Distance Metrics

Brian T. Luke (lukeb@ncifcrf.gov)

The final classifier currently used by the BioMarker Development Kit (BMDK) is a distance-dependent K-nearest neighbor classifier that only uses putative biomarkers after they have been examined for any possible sources of bias. The probability that a sample belongs to the same category as a neighbor is inversely proportional to the distance between the samples. In its current version, BMDK can use one of four different distance metrics, all based on a Euclidean distance, using the intensities of the N putative biomarkers. The general form of this distance between two samples is as follows.

$$d_{i,j} = \sqrt{\sum_{n=1}^{N} f(I_{i,n}, I_{j,n})^2}$$

In this equation, $I_{i,n}$ and $I_{j,n}$ are the intensities of the ith and j^{th} samples for the nth putative biomarker, and $f(I_{i,n},I_{j,n})$ is some function of these intensities.

The simplest distance metric is the standard Euclidean distance, or L_2-norm, where $f(I_{i,n},I_{j,n})$ is the absolute difference in intensities.

$$f(I_{i,n},I_{j,n}) = |I_{i,n} - I_{j,n}|$$

If two or more putative biomarkers are used in the classifier (N > 1) one of the features may have very little effect on the overall distance if its maximum intensity is significantly less than the other features. One method to remove this dependence is to use the relative difference in peak intensities.

$$f(I_{i,n},I_{j,n}) = 2|I_{i,n} - I_{j,n}| \,/\, (I_{i,n} + I_{j,n})$$

In this metric, the difference in intensities between two samples is divided by their average intensity.

Another procedure to the effect of different maximum intensities for the N putative biomarkers is to first scale the intensities. The first of these linearly scales each of N putative biomarkers so that they span the range [0.0,10.0] across all samples.

$$f(I_{i,n},I_{j,n}) = 10|I_{i,n} - I_{j,n}| \,/\, (I_{max,n} + I_{min,n})$$

In this equation, $I_{max,n}$ and $I_{min,n}$ are the maximum and minimum intensities for the nth putative biomarker, respectively.

The final procedure, sometimes called a scaled Euclidean distance, divides the difference in intensities by the standard deviation of all intensities for this putative biomarker, σ_n.

$f(I_{i,n},I_{j,n}) = |I_{i,n} - I_{j,n}| \,/\, \sigma_n$

This metric effectively converts all intensities to their *z*-value before their absolute difference is calculated.

Tests are currently being conducted to determine which distance metric is best using real datasets. All studies to date have used the standard Euclidean distance (L_2-norm).

(Last updated 4/30/07)