Computer Assisted Radiology and Surgery

Reliable Evaluation of Performance Level for Computer-Aided Diagnostic Scheme[1]

Qiang Li, PhD

Rationale and Objectives. Computer-aided diagnostic (CAD) schemes have been developed for assisting radiologists in the detection of various lesions in medical images. The reliable evaluation of CAD schemes is an important task in the field of CAD research.

Materials and Methods. Many evaluation approaches have been proposed for evaluating the performance of various CAD schemes in the past. However, some important issues in the evaluation of CAD schemes have not been systematically analyzed. The first important issue is the analysis and comparison of various evaluation methods in terms of certain characteristics. The second includes the analysis of pitfalls in the incorrect use of various evaluation methods and the effective approaches to the reduction of the bias and variance caused by these pitfalls. We attempt to address the first important issue in details in this article by conducting Monte Carlo simulation experiments, and to discuss the second issue in the Discussion section.

Results. No single evaluation method is universally superior to the others; different situations of CAD applications require different evaluation methods, as recommended in this article. Bias and variance in the estimated performance levels caused by various pitfalls can be reduced considerably by the correct use of good evaluation methods.

Conclusions. This article would be useful to researchers in the field of CAD research for selecting appropriate evaluation methods and for improving the reliability of the estimated performance of their CAD schemes.

Key Words. Computer-aided diagnosis; CAD; resubstitution; leave-one-out; hold-out; cross validation; bias; variance; generalization performance.

Computer-aided diagnostic (CAD) schemes have been developed for detecting various lesions in many medical imaging modalities, including conventional radiography, computed tomography, magnetic resonance imaging, and ultrasound imaging. An important issue for CAD schemes is the reliable evaluation of their performance levels. In early publications in CAD research, resubstitution (RS) method was commonly used for the evaluation of CAD schemes (1,2). Because the performance level estimated by use of the RS method is optimistically biased, investigators in recent years have begun to employ more reliable evaluation methods such as the leave-one-out (LOO), cross-validation (CV), hold-out (HO), and bootstrap (BS) methods. Some investigators have investigated the effect of the sample size on the bias or variance of the estimated performance for classifier (3,4) or for CAD schemes (5–8). However, to our knowledge, no investigator has systematically analyzed and compared these common evaluation methods in terms of multiple important

Acad Radiol 2007; 14:985–991

[1] From the Department of Radiology, University of Chicago, S. Maryland Avenue, Chicago, IL 60637. Received December 9, 2006; accepted April 29, 2007. This work was supported by USPHS grants CA62625, CA64370, and CA113820. Q. Li is a consultant to Riverain Medical Group, Miamisburg, OH. CAD technologies developed at the Kurt Rossmann Laboratories for Radiologic Image Research, the University of Chicago, have been licensed to companies including R2 Technologies, Riverain Medical Group, Deus Technologies, Median Technology, Mitsubishi Space Software Co., General Electric Corporation, and Toshiba Corporation. It is the policy of the University of Chicago that investigators disclose publicly actual or potential significant financial interests that may appear to be affected by research activities. **Address correspondence to:** Q.L. e-mail: qiangli@uchicago.edu

doi:10.1016/j.acra.2007.04.015

characteristics, in particular, the bias and generalization performance. As a result, the selection of an evaluation method is generally empirical, and sometimes even arbitrary, in many studies for CAD development. We aim to provide such a comprehensive analysis and comparison to help researchers select appropriate evaluation methods for improving the quality and reliability of performance evaluation of their CAD schemes.

An important characteristic for various evaluation methods is the bias in the estimated performance level of a CAD scheme. Some evaluation methods such as LOO, CV, and HO are almost unbiased if they are used appropriately. Incorrect use of these methods, however, can lead to significant biases in the estimated performance levels. Therefore, we identified a number of typical pitfalls in the incorrect evaluation methods for CAD schemes, and conducted experiments to demonstrate quantitatively the extent of bias or variance caused by each of the pitfalls (9). In addition, for promoting and using a high standard for reliable evaluation of CAD schemes, we attempted to make recommendations, whenever possible, for overcoming these pitfalls. Because this part the work has been published previously (9). We will summarize in the Discussion section some important points regarding the reduction of bias and variance.

GENERALIZATION PERFORMANCE, ESTIMATED PERFORMANCE, AND BIAS

For a CAD scheme trained with a finite sample, there are two performance levels. The first is the generalization performance which measures how well the CAD scheme would achieve for a population of patients (all future new cases). After a CAD scheme is trained, its generalization performance is a fixed value. Unfortunately, it can not be determined directly in practice because investigators are unable to obtain the population of patients when the CAD scheme is designed. Therefore, one often has to determine an estimated performance by applying the trained CAD scheme to a small test sample. The estimated performance is a random value dependent on the small test sample used, and therefore, can be higher or lower than the generalization performance. If on average, the generalization and estimated performance levels are the same, the estimated performance is unbiased; otherwise, it is biased. Different evaluation methods (RS, LOO, CV, HO, and BS) have different ways to select the test sample, and consequently, can be either biased or unbiased. The bias and generalization performance are two important concepts for evaluation of CAD schemes.

ANALYSIS AND COMPARISON OF VARIOUS EVALUATION METHODS

Methods

This study was conducted based on datasets obtained from four Monte Carlo simulation experiments with four different sample sizes of 220 (20 abnormals + 200 normals), 250 (50 + 200), 300 (100 + 200), and 400 (200 + 200). The prevalences of abnormal objects were thus 9.1%, 20.0%, 33.3%, and 50.0%. Each Monte Carlo experiment consisted of 100 trials. In each trial, we randomly generated a sample of synthetic objects with one of these sample sizes. Each object had six features. For normal objects, the six features each obeyed a Gaussian distribution with a mean of 0 and respective standard deviations of {0.7, 0.8, 0.9, 1.0, 1.1, 1.2}. For abnormal objects, the six features each obeyed a Gaussian distribution with respective means of {1.0, 1.0, 0.9, 0.9, 0.8, 0.8} and respective standard deviations of {1.2, 1.1, 1.0, 0.9, 0.8, 0.7}. The first and second features, the third and fourth, and the fifth and sixth were correlated with a correlation coefficient of 0.7, 0.5, and 0.3, respectively. There was no correlation between other pairs of features.

The sample of simulation data in each trial is partitioned into a training set and a test set for training and testing a CAD scheme, respectively. In the RS method, the entire sample is used for both training and testing of a CAD scheme. In a k-fold CV method, the entire sample is first randomly partitioned into k disjoint subsets of nearly equal size, and then each of the k subsets is used as a test set for evaluation of a CAD scheme trained on the other (k-1) subsets. When the size of the subset is equal to 1, the CV is equivalent to the LOO method. In the HO method, the entire sample is partitioned into two subsets (not necessarily, but often, of equal size), one of which is used only for training of the CAD scheme, and the other only for testing of the trained CAD to obtain the estimated performance. In the BS method (10), a training dataset is generated by sampling with replacement n times from the n available cases (n = 220, 250, 300, 400) in the entire dataset. The entire dataset is also employed as a test dataset for evaluating the performance level of a CAD scheme. In each trial, the BS sampling was repeated 100 times, and the average performance level of the 100 iterations was reported.