Classifying tissue samples from measurements on cells with within-class tissue sample heterogeneity

JOSE-MIGUEL YAMAL*

Division of Biostatistics, The University of Texas School of Public Health, Houston, TX, 77030 USA
jose-miguel.yamal@uth.tmc.edu

MICHELE FOLLEN

Department of Obstetrics and Gynecology, Drexel University College of Medicine, Philadelphia, PA 19102, USA

MARTIAL GUILLAUD

Department of Cancer Imaging, British Columbia Cancer Agency, Vancouver, British Columbia, V5Z 1L3 Canada

DENNIS D. COX

Department of Statistics, Rice University, Houston, TX, 77005 USA

SUMMARY

We consider here the problem of classifying a macro-level object based on measurements of embedded (micro-level) observations within each object, for example, classifying a patient based on measurements on a collection of a random number of their cells. Classification problems with this hierarchical, nested structure have not received the same statistical understanding as the general classification problem. Some heuristic approaches have been developed and a few authors have proposed formal statistical models. We focus on the problem where heterogeneity exists between the macro-level objects within a class. We propose a model-based statistical methodology that models the log-odds of the macro-level object belonging to a class using a latent-class variable model to account for this heterogeneity. The latent classes are estimated by clustering the macro-level object density estimates. We apply this method to the detection of patients with cervical neoplasia based on quantitative cytology measurements on cells in a Papanicolaou smear. Quantitative cytology is much cheaper and potentially can take less time than the current standard of care. The results show that the automated quantitative cytology using the proposed method is roughly equivalent to clinical cytopathology and shows significant improvement over a statistical model that does not account for the heterogeneity of the data.

Keywords: Automating cervical neoplasia screening; Clustering densities; Cumulative log-odds; Functional data clustering; Macro-level classification; Quantitative cytology.

1. INTRODUCTION

To motivate the class of problems we address in this manuscript, we consider the following application. We wish to classify patients as having or not having disease based on measurements on their cells. We can

*To whom correspondence should be addressed.

consider each patient as being a population and the cells as members of the population, where our goal is to classify the populations based on the member measurements. We will refer to the unit that we want to classify (patients or populations) as the “macro-level” and the unit on which the measurements are made (cells or members) as the “micro-level,” which are nested within the macro-level units. Thus, we are faced with a problem of making macro-level classification based on micro-level measurements and an inherent mismatch in classification goals.

Although the methodology developed has many applications, we illustrate with the detection of cervical neoplasia (cancer and precancer) from a cervicovaginal smear. Cervical cancer is a preventable disease if detected early. Thus, screening is the key to fighting this disease. Cervical cancer screening aims to determine whether the patient should receive a biopsy. Using information at the cellular level, the goal is to classify the patient’s entire tissue sample rather than just the individual cells. The current screening process suffers from insufficient numbers of trained cytotechnologists (especially in developing countries) and a high cost, which may deter use. The need for a more objective and accurate method has spawned the field of quantitative cytology, which quantifies the characteristics of cells for automated classification of the tissue sample.

Quantitative cytology could prove useful in other organ sites. For example, quantitative measurements of cell nuclei from fine needle aspirates have been used for breast cancer diagnosis (Mangasarian *and others*, 1995). Additionally, dentists are considering diagnosis using brushing of cells or mouthwashes (Sciubba, 1999; Christian, 2002; Boutaga *and others*, 2007). The goal will be to automatically diagnose the patient based on the collection of their cells. Other potential applications of the proposed methodology include flow cytometric measurements, periodontal research, classification of schools based on student test scores (Texas Schools Project, 2009, <http://www.utdallas.edu/research/tsp-erc/index.html>), and classification of cities based on a sample of residents.

There is limited statistical understanding of the best approach for such a classification problem (Tsybrovskyy and Berghold, 2003a; Schulerud *and others*, 1998). There have been 3 general approaches (i) classify the micro-level observations and use an *ad hoc* approach to classify the macro-level (Assailly and Zhang, 1989; Grohs and Husain, 1994; Korbelik *and others*, 1993), (ii) extract macro-level features from the micro-level measurements (Thiran and Macq, 1996; Lugli *and others*, 2007; Bashashati and Brinkman, 2009), and (3) use a statistical model that accounts for the nested structure of the data (Cadez *and others*, 1999; Swartz *and others*, 2005). Tsybrovskyy and Berghold (1999) showed that ignoring the macro-level as unit of analysis leads to an increase in bias and falsely lowered *p*-values. A multilevel or hierarchical method is the best approach to this problem, yet methods for classification have not been well developed—see also Tsybrovskyy and Berghold (2003a). If the problem is simplified by selecting only one kind of statistical unit, using the patient as the unit of analysis is the only way to avoid biased *p*-values. Nested analysis of variance (ANOVA) models have been applied to macro-level classification problems (Bartels, 1982), but classical nested ANOVA has many limitations including the handling of multivariate data (Tsybrovskyy and Berghold, 2003b).

