Comparison of Machine Learning and Traditional Classifiers in Glaucoma Diagnosis

Kwokleung Chan (kwchan@salk.edu) and Te-Won Lee (tewon@salk.edu)
Computational Neurobiology Laboratory, The Salk Institute

Pamela A. Sample (psample@eyecenter.ucsd.edu), Michael Goldbaum (mgoldbaum@ucsd.edu) and Robert N. Weinreb
Department of Ophthalmology, UCSD

Terrence J. Sejnowski
Computational Neurobiology Laboratory, The Salk Institute

Abstract. Glaucoma is a progressive optic neuropathy with characteristic structural changes in the optic nerve head reflected in the visual field. Visual field sensitivity test is hence commonly used in clinical setting to evaluate glaucoma. We compared the performance of a number of machine learning algorithms to STATPAC, which is a traditionally used statistical analysis package in the glaucoma community. The machine learning algorithms studied include Multi Layer Perceptron (MLP), Support Vector Machine (SVM), Mixture of Gaussian (MOG) and Mixture of Generalized Gaussian (MGG). MLP and SVM are classifiers that fall under the discriminative paradigm. Generative classifiers which first model the data probability density and then perform classification via Bayes' rule, though is more computationally demanding, usually give deeper insight into the structure of the data space. We have applied the mixture of Gaussian (MOG) and the mixture of generalized Gaussian (MGG), which is an extension of independent component analysis (ICA), to our glaucoma classification problem. Performance of the various classifiers are compared by area under the Receiver Operating Characteristic (ROC) curve and sensitivities (true positive rates) at some pre-chosen specificities (true negative rates). The machine learning type classifiers show improved performance over the best indices from STATPAC. Forward selection and backward elimination were implemented to identify the set of important variables in the diagnosis. This methodology further improves the classification rate and in practice it also reduces testing time requirements due to fewer visual field measurements.

Keywords: Standard automated perimetry, STATPAC, Bayes Rule

1. Introduction

In a two classes classification problem, we are given a training dataset \( \{x_i, y_i\}, i = 1, \ldots, N \) where \( x_i \in \mathcal{R}^D \) (could contain both continuous and discrete entries) is the input and \( y_i = \pm 1 \) is the output label. When performing classification, one approach is to first model the class conditional probability for each class \( p(x|C_{\pm}) \), and then employ the
Bayes’ rule

\[
P(C_{\pm}|x) = \frac{p(x|C_{\pm})P(C_{\pm})}{p(x)}
\]  

(1)

Under the Cox Axioms (Cox, 1961), the Bayes’ rule is the only consistent way to manipulate beliefs and plausibility, if they are represented by real numbers. Classification using equation (1) is also known as the generative paradigm, since the generation of the data point \( x \) is first modeled. This effectively reduces the problem of classification to that of modeling the class conditional probability distribution \( p(x|C_{\pm}) \) for the two classes.

However, it has always been difficult to model \( p(x|C_{\pm}) \) accurately. Naive Bayes’ classifier (Friedman et al., 1997) assumes independency between components of input \( x \). Modeling \( P(C_{\pm}|x) \) through \( p(x|C_{\pm}) \) is known to be inefficient (Mitchell, 1997) as it generally requires the estimation of more parameters. Take the example of doing the classification problem by modeling the two classes of data with Gaussian densities of same variance but different means. It takes \( D(D + 1)/2 + D + D \) parameters in this approach. The resulting classifier is well known to be a linear discriminate function \( u(x) = w \cdot x + b \) which only needs \( D + 1 \) parameters. For a dataset of finite size, this means that we have fewer data points for each parameter in the generative approach. This renders the generative methods less efficient, for the sole purpose of classification. And the resulting classifier will be less robust against outliers and noise in the data.

This suggests we may be better off using the discriminative approach in which the posterior probabilities \( P(C_{\pm}|x) \) are directly estimated. Logistic regression (Cox and Snell, 1989) is a famous example of the discriminative approach and is widely used in medical research. Decision tree is another kind of discriminative classifier. Recently attention has been shifted to neural network type classifiers (Bishop, 1995; Ripley, 1996) and the Support Vector Machine (SVM) (Vapnik, 2000). In some of these classifiers, there may even be no estimation of the posterior probabilities. The classifier simply returns the label \( y \) by applying discrimination functions on the input \( x \).

The advantage of discriminative classifiers is that they concentrate on the decision boundary and hence are robust against irrelevant outliers in the training data. However, they provide less insight to the structure of the data space. Hence it is difficult to handle data containing missing entries. The Multi-layer Perceptron and Support Vector Machine often serve as black boxes in classification and it is very difficult for humans to comprehend how the decision is made. In this paper, we will apply both the generative and discriminative classifiers to our glaucoma diagnosis problem and compare their performance.
2. Glaucoma

Glaucoma is a progressive optic neuropathy with characteristic structural changes in the optic nerve head reflected in the visual field (Hitchings and L., 1977). In the clinical setting, glaucoma is commonly evaluated using funduscopic examination of the optic disk and visual field testing (Johnson, 1997). Standard automated perimetry (SAP) is currently the visual function test most relied upon to measure visual function in glaucoma. Automated threshold perimetry gives detailed quantitative data.

