Modifying and Applying the kth Nearest Neighbor Algorithm to Enhance the Classification of Breast Cancer Tumors

Khurana, Rohit

The kth nearest neighbor algorithm (kNN) is a non-parametric method used for classification, in which a query point is compared to a data set of points with predetermined classes. The nearest neighbors are determined through the smallest Euclidean distance from this query point; its classification is then discerned by the label that is most abundant among its neighbors. This experiment sought to improve classification of breast cancer samples by modifying kNN. The standard kNN only classifies points based on distance, which could yield a high percentage error. Marginal points can be misclassified because outliers are given an equal weight to points more representative of their classes. A breast cancer data set, in which each patient had ten features, was utilized. There were two themes of kNN modification: modifying how nearest neighbors were determined (Point Radius Method) as well as determining the confidence of neighbors (Gaussian Class Confidence Weights, Fuzzy Algorithm, Local Gaussian Function, Coefficient Mean Distance). The point radius method achieved the highest percentage accuracy when k = 1. For both the Gaussian weighting scheme and Fuzzy Algorithm, error rates decreased as the number of nearest neighbors increased. Local Gaussian function witnessed the most stable percentage accuracies and outperformed standard kNN algorithm at many k values. Lastly, coefficient mean distance achieved high percentage accuracies for high k values (e.g. 173). The optimal range for neighbors, averaged across all modifications, was between k = 11 and k = 15.

Awards Won: