

Random Intensity Datasets: 30 Cases, 30 Controls

Brian T. Luke (lukeb@ncifcrf.gov)

These five pairs of datasets contain 300 features with 30 Cases and 30 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_30_1a	30	300	case_30_1a.txt	control_30_1a.txt	Tables
Random_Intensity_30_2a	30	300	case_30_2a.txt	control_30_2a.txt	Tables
Random_Intensity_30_3a	30	300	case_30_3a.txt	control_30_3a.txt	Tables
Random_Intensity_30_4a	30	300	case_30_4a.txt	control_30_4a.txt	Tables
Random_Intensity_30_5a	30	300	case_30_5a.txt	control_30_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
30_1a	21	146.7	151.1	157.8	200.0	200.0	200.0	200.0
30_2a	22	137.3	153.3	155.3	200.0	200.0	200.0	200.0
30_3a	22	147.4	137.3	145.7	196.7	200.0	200.0	200.0
30_4a	21	143.3	144.9	147.6	196.7	200.0	200.0	200.0
30_5a	16	144.2	146.7	None	200.0	200.0	200.0	200.0

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. It should be noted for Set 30_5a, no combination of three of the 16 putative biomarkers produced a classifier that contains 5% “undetermined” or less. 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 80%. The 3-feature classifier for Set 30_1a had a sensitivity of 76.7%, a specificity of 82.8% with a single Control receiving an “undetermined” classification.

In contrast, it was possible to construct a decision tree (DT) containing at most seven decision nodes that correctly classified all 60 samples for three of the five datasets; the other two datasets yielded a decision tree that misclassified one of the samples (Column 6). The final three columns of the preceding table show the best results for an MCA classifier using five, six, and seven features, respectively. In all cases, this algorithm was able to correctly classify all 60 samples after effectively separating the data into a training set containing 20 Cases and 20 Controls, and a testing set containing 10 Cases and 10 Controls.

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

(Last updated 9/1/07)