Artificial Intelligence in Medicine 58 (2013) 155–163

Contents lists available at SciVerse ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Subpopulation-specific confidence designation for more informative biomedical classification

Chuanlei Zhang[a], Ralph L. Kodell[b,*]

[a] Department of Applied Mathematics and Computer Science, Philander Smith College, 900 W. Daisy L. Gatson Bates Dr., Little Rock, AR 72202, United States
[b] Department of Biostatistics, #781, University of Arkansas for Medical Sciences, 4301 W. Markham Street, Little Rock, AR 72205, United States

ARTICLE INFO

Article history:
Received 18 September 2012
Received in revised form 25 April 2013
Accepted 27 April 2013

Keywords:
Cross-validation
Genomic prediction
Individualized therapy
Population heterogeneity

ABSTRACT

Objective: Although classification algorithms are promising tools to support clinical diagnosis and treatment of disease, the usual implicit assumption underlying these algorithms, that all patients are homogeneous with respect to characteristics of interest, is unsatisfactory. The objective here is to exploit the population heterogeneity reflected by characteristics that may not be apparent and thus not controlled, in order to differentiate levels of classification accuracy between subpopulations and further the goal of tailoring therapies on an individual basis.
Methods and materials: A new subpopulation-based confidence approach is developed in the context of a selective voting algorithm defined by an ensemble of convex-hull classifiers. Populations of training samples are divided into three subpopulations that are internally homogeneous, with different levels of predictivity. Two different distance measures are used to cluster training samples into subpopulations and assign test samples to these subpopulations.
Results: Validation of the new approach's levels of confidence of classification is carried out using six publicly available datasets. Our approach demonstrates a positive correspondence between the predictivity designations derived from training samples and the classification accuracy of test samples. The average difference between highest- and lowest-confidence accuracies for the six datasets is 17.8%, with a minimum of 11.3% and a maximum of 24.1%.
Conclusion: The classification accuracy increases as the designated confidence increases.

1. Introduction

Ensemble learning is the process by which multiple classifiers are combined to solve a problem in computational intelligence [1]. The idea of ensemble learning is to increase prediction accuracy by combining the strengths of a collection of simpler base models [2]. In binary classification problems, majority voting among ensemble members is a common approach for combining class labels to predict the class of an unknown sample [3]. A comprehensive review of ensemble-based methods is provided by Rokach [4].

In addition to simply obtaining a decision (i.e., prediction or classification) from a classifier, it is useful to have a measure of confidence in that decision [5]. With simple majority voting, the very structure of the vote naturally allows assigning a degree of confidence to a particular decision, in that high agreement among ensemble members tends to be associated with high confidence and vice versa [1]. This is related to using ranges of predictions expressed on a probability scale as indicators of prediction confidence, e.g., 0.7–1.0 or >0.8 implies high confidence [6,7]. Of course, a statistical lower confidence limit can be calculated for each decision assuming a binomial distribution; however, its utility will depend on both the size of the ensemble and the degree of independence among the ensemble members. It is also possible to estimate the posterior probability of the class chosen by the ensemble for any given test instance, and use that estimate as the confidence measure for that instance [8]. Alternatively, empirical measures of performance calculated over a set of representative samples are good indicators of the confidence that can be placed in an ensemble's decisions. In binary classification problems, for example, the positive predictive value and negative predictive value, which arise from a frequentist interpretation of Bayes' theorem, give overall measures of confidence in predictions made by the ensemble. These two measures, along with overall accuracy, sensitivity and specificity, are common performance measures for assessing the degree of confidence that can be placed in a classifier's decisions.

In this paper, we develop a subpopulation basis for calculating the performance measures commonly used to assess majority-voting-based classifications of test instances by ensembles. In particular, we derive a method for assigning subpopulation-specific

* Corresponding author. Tel.: +1 501 686 5353; fax: +1 501 526 6729.
E-mail addresses: rlkodell@uams.edu, kodell.r@att.net (R.L. Kodell).

0933-3657/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.artmed.2013.04.008