Using a statistical modeling approach, Swartz *and others* (2005) tackled the macro-level classification problem by obtaining estimates for the posterior log-odds of disease given the observed features at the micro-level and then combining those estimates to calculate the posterior log-odds of disease at the macro-level. With this cumulative log-odds (CLO) approach, they were able to use a high-dimensional feature set and avoid the *ad hoc* aspect of previous attempts. One of the assumptions in their method was independence and a common density of the feature vectors given the disease state. However, this assumption can be violated if between-patient variability causes heterogeneous populations. This can lead to different densities of the feature vector given the disease state.

Cadez *and others* (1999) used a hierarchical Bayesian model with 2 levels, first modeling the multivariate distribution at the micro-level with a mixture model, and then modeling the probability at the macro-level for each class. They did not address the case of heterogeneous data.

We give background on cervical neoplasia screening in Section 2, describe the CLO method in Section 3, propose an extension of that method for use with heterogeneous data in Section 4, present a simulation example comparing the proposed method latent-class cumulative log-odds (LACLO) to the CLO method in Section 5.1, present the results of using LACLO on real data Section 5.4, and discuss the results in Section 6.

2. CERVICAL NEOPLASIA SCREENING

Usage of Papanicolaou smears prevents cervical cancer by identifying premalignant lesions before they progress to invasive cancer. It uses a sample of cells from a patient's cervix that is examined under a microscope by a pathologist looking for dysplastic cells. Each sample is assigned a cytologic category, defined as either cancer, high-grade, low-grade, atypical squamous cells of undetermined significance (ASCUS), or negative. If the Papanicolaou smear is ASCUS or worse, a colposcopically directed biopsy is performed and the patient is treated if the histology result from the examination of the biopsy tissue is high-grade or worse.

It is important to have an easy means of conducting cervical cancer screening, especially in developing countries where resources and trained cytotechnologists (technologists who examine the cells microscopically for abnormalities) are scarce. [Grohs and Husain \(1994\)](#) proposed several approaches for controlling and potentially eradicating cervical cancer including the automation of any or all of the usual methods of performing Papanicolaou smears. The existing process suffers from much interobserver disagreement ([Stoler and Schiffman, 2001](#)). If we are to continue the use of the Papanicolaou smear as the primary method of screening, the need to increase the numbers of cytopathologists in the developing world is imperative. Nigeria's screening system illustrates this problem. I. F. Adewole, Professor of Obstetrics and Gynecology at the University of Ibadan, estimates that there are only about 12 trained cytopathologists for approximately 50 million women at risk, according to data from the Medical and Dental Council of Nigeria.

A way of automating the Papanicolaou smear process is through the use of quantitative cytology—classification of the sample from quantitative data obtained by measuring features on cell images. Details of the system used in this study are described in Section A.1 of the supplementary material available at *Biostatistics* online.

In a number of analyses, several features appear repeatedly as being predictive for dysplastic cells. One such feature, DNA index, correlates well with disease ([Yamal and others, 2004](#); [Guillaud and others, 2006](#); [Johnson and others, 1985](#); [Sun and others, 2005](#)), and is thus the focus in our analysis. DNA index is an approximate measure of the amount of nuclear DNA in the cell. The Feulgen stain used only stains the DNA in the nucleus and not the mitochondrial DNA. An increase in DNA index indicates either (i) that the normal cell is undergoing mitosis in which the amount of DNA chromatin will increase up to 2-fold followed by splitting into 2 cells or (ii) that the cell in question is actually an abnormal cell. A normal cell is expected to have a DNA index of around 1, and a cell that is about to split, around 2. If we observe a cell that has a DNA index much higher than 2, that cell is almost certainly aneuploid (having either fewer or more than the normal number of chromosomes), though not necessarily dysplastic—see [Griffiths \(1999\)](#) for details.

Details on the study design can be found in [Cantor and others \(2011\)](#). In our data, there were an average of around 2600 cells per patient (minimum 43, first quartile 1848, median 2717, third quartile 3662, and maximum 6569) and 56% of patients had all normal biopsies. Biopsies were taken from colposcopically identified abnormal sites, if the patient had colposcopic abnormalities, and 1 or 2 normal sites. A patient's histology was defined to be the worst histologic grade for all of their biopsies. The patient's histology, taken on all patients, was regarded as our gold standard.

For all analyses presented in this manuscript, we will dichotomize the disease status of the patient—class 1 is defined as a histological grade of high-grade disease or worse, and class 0 is defined as a

histological grade of low-grade disease or normal. We aim to predict the class (0 or 1) using this automated technology. We want to avoid overtreatment both because of possible adverse treatment effects and cost. Low-grade patients are followed in Canada and not treated. In the US, low-grade patients are not treated unless either the patients request treatment (because of great worry) or because the colposcopy is suspicious.

The comparison to the Papanicolaou smear accuracy was performed using receiver operating characteristic (ROC) analysis. A summary of the ROC curve across all thresholds is the area under the ROC curve (AUC); however, the AUC summarizes the information over areas of the ROC curve that are clinically not important, such as areas with low specificity. To focus on areas of clinical relevance, we calculate the area under a part of the ROC curve. The partial AUC is defined to be the area under the curve in a defined interval of either sensitivity or specificity, as discussed in [Dodd and Pepe \(2003\)](#). We used the interval between 80% and 100% specificity and used the partial AUC to compare the methods. We averaged the partial AUC over the interval by dividing by the length of the interval. We also compared methods by their sensitivity when the specificity was set at 90%.

In the process of selecting the best classifier, problems may result from overtraining—so that it works well on the data set that it was trained on but poorly on an independent data set. One solution is to use cross validation to obtain unbiased estimates of the classifier's performance. [Hastie and others \(2001\)](#) suggested that if there are sufficient data, the data can be divided into 3 sets: training, validation, and test sets. We chose to sample proportions of 40%, 30%, and 30% for the 3 data sets, respectively. The training set is used to estimate the parameters of a classifier. The validation set is used to obtain estimates of the trained classifier's performance using the parameters estimated from the training set and to select a classifier to apply to the test set. The test set is used to obtain an unbiased estimate of the chosen classifier's performance, as suggested by [Hastie and others \(2001\)](#).

To get an unbiased estimate of how well a method works on independent data, a threshold has to be chosen *a priori* and the results obtained using that threshold on the test set are then presented. The threshold of the score used to get 90% specificity was chosen from the validation set and applied to the final test set to get our final unbiased estimate of performance.

We compare all the methods to the accuracy of the Papanicolaou smear in the following way. The accuracy of the Papanicolaou smear is estimated to be 52% sensitivity at 90% specificity from our validation set (we use the validation data results since we will be comparing the classification methods on this data set). This is consistent with estimates by [Guillaud and others \(2006\)](#) which place it at 90% specificity and between 50–60% sensitivity.

3. CLO METHOD

[Swartz and others \(2005\)](#) obtained estimates of the posterior log-odds of disease given the observed features for individual cells and then combined the cell-level estimates to calculate the posterior log-odds of disease at the tissue-section level. Their CLO method is briefly described here. Let Y be the binary indicator of the true class of the population (patient), $Y \in \{0, 1\}$. Let $\mathbf{S}_i = x_{i1}, x_{i2}, \dots, x_{in_i}$ be the unordered n_i -tuple feature vectors measured on the n_i members (cells) from that population (patient). The n_i is assumed to be noninformative. The x can be univariate or multivariate features measured on each member. $\mathbf{S}_i$ is a vector of variable dimension of the patients' cell-level measurements. Note that the cells (members) are a sample from the large population of cervical cells in a patient. Let $\pi_0 = P(Y = 1)$ denote the prior probability (prevalence) of class 1 at the population level, and let $f(\mathbf{S}_i|Y = 1)$ denote the conditional distribution of the feature vector given the class Y . Let

$$\pi(\mathbf{S}_i) = \frac{\pi_0 f(\mathbf{S}_i|Y = 1)}{\pi_0 f(\mathbf{S}_i|Y = 1) + (1 - \pi_0) f(\mathbf{S}_i|Y = 0)}$$

denote the posterior probability of class 1 at the population level given the features from its members. A critical assumption here is that, conditional on the class, the feature vectors for the members from the same population are independent and identically distributed (i.i.d.) with density $f(x_i|Y)$ and, thus, can be written as

$$f(\mathbf{S}_i|Y = 1) = f(x_{i1}, \dots, x_{in_i}|Y = 1) = \prod_j f(x_{ij}|Y = 1), \quad (3.1)$$

and similarly given $Y = 0$. Using Bayes rule, the posterior log-odds of class 1 is

$$\text{logit}(\pi(\mathbf{S}_i)) = \text{logit}(\pi_0) + \sum_{j=1}^{n_i} \log \left(\frac{f(x_{ij}|Y = 1)}{f(x_{ij}|Y = 0)} \right). \quad (3.2)$$

The CLO score can be estimated using the ratio of the densities $\frac{f(x_{ij}|Y=1)}{f(x_{ij}|Y=0)}$ in (3.2).

Although there are superficial similarities between the CLO and naïve Bayes classifiers, there are important differences that distinguish the class of problems and assumptions. The main assumption in naïve Bayes is independence of the components of feature vector given the class. In contrast, CLO assumes independence of a random number of micro-level feature vectors within the macro-level unit given the class. In naïve Bayes, the unit of measurement and unit of classification are the same, which is not true in the CLO setting.

