International Journal of Innovative
Computing, Information and Control
Volume 4, Number 2, February 2008

 ISSN 1349-4198
pp. 369–379

COMBINATION OF MULTIPLE REAL-VALUED NEAREST NEIGHBOR CLASSIFIERS BASED ON DIFFERENT FEATURE SUBSETS WITH FUZZY INTEGRAL

Li-Juan Wang[1,2], Xiao-Long Wang[1] and Yuan-Chao Liu[2]

[1]Department of Computer Science and Technology
Harbin Institute of Technology
Harbin 150001, P. R. China

[2]Faculty of Mathematics and Computer Science
Hebei University
Baoding 071002, P. R. China
ljwang@insun.hit.edu.cn

Received December 2006; revised April 2007

Abstract. *Generally, the curse of dimensionality leads to great bias of NNC. However, in this paper, multiple real-valued NNCs based on different feature subsets are combined by fuzzy integral so that the bias of NNC in high dimensionality is minimized, which is called FI-MRNNC. In FI-MRNNC, the feature set is partitioned into several low dimensionality feature subsets, where fuzzy measure is used to measure the importance of each feature subset and the interaction between feature subsets in its decision making process. According to the FI-MRNNC's classification accuracy, GA not only partitions the feature set into several feature subsets but also defines a density value for the corresponding feature subset. Experimental results on some UCI databases illustrate that FI-MRNNC can reduce the bias of NNC, especially in high dimensionality.*
Keywords: Real-valued nearest neighbor classifier, Feature subset partition, Fuzzy measure, Fuzzy integral, Multiple-classifier combination, GA

1. **Introduction.** Nearest neighbor classifier (NNC) is one of the simple, popular and non-parametric classifiers. NNC simply stores the training data and retrieves local information to classify the unknown data, which does not need the training phase to find the complex target function on the whole data set.

Despite its simplicity, NNC in high dimensional feature space is severely biased with finite training data due to the curse of dimensionality [1]. There are several methods proposed to minimize the bias of NNC, such as Re-sampling technique [2,3], Large margin nearest neighbor classifiers [4,5] and Synthetic pattern generation method [6,7]. A straight method to this problem is to partition the high dimensionality feature set into several low dimensionality feature subsets, each feature subset independently classified by one NNC. The final decisions of multiple NNCs are aggregated by the combination technique.

Bryll [8] thinks that the performance of feature subset partition method is superior to that of the data partition method in ensemble learning. Bremain [9] experimentally demonstrates that NNC, unlike decision tree, stores a large number of prototypes and is stable to the change of training data generated by Bagging and Boosting. Langley and Iba [10] find out that NNC is drastically varied with the addition of irrelevant features. Bay [11] combines multiple NNCs each using a random subset of features (MFS) to improve the performance of standard NNC.

In the preceding method, because the target function of NNC is a discrete-valued function, the final decisions of multiple NNCs based on different feature subsets are combined