PREDICTING CLASSIFIER PERFORMANCE WITH A SMALL TRAINING SET: APPLICATIONS TO COMPUTER-AIDED DIAGNOSIS AND PROGNOSIS

Ajay Basavanhally, Scott Doyle, Anant Madabhushi *

Rutgers, The State University of New Jersey
Department of Biomedical Engineering
Piscataway, NJ, USA

ABSTRACT

Selection of an appropriate classifier for computer-aided diagnosis (CAD) applications has typically been an ad hoc process. It is difficult to know *a priori* which classifier will yield high accuracies for a specific application, especially when well-annotated data for classifier training is scarce. In this study, we utilize an inverse power-law model of statistical learning to predict classifier performance when only limited amounts of annotated training data is available. The objectives of this study are to (a) predict classifier error in the context of different CAD problems when larger data cohorts become available, and (b) compare classifier performance and trends (both at the sample/patient level and at the pixel level) as additional data is accrued (such as in a clinical trial). In this paper we utilize a power law model to evaluate and compare various classifiers (Support Vector Machine (SVM), C4.5 decision tree, k-nearest neighbor) for four distinct CAD problems. The first two datasets deal with sample/patient-level classification for distinguishing between (1) high from low grade breast cancers and (2) high from low levels of lymphocytic infiltration in breast cancer specimens. The other two datasets are pixel-level classification problems for discriminating cancerous and non-cancerous regions on prostate (3) MRI and (4) histopathology. Our empirical results suggest that, given sufficient training data, SVMs tend to be the best classifiers. This was true for datasets (1), (2), and (3), while the C4.5 decision tree was the best classifier for dataset (4). Our results also suggest that results of classifier comparison made on small data cohorts should not be generalized as holding true when large amounts of data become available.

1. INTRODUCTION

Most computer-aided diagnosis (CAD) systems typically involve a supervised classifier that needs to be trained on a set of annotated examples. These training samples are usually provided by a medical expert, who labels the samples according to their class. Unfortunately, in many biomedical applications, training data is not abundant either due to the cost involved in obtaining expert annotations or because of overall data scarcity. Therefore, classifier choices are often made based on classification results from a small number of training samples, which relies on the assumption that the selected classifier will exhibit the same performance when exposed to larger datasets. We aim to demonstrate in this study that classifier trends observed on small cohorts may not necessarily hold true when larger amounts of data become available. If evaluating a CAD classifier in the context of a clinical trial (where data becomes available sequentially), one needs to be wary of choosing a classifier based on performance on limited training data. The optimal classifier could change as more data becomes available mid-way through the trial, at which point one is saddled with the initial classifier. Furthermore, the selection of an optimal classifier for a specific dataset usually requires large amounts of annotated training data [1] since the error rate of a supervised classifier tends to decrease as training set size increases [2].

The objectives of this work are to address certain key issues that arise early in the development of a CAD system. These include:

1. Predicting error rate associated with a classifier assuming that a larger data cohort will become available in the future, and
2. Comparing performance of classifiers, at both the sample- and pixel-levels, for large data cohorts based on accuracy predictions from smaller, limited cohorts.

The specific translational implications of this study will be relevant in (a) better design of clinical trials (especially pertaining to CAD systems), and (b) enabling a power analysis of classifiers operating on the pixel level (as opposed to patient/sample level), which cannot be currently done via standard sample power calculators.

The methodology employed in this paper is based on the work in [3] where an inverse power law was used to model the change in classification accuracy for microarray data as a function of training set size. In our approach, a subsampling procedure is used to create multiple training sets at various sizes. The error rates resulting from evaluation of these training samples is used to determine the three parameters of the power-law model (rate of learning, decay rate, and Bayes error) that characterize the behavior of the error rate as a function of training set size. By calculating these parameters for various classifiers, we can intelligently choose the classifier that will yield the optimal accuracy for large training sets. This approach will also allow us to determine whether conclusions derived from classifier comparison studies involving small datasets, are valid in circumstances where larger data cohorts become available. In this work, we apply this method to predict the performance of four classifiers: Support Vector Machine (SVM) using radial basis function, SVM using linear kernel, k-nearest neighbor, and C4.5 decision tree on four different CAD tasks (see Table 1 and Section 3).

2. EXPERIMENTAL DESIGN

2.1. Overview of Prediction Methodology

The general procedure for estimating performance comprises the following steps: (1) Generate training sets that will be used to calculate

*This work was made possible via grants from the Wallace H. Coulter Foundation, New Jersey Commission on Cancer Research, National Cancer Institute (R01CA136535-01, R21CA127186-01, R03CA128081-01), the Cancer Institute of New Jersey, and Bioimagene Inc.

Notation	Description	Samples (ω_1 / ω_2)	Features for Classifier
$\mathcal{D}_1$	Breast: Cancer Grade	34 (16 / 18)	Spatial arrangement of cancer nuclei
$\mathcal{D}_2$	Breast: Extent of Lymphocytic Infiltration	41 (21 / 20)	Spatial arrangement of lymphocyte nuclei
$\mathcal{D}_3$	Prostate: Cancer Detection on DCE-MRI	18 (450 / 450)	Intensity and texture
$\mathcal{D}_4$	Prostate: Cancer Detection on Histopathology	60 (30 / 30)	Intensity and texture

Table 1. List of the breast cancer and prostate cancer datasets used in this study. Although the number of samples for $\mathcal{D}_3$ is listed in terms of MRI slices, it is important to note that classification is performed at the pixel level represented by the values listed under ω_1 and ω_2.

the model parameters; (2) Ensure that the number of training samples in each set is statistically significant; (3) Calculate the parameters of the power law model (Equation 3) for each classifier; (4) Plot the model at increasing training set sizes to determine the optimal classifier for the dataset, as well as the expected performance for a large training cohort; (5) Verify the power law estimate empirically using larger training sets.

2.2. Subsampling Procedure to Create Initial Training Sets

To accurately extrapolate classifier performance for large training set sizes, we must first measure classification accuracy for a number of smaller training set sizes. However, classification accuracy may vary greatly for small training set sizes and lead to a miscalculation of the model parameters. Therefore, we must ensure that our initial training set is large enough to significantly estimate the model parameters. This is done by comparing the accuracy of the classifier trained with real data against a classifier trained with data that has randomly assigned labels. If the difference between the error rates is statistically significant, we can confidently calculate the model parameters.

We first divide a dataset $\mathcal{D}$ containing ℓ samples into a training pool $\mathcal{N}$ and a testing set $\mathcal{T}$, where $\mathcal{N}$ contains three-fourths the samples in $\mathcal{D}$ and $\mathcal{T}$ contains the remaining one-fourth. We denote the class label of a sample $x \in \mathcal{D}$ by ω_1 (representing the diseased or "high" class) or ω_2 (representing the normal or "low" class).

A set of training set sizes $\mathbf{N} = \{n_1, n_2, \ldots, n_6\}$ are selected, where each training set size n falls within the interval $[1, |\mathcal{N}|]$ and $|\cdot|$ denotes set cardinality. For each $n \in \mathbf{N}$, the training pool $\mathcal{N}$ is sampled randomly T_1 times. Each of these random subsets of the training are denoted by $\mathcal{S}_{n,i}$, where $n \in \mathbf{N}$ and $i \in \{1, 2, \ldots, T_1\}$. Thus, there are a total of $6 \times T_1$ training sets generated by this subsampling procedure. Each $\mathcal{S}_{n,i}$ is used to train a classifier $\Omega_{n,i}$, which is evaluated on the entire testing set $\mathcal{T}$. The error rate for classifier $\Omega_{n,i}$ is denoted $e_{n,i}$, and the mean error rate for each n is denoted:

$$\bar{e}_n = \frac{1}{T_1} \sum_i e_{n,i}. \tag{1}$$

2.3. Estimation of Statistical Significance

To verify the significance of the mean error rate $\bar{e}_n$, we compare the performance of training set $\mathcal{S}_{n,i}$ against the performance of randomly labeled training data, thus ensuring that our training sets $\mathcal{S}_{n,i}$ yield a statistically sound basis for calculating the power law parameters. The motivation for this approach is that a randomly trained classifier corresponds to the "intrinsic error" of the classifier. This procedure is summarized by the following steps.