The density functions are approximated using a kernel density estimate, $\hat{f}_h(x) = \frac{1}{Nh} \sum_{i=1}^N K\left(\frac{x-x_i}{h}\right)$, where K is some kernel, h is the bandwidth, and $x_1, x_2, \dots, x_n$ are i.i.d. observations (see, e.g., [Mclachlan, 1992](#)). We used the biweight kernel, defined by $K_b(x) = \frac{15}{16}(1-x^2)^2 I_{|x| \leq 1}$. The bandwidth or smoothing parameter h affects the CLO method classification accuracy. The bandwidth is chosen by maximizing the classification accuracy in a grid of values.

The likelihood ratio is calculated by dividing 2 kernel density estimates. When the denominator of this ratio is close to 0, the value of the likelihood ratio is unstable. By mixing the density estimate with a uniform distribution before dividing the 2 densities, the likelihood ratio can be stabilized. Let ψ be the mixing parameter for the uniform distribution where $0 \leq \psi \leq 1$. The new mixed likelihood ratio, denoted lr_ψ , is given by

$$\text{lr}_\psi = \frac{\psi \times g(x) + (1 - \psi) \times f(x|Y = 1)}{\psi \times g(x) + (1 - \psi) \times f(x|Y = 0)},$$

where $g(x)$ is the uniform distribution. For the support of the uniform distribution, we used either the sample range of the training set or the interval $[0, 5]$ since it is unlikely for the DNA index of a cell to be higher than 5 unless it is an image of overlapping cells or debris. There is a trade-off between stabilization of the likelihood ratios and how well the CLO method will be able to discriminate between the classes.

4. EXTENDING THE CLO METHOD TO HETEROGENEOUS DATA

Our data did not appear to be homogeneous across patients within the same disease class, which violated the assumption in the CLO method that, conditional on the class, the feature values are i.i.d. with a common distribution across all populations of that class. For the feature DNA index, we noted that some of the density estimates were unimodal and others had additional modes, usually around 2. Figure 1 shows the density estimates for 2 diseased patients, both with a mode around 1 for the noncycling normal cells, but the second patient (dotted line) has 2 additional modes that includes cycling cells and other cells with a DNA index of around 4.5. The density estimate for all cells (pooled across all patients from the diseased class) suggests that most diseased patients did not have as many cells with DNA index greater than 2 as patient 2 had. The figure suggests the data were heterogeneous. We performed a permutation test to

Fig. 1. Density of pooled cells from all diseased patients in the training set and from two individual diseased patients. The dashed-dotted (blue online) line sample patient had an additional mode around DNA Index 4.5 whereas the dashed (red online) line sample patient did not.

test for heterogeneity in our data ($p < 0.001$). Details can be found in Section B.1 in the supplementary material available at *Biostatistics* online.

We model the between-patient heterogeneity by assuming the existence of an unobserved latent variable U , and assume that the features are i.i.d. given the class and the latent variable. We claim that the cell-level measurements may be regarded as i.i.d. since the cells are well mixed in the liquid suspension before being placed on the slide, there is no information in the ordering of the measurements nor in the location of the cells on the slide. Thus, the cell-level measurements may be modeled as realizations of exchangeable random variables, and according to the generalized version of de Finetti's theorem (Hewitt and Savage, 1955), are conditionally i.i.d. given the patient. We further model patient to patient variability with the latent variable. Additionally, we use a discrete latent variable to approximate continuous or more complicated latent variables.

Denote the probability mass function of the latent variable given the class by $f(u|Y)$. Given our assumption that conditional on both U and Y , the micro-level observations within a macro-level unit were i.i.d., the density for the feature vectors given the disease state are given by

$$\begin{aligned} f(x_{i1}, \dots, x_{in_i}|Y) &= \sum_u f(u|Y) f(x_{i1}, \dots, x_{in_i}|u, Y) \\ &= \sum_u f(u|Y) \prod_j f(x_{ij}|u, Y). \end{aligned}$$

It follows that the log-odds of having disease given our feature vector is

$$\begin{aligned} \text{logit}(\pi(\mathbf{S}_i)) &= \text{logit}\left(\frac{\pi_0 f(\mathbf{S}_i|Y=1)}{\pi_0 f(\mathbf{S}_i|Y=1) + (1-\pi_0) f(\mathbf{S}_i|Y=0)}\right) \\ &= \text{logit}(\pi_0) + \log\left(\frac{\sum_u f(u|Y=1) \prod_j f(x_{ij}|u, Y=1)}{\sum_v f(v|Y=0) \prod_j f(x_{ij}|v, Y=0)}\right). \end{aligned}$$

We refer to this method as the LACLO method.

