Dataset complexity in gene expression based cancer classification using ensembles of k-nearest neighbors.

Author: Okun Oleg
Priisalu Helen

Author Affiliation: University of Oulu, Department of Electrical and Information Engineering, P.O. Box 4500, Oulu 90014, Finland. oleg@ee.oulu.fi

Source: Artif Intell Med. 2009 Feb-Mar;45(2-3):151-62

Language: English

Publication Type: Article

Keywords: Gene Expression
Humans
Neoplasms - classification - genetics

Abstract: OBJECTIVE: We explore the link between dataset complexity, determining how difficult a dataset is for classification, and classification performance defined by low-variance and low-biased bolstered resubstitution error made by k-nearest neighbor classifiers. METHODS AND MATERIAL: Gene expression based cancer classification is used as the task in this study. Six gene expression datasets containing different types of cancer constitute test data. RESULTS: Through extensive simulation coupled with the copula method for analysis of association in bivariate data, we show that dataset complexity and bolstered resubstitution error are associated in terms of dependence. As a result, we propose a new scheme for generating ensembles of classifiers that selects subsets of features of low complexity for ensemble members, which constitutes the accurate members according to the found dependence relation. CONCLUSION: Experiments with six gene expression datasets demonstrate that our ensemble generating scheme based on the dependence of dataset complexity and classification error is superior to a single best classifier in the ensemble and to the traditional ensemble construction scheme that is ignorant of dataset complexity.

PubMed ID: 18790620 View in PubMed