1. For each $\mathcal{S}_{n,i}$ we generate T_2 random training sets $\mathcal{S}^{\text{ran}}_{n,i,j}$, for $j \in \{1, 2, \ldots, T_2\}$, where the label of each sample has been randomly assigned.
2. This generates an additional $6 \times T_1 \times T_2$ randomized training sets, each of which is used to train a classifier $\Omega^{\text{ran}}_{n,i,j}$ and produce an error rate $e^{\text{ran}}_{n,i,j}$.
3. For each n, we calculate the following relation between the randomly- and correctly-labeled classifiers:

$$P_n = \frac{1}{T_1 \times T_2} \sum_i^{T_1} \sum_j^{T_2} \theta(\bar{e}_n - e^{\text{ran}}_{n,i,j}), \tag{2}$$

where $\theta(z) = 1$ if $z \geq 0$ and 0 otherwise.

If the randomly-trained classifiers consistently yield lower error rates than the correctly-trained classifiers, the value of P_n will increase. If $P_n \geq 0.05$, then there is no statistically significant difference between the random and correct training sets; thus, the accuracy of the classifier cannot be reliably estimated. For model-fitting, we only use those $n \in \mathbf{N}$ for which $P_n < 0.05$, i.e. the set of significant training set sizes $\mathbf{M} \subset \mathbf{N}$. It is important to note that the robustness of the classifier is also validated implicitly by the fact that the power law model requires a minimum of three training set sizes, i.e. $|\mathbf{M}| \geq 3$, to extrapolate a performance curve.

2.4. Estimation of Power Law Model Parameters

The power-law model [3] describes the relationship between error rate and training set size:

$$\bar{e}_n = a n^{-\alpha} + b, \tag{3}$$

where $\bar{e}_n$ is the error rate for training set size n, a is the learning rate, α is the decay rate, and b is the Bayes error, which is the error rate given an infinite amount of training data and is considered to be the lowest possible error [2]. The model parameters a, α, and b can be found via the constrained non-linear minimization

$$\min_{a,\alpha,b} \sum_{m=1}^{|\mathbf{M}|} (a n_m^{-\alpha} + b - \bar{e}_n)^2, \tag{4}$$

where $a, \alpha, b \geq 0$. In this paper, the MATLAB function *fmincon* was used to perform this calculation.

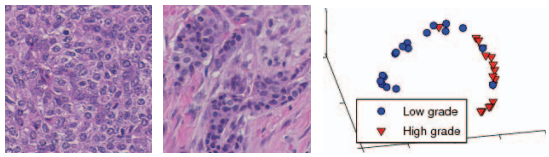
3. DESCRIPTION OF PROBLEMS AND DATASETS

To evaluate our prediction methodology we considered 4 different CAD problems, all of which had limited amounts of data. The objective in all 4 problems was to (a) predict classifier error rate, and (b) compare classifier trends, assuming a pre-defined number of training samples were to become available.

3.1. Dataset $\mathcal{D}_1$: ER+ Breast Cancer Grading

Bloom-Richardson (BR) grade is known to be correlated with breast cancer (BC) prognosis [4]. Grade determination is however subject

to high inter- and intra-pathologist variability. Therefore, a CAD algorithm that automatically determines BR grade would be invaluable to clinicians. In this dataset [5], hematoxylin and eosin (H & E) stained estrogen receptor-positive (ER+) breast biopsy slides were digitized at 20x optical magnification using a whole-slide digital scanner. For our study, each image is classified as either high-grade (BR grade $>$ 6, Figure 1(a)) or low-grade (BR grade $<$ 6, Figure 1(b)). For each histopathology image, all cancer nuclei were first automatically detected [5]. The centroids of the detected nuclei were used as nodes to construct the Voronoi Diagram, Delaunay Triangulation, and Minimum Spanning Tree graphs for each image. A total of 12 architectural features were extracted to characterize the spatial arrangement and density of the nuclei. Graph Embedding, type of non-linear dimensionality reduction, was used to embed the features into a 3-dimensional space as shown in Figure 1(c). The 3D embedding, along with the ground truth labels, allows us to visualize the dataset and the discriminability of its features. In this work a classifier was trained using the 12 architectural features to identify each image as being either high or low grade.