When there are many micro-level observations, the product $\prod_j f(x_{ij}|u, Y=1)$ can easily give exponential overflow or underflow, so we rewrite this equation to obtain more stable results. See Section B.2 of the supplementary material available at *Biostatistics* online.

We use a functional clustering method to estimate the densities $f(x|u, Y)$ and probabilities $f(u|Y)$, where u ranges over a finite set. This is implemented by computing a density estimate for each patient and then applying a clustering algorithm to the densities. Since nonparametric density estimates are functional data, this is functional data clustering. Most clustering algorithms require a similarity measure between the units being clustered. We investigated several such measures and settled on Euclidean metric on vectors obtained by evaluating the densities on a grid, which approximates the L_2 distance between the densities. We also used the K -means algorithm (MacQueen, 1966) although other algorithms, such as hierarchical clustering could be used.

The univariate density estimate of each patient's cells' DNA index was placed on a common, evenly spaced, 500-point grid either in the range of the training data or a range from 0 to 5. The density vectors were then clustered using K -means clustering, with $K = K_1$ within the disease group and $K = K_0$ within the nondiseased group. Once we had the clusters of patients, the proportion of patients put into a cluster was our estimate for the probability $f(u|Y)$ of belonging in that cluster. The cells from all the patients clustered together were pooled to make a density estimate $f(x|u, Y)$ for the cluster corresponding to u . These densities $f(x|u, Y)$ (for given u and Y) are then used to evaluate new observations in (4.1). The idea was that the distributions within each cluster were now approximately homogeneous. For a new patient, each cell's density was calculated for each cluster for both disease states and the weighted sum was computed, weighted by the probability of belonging in that cluster as in (4.1).

Several factors can influence the classification performance, including the choices for ψ (the mixing parameter of the uniform distribution), the bandwidth h for the kernel densities, the number of clusters in each disease state, transformations of the variable (which amplify subtle differences which in turn can be detected by applying K -means clustering), and limiting the range of the variable (denoted as *rng*). To determine which choices of the parameters to use, we tried several choices of the parameters and saw how well the classifier worked on independent data. The ψ parameter was varied as 10 evenly spaced points between 0 and 0.02. h was varied as 10 evenly spaced points between 0.05 and 0.12 (values near the bandwidth chosen by Silverman's rule of thumb, Silverman, 1986, for the patient with the number of cells). The number of clusters in each disease state was set as $1 \leq K_m \leq 10$ for $m = 1, 2$. As shown in the next section, this range of K_m 's was suitable for our data. Transformations can help to amplify any subtle differences, which in turn can be detected by applying K -means clustering. The transformations done on DNA index included the log transformation, square-root transformation, and the identity transformation.

rng was set as either the whole range of the training data DNA indices or the range with upper limit truncated to 5. With so many factors, it was not possible to compute all combinations. We restricted our search to a 2-step optimization. At each step, the “optimal” parameters were chosen as the ones that had the largest partial AUC in the validation set. We first looked at every combination of the 3 transforms, all values of (K_1, K_0) , and both $rng = [0, 5]$ and $rng =$ the range of the training data, with given h and ψ . We then fixed the parameters K_1, K_0, rng and the transformation and optimized on h and ψ .

The R statistical package was used for the analysis (R Development Core Team, 2010). Code is available on request.

5. RESULTS

To get a better sense of how heterogeneous data can affect the CLO method, we considered the following simulation. We constructed training, validation, and test sets with the same number of diseased and nondiseased patients as in our data (see Table 1 in the supplementary material available at *Biostatistics* online) and with 1000 cells per patient. The patient distributions were modeled using a mixture of normal distributions with 3 components with modes at $\mu = (1, 2, 3)$. The mode at 1 models the normal cells, the mode at 2 models the cycling cells, and the mode at 3 models aneuploid cells. The variances were set as $\sigma_1^2 = (0.1, 0.1, 0.4)$ for class 1 and $\sigma_0^2 = (0.1, 0.1, 0.9)$ for class 0 and the weights were generated using the stick-breaking approach proposed by Sethuraman (1994), using independent beta random variables. We chose the parameters for the beta distributions to emulate our data, giving more weight to the component with mode around one, and having a larger variance for the class 0 third component than the class 1 third component. The weights of the second and third component of the mixture of normal distributions were drawn from a $beta(10, 40)$ for the class 1 data and $beta(1, 4)$ for the class 0 data. The LACLO and the CLO methods used the training data to build the model and were optimized by choosing the parameters that resulted in the largest partial AUC in the validation set. The optimal parameters for the LACLO method were $K_0 = 7$, $K_1 = 10$, $\psi = 0$, and $h = 0.089$. The optimal bandwidth for the CLO method was 0.01 times the Silverman’s rule of thumb for the density of each class. The model was then trained on the combined training and validation sets using these parameters and applied to the test set to get an unbiased estimate of the performance of the methods. Figure 2 shows the ROC curves of the LACLO and CLO methods applied to the test data. The LACLO method outperformed the CLO method. The difference in the partial AUC between the LACLO and CLO method on the test set was statistically significant using 10,000 bootstraps ($p < 0.001$, see Section B.3 of the supplementary material available at *Biostatistics* online.).

