Distance-Dependent K-Nearest Neighbors

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 State as a neighbor is inversely proportional to the distance between the samples. In practice, determining the number of States in each category occurs after the best classifier is determined for a given number of putative biomarkers, so the search for the best classifier initially uses categories, not States. In addition, the number of neighbors examined, *K*, is usually large (at least six) so that the local region around the sample being classified is sufficiently large.

If the k^{th} nearest neighbor is a distance d_k away from the sample being examined in the *N*-dimensional feature space determined by the *N* putative biomarkers used in the classifier, and *ck* is the category of this neighbor, the un-normalized probability that this sample belongs to this category is given by the following expression.

$$P_{ck} = \alpha / d_k$$

The scaling factor α is controlled by a user-supplied fractional value *f*, which represents the fraction of the maximum possible distance between samples for which the probability drops to 0.5. This maximum possible distance represents the distance of the diagonal, d_d, for the *N*-dimensional rectangle containing all points. This is determined by setting one hypothetical point at the minimum intensities of all *N* putative biomarkers and a second hypothetical point at the maximum intensities and finding their distance.

$P_{ck} = 0.5$ when $d_k = f(d_d)$

The scaling factor is then

$$\alpha = f(d_d) / 2.0$$

Associated with each probability that it belongs to the same category as a neighbor is an un-normalized probability that the classification is “Undetermined”, $P_{u,k}$.

$$\begin{aligned} P_{u,k} &= U && \text{if } P_{ck} < (1 - 2U) \\ &= (1 - P_{ck})/2 && \text{if } (1\text{-}2U) \le P_{ck} \le 1.0 \\ &= 0 && \text{if } P_{ck} > 1.0 \end{aligned}$$

If $U = 0.1$, for example, then $P_{u,k}$ equals 0.1 if P_{ck} is less than 0.8; it linearly drops from 0.1 to 0.0 as P_{ck} increases from 0.8 to 1.0; and is 0.0 whenever P_{ck} is greater than 1.0. This term is included so that if one of the removed samples is an outlier (has an intensity

that is significantly far from the centroids) it will receive a classification of "Undetermined".

If the dataset contains C categories, the normalized probability that it belongs to a given category, $P_{c'}^{n}$, is therefore

$$P_{c'}^{n} = P_{c'} \Big/ \left\{ \sum_{c=1}^{C} P_c + \sum_{k=1}^{K} P_{u,k} \right\}$$

In the denominator of this expression is the sum over the un-normalized probabilities that the sample belongs to any of the *C* categories and the sum of the un-normalized probability that the classification is "Undetermined" over all *K* neighbors. The normalized probability that the classification is "Undetermined" is given by the following expression.

$$P_{u}^{n} = \sum_{k=1}^{K} P_{u,k} \Big/ \left\{ \sum_{c=1}^{C} P_c + \sum_{k=1}^{K} P_{u,k} \right\}$$

If any value of $P_{c'}^{n}$ is 0.5 or greater, then the sample is assigned to this category; otherwise it receives a classification of "Undetermined".

(Last updated 4/30/07)