Return to Contents PDF Version

Random Intensity Datasets: 42 Cases, 42 Controls

Brian T. Luke (lukeb@ncifcrf.gov)

These five pairs of datasets contain 300 features with 42 Cases and 42 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_42_1a	42	300	case_42_1a.txt	control_42_1a.txt	Tables
Random_Intensity_42_2a	42	300	case_42_2a.txt	control_42_2a.txt	Tables
Random_Intensity_42_3a	42	300	case_42_3a.txt	control_42_3a.txt	Tables
Random_Intensity_42_4a	42	300	case_42_4a.txt	control_42_4a.txt	Tables
Random_Intensity_42_5a	42	300	case_42_5a.txt	control_42_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
42_1a	27	142.9	135.7	113.0	188.1	195.2	197.6	197.6
42_2a	16	121.7	136.4	130.2	188.1	197.6	197.6	197.6
42_3a	22	131.4	136.2	137.3	187.1	195.2	197.6	197.6
42_4a	26	135.7	135.7	135.7	183.3	195.2	197.6	197.6
42_5a	21	131.0	134.7	129.8	190.5	195.2	197.6	197.6

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.

For these five datasets, none of the final BMDK classifiers produced a sensitivity and specificity above 70%. The 3-feature classifier for Set 42_3a had a sensitivity of 71.8%, a specificity of 69.0% with a 3.6% of the samples (three samples) receiving an “undetermined” classification.

The best DT classifiers (Column 6) containing up to seven decision nodes misclassified between four and seven samples, yielding an average sensitivity and specificity between 91.7 and 95.2% across all 84 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 misclassified either one or two samples, yielding an average sensitivity and specificity of at least 96.7%. The 6- and 7-feature classifiers always misclassified only a single subject producing an average sensitivity and specificity of 98.8%. In all cases, the classifier was constructed after effectively separating the data into a training set containing 28 Cases and 28 Controls, and a testing set containing 14 Cases and 14 Controls.

It is clear that the fingerprint-based methods are able to classify these samples to a very high accuracy, even though these datasets are constructed to contain no biological information.

(Last updated 9/1/07)