The CLO method was applied to the real data from our study using the DNA index as a predictor. Density estimates were computed for $f(x|Y = y)$, $y = 0, 1$ from the training data. The prior prevalence of disease was 14%, estimated from the training data. The uniform distribution mixing parameter ψ was varied from 0 to 0.02 (10 evenly spaced points). The bandwidth was varied by multiplying the Silverman’s

Table 1. *Summary of estimated sensitivities for detecting High Grade Squamous Intraepithelial Lesion or worse versus Low Grade Squamous Intraepithelial Lesion or better on Test Set using the threshold that gave 90% specificity in the validation set*

Method	Sensitivity (95% CI)	Specificity (95% CI)	AUC (95% CI)	Partial AUC (0.8-1.0 specificity)
CLO	47% (37%, 57%)	90% (87%, 93%)	0.71 (0.65, 0.77)	0.41
LACLO	52% (42%, 62%)	92% (89%, 94%)	0.76 (0.7, 0.82)	0.47

Fig. 2. ROC Plot of Latent-class CLO Method versus CLO Method in a simulation example.

bandwidth by a constant, ranging from 0.2 to 20 (11 values), for the normal and diseased patients separately (Silverman, 1986). We then computed the posterior log-odds for each ψ , bandwidth adjustment, and range choice.

Figure 3 shows the ROC curve of CLO using the range [0,5], $\psi = 0$, and bandwidth factor 10. The best classifier we were able to find with CLO had an AUC of 0.71, a partial AUC of 0.41, and 47% sensitivity for 90% specificity.

In the first step of the optimization process, we chose between 1 and 10 groups in each disease state to calculate the LACLO score. We set the value of the bandwidth parameter, h , to be equal to 0.086 using Silverman's rule of thumb for the patient with the largest number of cells. The uniform distribution mixing parameter ψ was set to be equal to 0.01. The range rng was set to be either the interval [0,5] or the whole sample range of the training data. Using the whole range of DNA index showed an improvement over the range [0,5]. It is not surprising that the LACLO method worked well with the whole range of DNA index since LACLO can assign a cluster to the cases that had cells with a DNA index higher than 5.

Fig. 3. ROC curve of CLO and LACLO methods using DNA Index on validation data.

The best partial AUC was 0.53 using 4 clusters in the nondiseased group and 5 clusters in the diseased group, the whole range, and the identity transformation.

In the second step of the optimization process, we fixed $K_0 = 4$, $K_1 = 5$, rng = the training data sample range, used the identity transformation, and optimized over the ψ and h parameters. The bandwidth h ranged from 0.05 to 0.12 and the amount to mix with a uniform distribution ψ ranged from 0 to 0.02. We maximized the partial AUC in the validation data using parameters $h = 0.081$ and $\psi = 0$ (sensitivity of 65% at 90% specificity).

At 90% specificity, this represented a 18% increase in sensitivity compared to the CLO method. The ROC curve is presented in Figure 3. Note that the LACLO method with one cluster in each disease state is equivalent to CLO.

Figure 4 shows the density estimates for the 4 clusters used in the non-diseased class and the 5 clusters in the diseased class. The graphs are truncated between DNA index 1.25 and 5 since the density estimates were most different from each other in this range. There was much more variation among the density

Fig. 4. Estimated density (on limited range) of clusters in the non-diseased and diseased groups.

estimates in the diseased group than among those in the nondiseased group. The first cluster represents the patients who apparently had more cycling cells (between DNA indices 1 and 2), possibly due to the Hayflick limit where a high grade squamous intraepithelial lesion is likely to harbor a large aneuploid population around 1.7 as described by [Rasnick \(2000\)](#). The second disease cluster had the greatest proportion of cells with DNA index around 2 as well as many cells with a DNA index of around 4. The diseased and nondiseased representative density estimates were very different in appearance.

Using the threshold of the posterior log-odds score used to get 90% specificity from the validation set, we applied the methods to the final test set to get our final unbiased estimate of its performance, shown in Table 1. To compare 2 ROC curves in the test set, we constructed bootstrap confidence intervals of the difference between the partial AUCs ([Dodd and Pepe, 2003](#)). The LACLO method's sensitivity decreased from 66% to 52%, yet the specificity increased slightly. On the test set, the LACLO ROC curve was very similar to that of clinical cytology. The LACLO method had an AUC of 0.76 versus 0.71 from CLO ($p = 0.075$). The partial AUC was statistically significantly higher for the LACLO method (0.47 vs. 0.41, $p = 0.01$). The ROC curves of the CLO and LACLO methods are shown in Figure 5.