(a) (b) (c)

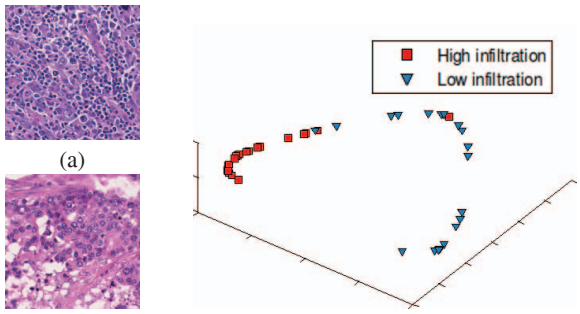
Fig. 1. BC histopathology samples from $\mathcal{D}_1$ denoting (a) high and (b) low-grade tumors. (c) A 3D Graph Embedding of the architectural features illustrates the clear separation between high and low grade tumors.

3.2. Dataset $\mathcal{D}_2$: Lymphocytic Infiltration in Breast Cancer

The extent of lymphocytic infiltration (LI) in HER2+ breast cancers has recently been linked to likelihood of tumor recurrence and distant metastasis. While the presence of LI can be assessed qualitatively by pathologists, a quantitative and reproducible CAD system will provide greater insight into the relationship between LI and prognosis. This dataset [6] comprises H & E stained HER2+ breast biopsy tissue prepared and digitized in a manner similar to $\mathcal{D}_1$. Regions of interest were selected by an expert pathologist and labeled as having either high (Figure 2(a)) or low (Figure 2(b)) LI extent. The CAD system first automatically detects the lymphocyte nuclei and treats their centroids as nodes to construct three graphs (similar to $\mathcal{D}_1$) [6]. For each image, the graphs were used to extract a set of 50 architectural features quantifying the spatial arrangement, density, and nearest neighbor statistics [6]. Similar to $\mathcal{D}_1$, Graph Embedding is used to visually represent the separation between high and low LI extent (Figure 2(c)). In this work a classifier was trained using the 50 architectural features to identify each image as having either high or low LI extent.

3.3. Dataset $\mathcal{D}_3$: Detection of Cancer in MRI

The ability to detect and localize prostate cancer via radiological studies will help clinicians provide more targeted therapies to prostate cancer patients. In dynamic contrast enhanced magnetic resonance imaging (DCE-MRI) studies, a contrast agent is injected into the patient, which alters the intensity profile of the MRI image over time. In this dataset, we combine the functional information from the DCE images with structural information from T2-weighted



(a) (b) (c)

Fig. 2. BC histopathology samples from $\mathcal{D}_2$ denoting (a) high and (b) low LI extent. (c) A 3D Graph Embedding of the samples shows clear separation between samples with high and low LI extent.

3 Tesla MRI images, both captured *in vivo* [7]. A set of 14 features corresponding to (a) pixel intensities of the DCE images and (b) pixel intensity textural features from the T2-weighted MRI images is extracted from each study. The features are used to create a probability map, whereby lighter pixels represent likely cancerous regions.Unlike datasets $\mathcal{D}_1$ and $\mathcal{D}_2$, the data in $\mathcal{D}_3$ is classified on a pixel-wise basis as either cancer or non-cancer. Classification accuracy is evaluated by pixel-by-pixel comparison with a ground truth for cancer extent determined via registration of whole mount histology specimens obtained via radical prostatectomy and the corresponding *in vivo* MRI.