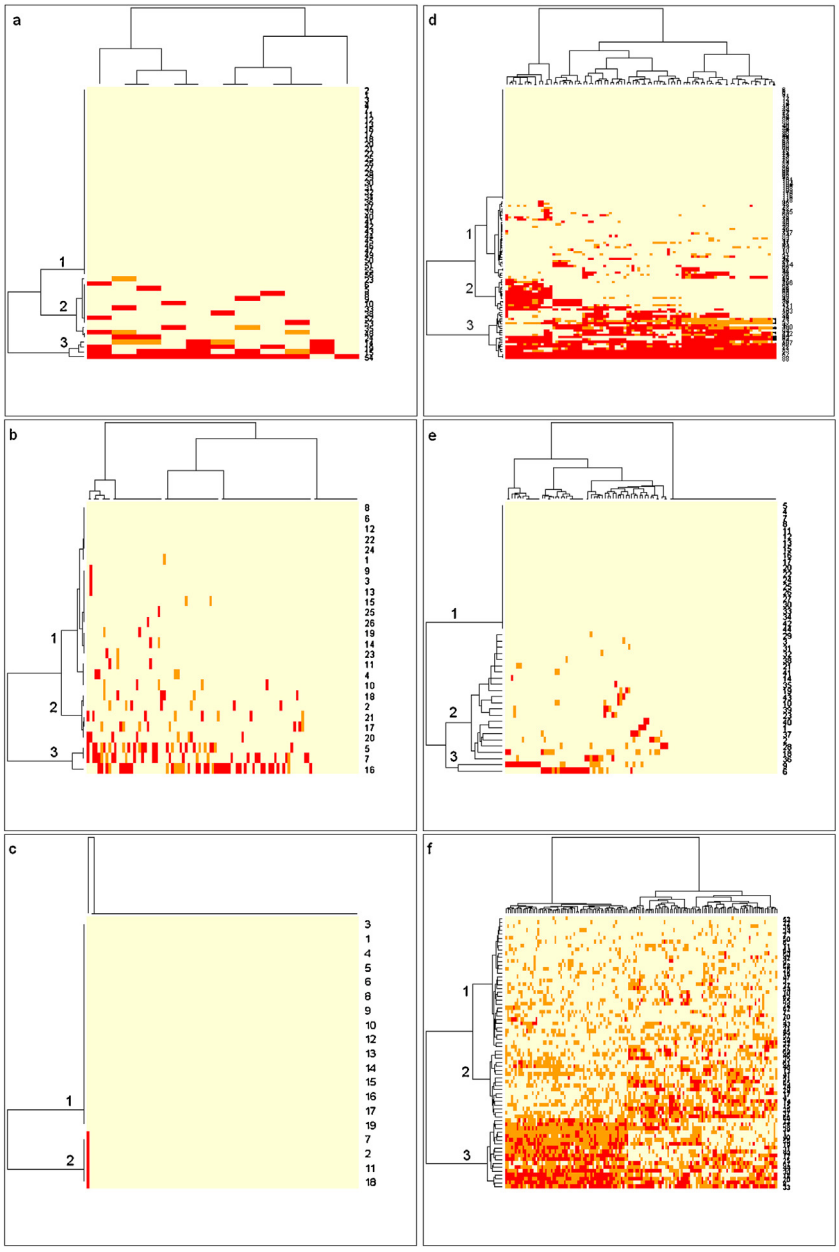

Fig. 1. Each heat map represents training samples from one representative block of one 10-fold CV, where panels (a) (colon cancer), (b) (non-classic glioma), and (c) (classic glioma) were produced using Method 1 and panels (d) (gene imprinting), (e) (soft tissue tumors), and (f) (breast cancer) were produced using Method 2. The colored rectangular spots on each heat map represent voting accuracies of an ensemble of classifiers used to classify the training set data. The classifiers in the ensemble are represented by the columns, and the samples in the training set are represented by the rows. A correct vote by an ensemble member is indicated by light yellow, an incorrect vote by red, and an abstention by orange. Dendrograms resulting from hierarchical clustering column-wise and row-wise are shown. The tree that represents the subpopulations of samples, clustered into either two or three subpopulations according to voting accuracies, is shown on the left side of the heat map and the tree that represents the ensemble classifiers is shown on the top of the heat map.

levels of confidence (highest, intermediate, or lowest) to classifications of unknown samples. We assume that the population of samples being classified is heterogeneous, but can be divided into subpopulations that are internally homogeneous. The heat maps in Fig. 1, to be described fully in Section 3, illustrate the approach. Each heat map resulted from a single run of the convex-hull, selective-voting algorithm of Kodell et al. [9] on an approximate 90% sample from one of six datasets in a 10-fold cross-validation. The colored rectangular spots on the heat maps represent voting accuracies of an ensemble of classifiers used to classify a training set of samples. The classifiers in an ensemble are represented by the columns, and the samples in the training set are represented by the rows, the latter having been clustered into either two or three subpopulations according to voting accuracies. A correct vote by an ensemble member is indicated by light yellow, an incorrect vote by red, and an abstention by orange. Dendrograms produced by hierarchical clustering row-wise and column-wise are shown. The tree that represents the subpopulations of samples is shown on the left side of the heat map and the tree that represents the ensemble of classifiers is shown on the top of the heat map. As the heat maps indicate, subpopulation 1 has the highest accuracy, subpopulation 2 has intermediate accuracy, and subpopulation 3 (if available) has the lowest accuracy.

With regard to practical application in a clinical setting, knowing the level of predictivity associated with a patient's algorithm-derived diagnosis, prognosis, or predicted response to treatment