Fig. 5. ROC Plot of CLO and Latent-class CLO Methods on test data using variable DNA Index. The LACLO method used 4 clusters in the normal patients and 5 clusters in the diseased patients, $\psi = 0$, & $h = 0.081$.

6. DISCUSSION

Our primary goal was to create a general statistical framework for macro-level classification, especially for heterogeneous data with between-macro-level variation. The LACLO method showed an increase in classification performance over the CLO method in simulations and our application.

A potential drawback to the LACLO method is the number of tuning parameters: the number of clusters in each class, the bandwidth of the kernel densities, and the mixing parameter with the uniform distribution. Care has to be taken to not overtrain the algorithm. We used the strategy of separating the data into training, validation, and test sets. Other cross-validation methods could be employed. We observed that there was a wide range of values of the parameters where we got similar performance to that using the optimized parameters, suggesting that the results were not very sensitive to the exact choice of parameters. The extension to high-dimensional data is a challenge. Preliminary investigation of the use of an additional

variable with DNA index has produced promising results, but further research will need to be done for higher dimensional data.

From a clinical standpoint, we have developed an algorithm that performs at least as well as conventional cytology in a mix of diagnostic and screening patients. A comparison in the subset of screening patients was not possible due to the low prevalence of disease. Nonetheless, our results are important because of the obvious immediate benefits for automated screening for cervical cancer.

SUPPLEMENTARY MATERIAL

Supplementary material is available at <http://biostatistics.oxfordjournals.org>.

ACKNOWLEDGMENTS

We thank E. Neely Atkinson and David W. Scott for their input to this analysis and Brian Crain for editing this document. *Conflict of Interest*: None declared.

FUNDING

National Cancer Institute (P01-CA-82710-09); Dennis Cox's research was supported in part by National Science Foundation (NSF) (DMS-05-05584).

REFERENCES

- ASSAILLY, J. AND ZHANG, J. (1989). Contribution of quantitative cytology and ploidy in the diagnosis and monitoring of tumors of the bladder. Study of 52 cases using the Samba analyser. *Journal d'urologie* **95**, 3–9.
- BARTELS, P. (1982). Numerical evaluation of cytologic data XI. Nested designs in multivariate analysis of variance. *Analytical and Quantitative Cytology* **4**, 81–89.
- BASHASHATI, A. AND BRINKMAN, R. (2009). A survey of flow cytometry data analysis methods. *Advances in Bioinformatics* **2009**, 1–19.
- BOUTAGA, K., SAVELKOU, P., WINKEL, E. AND VAN WINKELHOFF, A. (2007). Comparison of subgingival bacterial sampling with oral lavage for detection and quantification of periodontal pathogens by real-time polymerase chain reaction. *Journal of Periodontology* **78**, 79–86.
- CADEZ, I., MCLAREN, C., SMYTH, P. AND MCLACHLAN, G. (1999). Hierarchical models for screening iron deficiency anemia. In: Bratko, I. and Dzeroski, S. (editors), *Proceedings of the 16th International Conference on Machine Learning*. San Francisco, CA: Morgan Kaufmann, pp. 77–86.
- CANTOR, S., YAMAL, J.-M., GUILLAUD, M., COX, D., ATKINSON, E., BENEDET, J., MILLER, D., EHLEN, T., MATISIC, J., VAN NIEKERK, D. and others (2011). Accuracy of optical spectroscopy for the detection of cervical intraepithelial neoplasia: testing a device as an adjunct to colposcopy. *International Journal of Cancer* **128**, 1151–1168.
- CHRISTIAN, D. (2002). Computer-assisted analysis of oral brush biopsies at an oral cancer screening program. *Journal of the American Dental Association* **133**, 357–362.
- DODD, L. AND PEPE, M. (2003). Partial AUC estimation and regression. *Biometrics* **59**, 614–623.
- GRIFFITHS, A. (1999). *An Introduction to Genetic Analysis*. San Francisco, CA: W.H. Freeman.
- GROHS, O. AND HUSAIN, O. (1994). *Automated Cervical Cancer Screening*. New York: Igaku-Shoin.
- GUILLAUD, M., BENEDET, J., CANTOR, S., STAERKEL, G., FOLLEN, M. AND MACAULAY, C. (2006). DNA ploidy compared with human papilloma virus testing (hybrid capture II) and conventional cervical cytology as