3.4. Dataset $\mathcal{D}_4$: Detection of Cancer in Prostate Histology

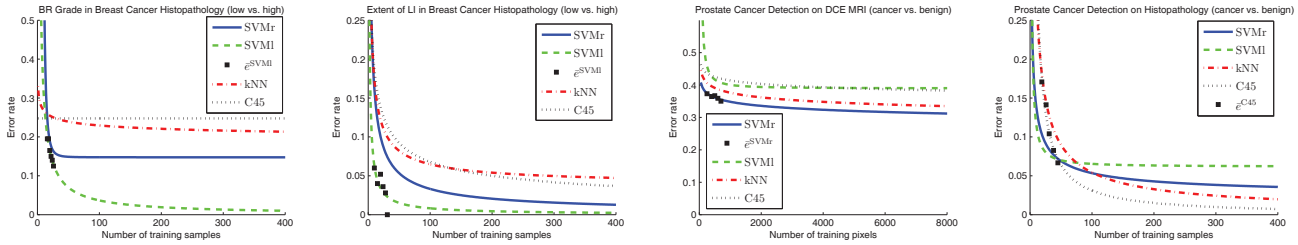
An automated method for detecting prostate cancer on biopsy specimens will improve productivity by helping pathologists focus their efforts solely on samples with cancer and ignore those without. H & E stained needle-core biopsies of prostate tissue were digitized at 20x optical magnification on a whole slide digital scanner, similar to the method used in $\mathcal{D}_1$ and $\mathcal{D}_2$. Ground truth, i.e. regions corresponding to prostate cancer, was manually delineated by a pathologist. For this dataset, the images were divided into 30-by-30-pixel grids and a set of 14 texture features was extracted from each region [8]. Similar to $\mathcal{D}_1$ and $\mathcal{D}_2$, the objects of classification are individual tissue regions, which were considered cancerous if greater than 50% of the pixels in the region were labeled as cancerous and labeled as non-cancerous otherwise.

4. EXPERIMENTAL RESULTS AND DISCUSSION

The performance of Support Vector Machine, both with a linear kernel (SVMl) and a radial basis function kernel (SVMr), C4.5 decision tree (C45), and k-nearest neighbor (kNN) classifiers were compared. For a given dataset, the same training sets $\mathcal{S}_{n,i}$ and testing set $\mathcal{T}$ were used for all classifiers.

4.1. Estimating Error Rate for Large Training Cohort

The extrapolated performance curves (Figure 3) from the best classifiers for each dataset (discussed in Section 4.2) are used to predict error rates for future experiments involving larger data cohorts. Assuming a training set comprised of 200 samples, the datasets $\mathcal{D}_1$ with SVMl, $\mathcal{D}_2$ with SVMl, $\mathcal{D}_3$ with SVMr, and $\mathcal{D}_4$ with C45 can be expected to produce error rates of 0.0192, 0.00444, 0.198, and 0.0151, respectively. Note that, since $\mathcal{D}_3$ utilizes pixel-wise classification



(a) Breast Cancer Grading ($\mathcal{D}_1$) (b) Lymphocytic Infiltration ($\mathcal{D}_2$) (c) Prostate Cancer MRI ($\mathcal{D}_3$) (d) Prostate Cancer Histology ($\mathcal{D}_4$)

Fig. 3. Extrapolated classification performance curves are shown for SVM using both radial basis function (SVMr) and linear kernel (SVMl), k-nearest neighbor (kNN), and C4.5 decision tree (C45) classifiers on all 4 CAD datasets. On each plot the black squares represent the actual mean error rates $\bar{e}_n$ corresponding to the classifier that produces the lowest extrapolated error rates for large training set sizes.

(Figure 3(c)), we assume that one sample contains approximately 4,500 pixels.

