three of the putative biomarkers, respectively. For the DT and MCA classifiers, the quality is the sum of the sensitivity and specificity.

None of the final BMDK classifiers produced a sensitivity and specificity above 65%.

The best DT classifiers (Column 6) containing up to seven decision nodes yielded an average sensitivity and specificity between 70.7 and 77.7% for the 300 samples. The final three columns of the preceding table show the best results for an MCA classifier using five, six, and seven features, respectively. The best 5-feature classifiers had an average sensitivity and specificity of at least 90.0%, while the 6- and 7-feature classifiers has an average sensitivity and specificity of at least 91.3 and 92.7%, respectively. In all cases, the MCA classifier was constructed after effectively separating the data into a training set containing 100 Cases and 100 Controls, and a testing set containing 50 Cases and 50 Controls.

While there is a significant decrease in the quality of the DT classification, it is clear that the fingerprint-based methods are able to classify these samples to a higher accuracy than the biomarker-based, even though these datasets are constructed to contain no biological information. These fingerprint-based methods are designed so that the average sensitivity and specificity cannot decrease if more features are used. The observed decrease for the 6-feature classifier in Set 150_2a relative to the best 5-feature classifier is simply due to an incomplete sampling of possible feature-combinations. Conversely, for three of the five sets of data the 3-feature BMDK classifier did not perform better than the 2-feature classifier. If a minimum required "accuracy for publication" is set at 85%, for example, a decision tree with more than seven decision nodes would be required, while even a 5-feature MCA classifier has sufficient accuracy. No biomarker-based classifier would ever be able to obtain this accuracy for these datasets.

(Last updated 9/1/07)