- a primary screening test for cervical high-grade lesions and cancer in 1555 patients with biopsy confirmation. *Cancer* **107**, 309–318.
- HASTIE, T., TIBSHIRANI, R. AND FRIEDMAN, J. (2001). *The Elements of Statistical Learning*. New York: Springer.
- HEWITT, E. AND SAVAGE, L. J. (1955). Symmetric measures on Cartesian products. *Transactions of the American Mathematical Society* **80**, 470–501.
- JOHNSON, T., WILLIAMSON, K., CRAMER, M. AND PETERS, L. (1985). Flow cytometric analysis of head and neck carcinoma DNA index and S-fraction from paraffin-embedded sections: comparison with malignancy grading. *Cytometry* **6**, 461–470.
- KORBELIK, J., MATISIC, J., MACAULAY, C., GARNER, D. AND PALCIC, B. (1993). Discrimination of mildly dyskaryotic squamous cervical cells with or without Koilocytosis using nuclear features measured by quantitative histology and automated quantitative cytology. *Acta Cytologica* **37**, 602.
- LUGLI, E., PINTI, M., NASI, M., TROIANO, L., FERRARESI, R., MUSSI, C., SALVIOLI, G., PATSEKIN, V., PAUL ROBINSON, J., DURANTE, C. and others (2007). Subject classification obtained by cluster analysis and principal component analysis applied to flow cytometric data. *Cytometry Part A* **71A**, 334–344.
- MACQUEEN, J. (1966). Some methods for classification and analysis of multivariate observations. In: Cam, L. L. and Neyman, J. (editors), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, CA: University of California Press, pp. 281–297.
- MANGASARIAN, O., STREET, W. AND WOLBERG, W. (1995). Breast cancer diagnosis and prognosis via linear programming. *Operations Research* **43**, 570–577.
- MCLACHLAN, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. New York: Wiley-Interscience.
- R DEVELOPMENT CORE TEAM (2010). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <http://www.R-project.org>.
- RASNICK, D. (2000). Auto-catalysed progression of aneuploidy explains the Hayflick limit of cultured cells, carcinogen-induced tumours in mice, and the age distribution of human cancer. *Biochemical Journal* **348**, 497–506.
- SCHULERUD, H., KRISTENSEN, G., LIESTOL, K., VLATKOVIC, L., REITH, A., ALBREGTSEN, F. AND DANIELSEN, H. (1998). A review of caveats in statistical nuclear image analysis. *Analytical Cellular Pathology* **16**, 63–82.
- SCIUBBA, J. (1999). Improving detection of precancerous and cancerous oral lesions. *Journal of the American Dental Association* **130**, 1445–1457.
- SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4**, 639–650.
- SILVERMAN, B. (1986). *Density Estimation*. London: Chapman and Hall.
- STOLER, M. AND SCHIFFMAN, M. (2001). Interobserver reproducibility of cervical cytology and histological interpretations. *Journal of the American Medical Association* **11**, 1500–1505.
- SUN, X., WANG, J., GARNER, D. AND PALCIC, B. (2005). Detection of cervical cancer and high grade neoplastic lesions by a combination of liquid-based sampling preparation and DNA measurements using automated image cytometry. *Cellular Oncology* **27**, 33–41.
- SWARTZ, R., WEST, L., BOIKO, I., MALPICA, A., GUILLAUD, M., MACAULAY, C., FOLLEN, M. AND COX, D. (2005). Classification using the cumulative log-odds in the quantitative pathological diagnosis of adenocarcinoma of the cervix. *Gynecologic Oncology* **99**(Suppl 1), S24–S31.
- THIRAN, J. AND MACQ, B. (1996). Morphological feature extraction for the classification of digital images of cancerous tissues. *IEEE Transactions on Biomedical Engineering* **43**, 1011–1020.

- TSYBROVSKYY, O. AND BERGHOLD, A. (1999). Primary unit for statistical analysis in morphometry: patient or cell? *Analytical Cellular Pathology* **18**, 191–202.
- TSYBROVSKYY, O. AND BERGHOLD, A. (2003a). Application of multilevel models to morphometric data. Part 1. Linear models and hypothesis testing. *Analytical Cellular Pathology* **25**, 173–185.
- TSYBROVSKYY, O. AND BERGHOLD, A. (2003b). Application of multilevel models to morphometric data. Part 2. Correlations. *Analytical Cellular Pathology* **25**, 187–191.
- YAMAL, J.-M., COX, D., HITTELMAN, W., BOIKO, I., MALPICA, A., GUILLAUD, M., MACAULAY, C., FOLLEN, M. AND VLASTOS, A. (2004). Quantitative histopathology and chromosome 9 polysomy in a clinical trial of 4-HPR. *Gynecologic Oncology* **94**, 296–306.

[Received February 22, 2010; revised April 15, 2011; accepted for publication April 19, 2011]