4.2. Optimal Classifier Selection

We define the "optimal" classifier as one that produces the lowest error rates for large training set sizes. The importance of extrapolating classifier performance is illustrated in Figure 3(a), where C45 and kNN both perform better than SVMl while training cohort size $n \leq 13$. After $n > 13$, however, SVMl is predicted to perform better than the other classifiers. The same phenomenon can be observed in Figure 3(d), where SVMl consistently provides the best performance while $n \leq 46$. As the training set grows, it becomes apparent that C45 will be the optimal classifier when larger data cohorts are used in future studies. Conversely, note that training cohort size does not always affect the selection of an optimal classifier. For example, the optimal classifiers for Figures 3(b), (c) are SVMl and SVMr, respectively, for all training set sizes.

5. CONCLUDING REMARKS

Over the last few years there has been an explosion in the development of CAD systems for radiology and histology applications [5, 6, 7, 8]. While novel methods have been developed, extensive annotated databases are still in the process of being compiled, both for classifier training and evaluation. Evaluation on most of these CAD systems has been limited to small cohorts, and procedures for classifier selection have been either ad hoc or based on comparison studies involving a small number of training samples. Instead, classifier selection needs to be dictated by performance on a large cohort of data. The problems explored in this paper are common to many CAD applications:

1. Given a limited amount of training data, are there strategies we can employ to predict with a high degree of confidence the classifier that would perform best for a specific CAD task when provided with a larger cohort of data?
2. For CAD applications (e.g. tumor volume estimation) where the classifier is required to make decisions at the pixel level, is there a methodology that allows us to predict the power of the classifier in separating benign and cancerous pixels?

Note that while one might argue that classifier predictions on larger cohorts of data could also be made via traditional sample power calculations, it should be noted that the confidence levels associated with predictions made from these calculations (using small cohorts) are low. The approach presented in this paper, which employs a bootstrap-based subsampling procedure, allows for predicting classifier performance assuming a larger cohort of data was available, with a higher degree of confidence. Additionally this approach allows for performance analysis of pixel-level classifiers; these cannot be directly accommodated via traditional sample power calculations where the assumption is that the decision is made at the patient level.

We also acknowledge some limitations of this study. Firstly, the classifier comparison trends predicted were not actually validated on larger cohorts. The reason for this is simply that those larger datasets are not yet available. Secondly, classifier trends are also a function of training data quality (e.g. accuracy of annotations), choice of features, and number of classes, none of which were analyzed in this study. In addition, the effect of variations in parameters T_1 and T_2 are not studied in this work. We intend to explore these issues in future work.

6. REFERENCES

[1] L. Didaci, G. Giacinto, F. Roli, and G.L. Marcialis, "A study on the performances of dynamic classifier selection based on local accuracy estimation," *Pattern Recognition*, vol. 38, 2005.

[2] Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*, Wiley, 2001.

[3] S. Mukherjee, P. Tamayo, S. Rogers, R. Rifkin, A. Engle, C. Campbell, T.R. Golub, and J.P. Mesirov, "Estimating dataset size requirements for classifying dna microarray data," *J. Comput. Biol.*, vol. 10(2), pp. 119–142, 2003.

[4] H.J. Bloom and W.W. Richardson, "Histological grading and prognosis in breast cancer; a study of 1409 cases of which 359 have been followed for 15 years," *Br. J. Cancer*, vol. 11, no. 3, pp. 359–377, 1957.

[5] A. Basavanhally, A. Madabhushi, et al., "Computer-aided prognosis of er+ breast cancer histopathology and correlating survival outcome with oncotype dx assay," *ISBI*, pp. 851–854, 2009.

[6] A. Basavanhally, A. Madabhushi, et al., "Computerized image-based detection and grading of lymphocytic infiltration in her2+ breast cancer histopathology," *IEEE Trans. on Biomed. Eng.*, 2009 (in press).

[7] S. Viswanath, A. Madabhushi, et al., "Integrating structural and functional imaging for computer assisted detection of prostate cancer on multi-protocol in vivo 3 tesla mri," in *SPIE Medical Imaging*, 2009.

[8] S. Doyle, A. Madabhushi, et al., "Automated grading of prostate cancer using architectural and textural image features," in *ISBI*, 2007, pp. 1284–1287.