

Effective Nearest Neighbor Methods for Multiclass Cancer Classification Using Microarray Data

Satoshi Nijjima¹

Satoru Kuhara²

nijjima@grt.kyushu-u.ac.jp

kuhara@grt.kyushu-u.ac.jp

¹ Department of Bioinformatics, Graduate School of Systems Life Sciences, Kyushu University, Hakozaki 6-10-1, Higashi-ku, Fukuoka 812-8581, Japan

² Faculty of Agriculture, Kyushu University, Hakozaki 6-10-1, Higashi-ku, Fukuoka 812-8581, Japan

Keywords: cancer classification, gene selection, nearest neighbor methods

1 Introduction

In cancer classification problems using microarray data, an increasing number of studies have successfully demonstrated the effectiveness of state-of-the-art supervised machine learning methods such as support vector machines (SVMs). On the other hand, recent comparative studies suggest that, as far as gene selection is applied reasonably, simple and classical classifiers such as k -nearest neighbor (k -NN) perform as well as or even better than more complex methods including SVMs (e.g. [1, 3]). Thus, a promising approach to obtaining higher classification performance would be to apply modified NN methods. This study explores the applicability of two modified NN methods called k -discriminant adaptive nearest neighbor (k -DANN) [2] and K -local hyperplane distance nearest neighbor (HKNN) [4], both of which are known to be effective for high-dimensional data. To this end, we performed a comparative study on multiclass cancer classification using five public datasets, five classification methods including one-versus-all SVM (OVASVM) and one-versus-one SVM (OVOSVM) with linear and nonlinear kernels, in combination with four gene selection criteria.

2 Methods

For the standard k -NN method to be effective, the class conditional probabilities are assumed to be locally approximately constant. In a high-dimensional space, however, this assumption is severely violated. Such inhomogeneity often deteriorates the performance of k -NN. The k -DANN method adaptively selects a distance measure, making the class conditional probabilities locally more homogeneous. Whereas k -NN searches k nearest neighbors in a hypersphere, k -DANN in a hyperellipsoid.

If one thinks of each class as a low-dimensional manifold embedded in a high-dimensional space, it is natural to assume that this manifold is locally linear. Since microarray samples are sparsely distributed, missing samples would appear as *holes* introducing artifacts in the decision boundary of k -NN, which leads to poor generalization performance. To overcome this drawback, the HKNN method *fantasizes* the missing points, based on a local linear approximation of the manifold of each class.

3 Results

The public datasets used in this study are: CNS (Pomeroy *et al.*, 2002), NCI60 (Ross *et al.*, 2000), AC (Garber *et al.*, 2001), GNF (Su *et al.*, 2001), and ALL (Yeoh *et al.*, 2002). For the CNS, NCI60,

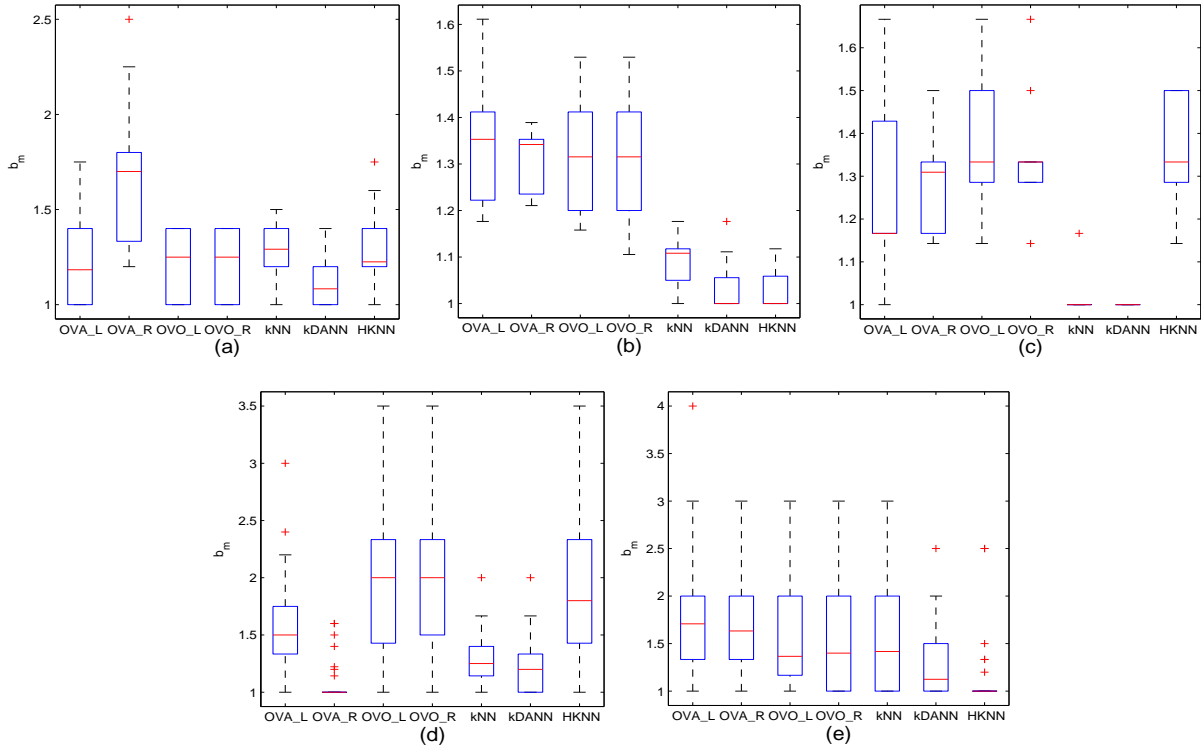


Figure 1: Performance distributions for the (a) CNS (b) NCI60 (c) AC (d) GNF (e) ALL datasets; OVA_L and OVA_R denote OVASVM with linear kernel and OVASVM with RBF kernel, respectively.

and AC datasets, classification was performed using various numbers of top-ranked genes selected by the BW ratio [1], and the performance of each classifier was evaluated using leave-one-out cross-validation (LOOCV). For the GNF and ALL datasets, information gain, twing rule, and sum of variances were employed as gene selection criteria, and the classification performance was evaluated using independent test samples. Figure 1 plots for each dataset the performance distributions, defined as $b_m = e_m / \min_{\ell \in \{1, \dots, 7\}} e_\ell$, where e_m denotes the error rate of a classification method m when a particular number of genes are used. These results have demonstrated that k -DANN performs better than k -NN, and often outperforms the multiclass SVMs. Although HKNN has shown even poorer performance than k -NN, it appears to be effective when a relatively large number of training samples per class are available. The detailed results as well as the algorithms used for k -DANN and HKNN will be presented in our poster.

References

- [1] Dudoit, S., Fridlyand, J., and Speed, T., Comparison of discrimination methods for the classification of tumors using gene expression data, *J. Am. Stat. Assoc.*, 97:77–87, 2002.
- [2] Hastie, T. and Tibshirani, R., Discriminant adaptive nearest neighbor classification, *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(6):607–615, 1996.
- [3] Li, T., Zhang, C., and Ogihara, M., A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression, *Bioinformatics*, 20(15):2429–2437, 2004.
- [4] Vincent, P. and Bengio, Y., K-local hyperplane and convex distance nearest neighbor algorithms, *Advances in Neural Information Processing Systems 14*, 985–992, 2002.