In SAP, a target $0.47^\circ$ in diameter of variable intensity is flashed for 200 msec against a background of 31.5 apostilbs (10 candelas per meter squared (cd/m$^2$)). The most commonly used procedure worldwide is the full threshold standard automated perimetry test, program 24-2 or 30-2 of the Humphrey Visual Field Analyzer (HFA, Humphrey-Zeiss, Dublin, CA). The target is randomly presented to 52 locations with a HFA over $24^\circ$ at 2 decibels resolution. The displayed outputs (figure 1) are the absolute sensitivity, the gray scale, the age-corrected total deviation (numerical and probability), and the pattern deviation (numerical and probability), the glaucoma hemifield test result (GHT), and the global indices (MD, PSD, SF & CPSD). The age corrected total deviation is the absolute sensitivity subtracted from an age-matched normal surface. The pattern deviation is the total deviation compensated by global depression to account for cataracts or other non glaucoma conditions that may globally depress the visual field. The initial output of the Humphrey Field Analyzer is the absolute sensitivity at each visual field location. This output consists of the raw sensitivity threshold values, represented in decibels (dB) relative to the maximum intensity of the machine (set a 0 dB) with a minimum of 40 dB. These values, together with the age of the patient, will then constitute the raw input of our classifiers.

3. Classifiers

3.1. Multi-layer Perceptron

The Multi-layer Perceptron (MLP) (Bishop, 1995), also termed feed forward network, is a generalization of the single-layer perceptron studied by (Rosenblatt, 1958). The MLP is a universal approximator to any real valued functions. In fact, a feed forward network of just two layers (not including the input layer) is enough to approximate any continuous function (Blum and Li, 1991). The MLP has been successfully applied
Figure 1. A sample STATPAC printout from the HFA. Top row: absolute sensitivities and gray scale plot over the 54 locations on the retina. Middle: age corrected total deviation and pattern deviation (total deviation compensated by global depression). Bottom: probability plots of total deviation and pattern deviation.
to a wide class of problems such as face recognition (Samal and Iyengar, 1992) and optical character recognition (Baird, 1993).

In the setting of two class classification problem, we have for a given input \( x = (x_1, \ldots, x_D)^T \),

\[
z_j = g \left( \sum_{d=1}^{D} w_{jd} x_d + w_{j0} \right)
\]

\[
f = h \left( \sum_{j=1}^{J} v_j z_j + v_0 \right)
\]

\( z_j, j = 1, \ldots, J \) are the activations of the hidden layer units. \( w_{jd} \) are the weights between the input and the hidden layer. Similarly, \( v_j \) are weights connecting the hidden layer to the output unit \( f \). \( w_{j0} \) and \( v_0 \) are correspondingly the biases for the hidden and output units. \( h(t) \) and \( g(t) \) are continuous sigmoid function, usually of the form \( \tanh(t) \) or the logistic function \( 1/(1 + e^{-t}) \).

MLP is one of the most popular architectures among other neural networks such as the radial basis function (Broomhead and Lowe, 1988) for its efficient training by error backpropagation (Rumelhart et al., 1986). The desired error function in classification is however not the mean squared error (MSE), but the negative log likelihood function

\[
\log L = - \sum_{i=1}^{N} y_i \log f_i + (1 - y_i) \log (1 - f_i)
\]

Here it is assumed that the logistic function is used for \( h(t) \) and the output label \( y \) takes the values of \( \{1, 0\} \), instead of \( \{+1, -1\} \). Despite having a different error function, the equations in the error backpropagation remain unchanged.

### 3.2. Support Vector Machine

Support Vector Machine (SVM), is a newly developed technique for solving a variety of classification and regression problems (Vapnik, 1998; Vapnik, 2000). It exploits the statistical learning theory to minimize the generalization error when training a classifier. SVMs have demonstrated to produce good generalization performance in face recognition (Osuna et al., 1997) and text categorization (Dumais et al., 1998) and OCR (Schölkopf et al., 1996; Burgers and Schölkopf, 1997).

The general form of the decision function \( f(x) \) of a SVM is given by

\[
f(x) = \text{sign}(u(x))
\]

\[
u(x) = \sum_{i=1}^{N} \alpha_i y_i k(x, x_i) + b
\]
\[
u(x) = \sum_{i=1}^{N} \alpha_i y_i k(x, x_i) + b
\]

where \( k(x_i, x_j) \) is known as the kernel function and has the below commonly used form,

\[
\begin{align*}
k(x_i, x_j) &= x_i \cdot x_j \\
k(x_i, x_j) &= (x_i \cdot x_j + 1)^p \\
k(x_i, x_j) &= \exp[-|x_i - x_j|^2/(2\sigma^2)] \\
k(x_i, x_j) &= \tanh(\kappa x_i \cdot x_j + \theta)
\end{align*}
\]  

\( \alpha_i \)'s are chosen by the SVM through training, subjected to constraints \( \sum \alpha_i y_i = 0 \) and \( 0 \leq \alpha_i \leq A \). \( A \) is a user defined penalty term controlling the generalization performance of the SVM. Upon training, only a fraction of the \( \alpha \)'s will be non-zero, whose corresponding \( x \)'s are known as support vectors since only they will contribute to the decision function through \( u(x) \) in equation (6). The architecture of the SVM in classification is shown in figure 2.

Instead of a hard decision function as in equation (5), following (Platt, 1999), we could also convert the SVM output \( u(x) \) into a probabilistic one by a logistic transformation of parameters \( a \) and \( b \),

\[
P(C_+ | x) = \frac{1}{1 + \exp[a u(x) + b]}
\]  

\[ (11) \]
3.3. Mixture of Gaussians

As mentioned in the introduction, the generative approach is to model the class conditional density \( p(x|C_\pm) \). Since the input of the glaucoma data contains continuous valuable in a 53 dimensional space, \( x \in \mathcal{R}^{53} \), we may want to model \( p(x|C_\pm) \) each by a normal multivariate density. This would result in a linear discriminant function (LDF) or a quadratic discriminant function (QDF), depending on if the two normal densities are constrained to have the same covariance or not. However, many careful studies show that most real distributions usually do not follow a normal distribution but have slightly heavier tails, skewed or even bi-modal structure. It is likely that a single Gaussian is not flexible enough to model adequately the distribution of data.

In simple non-parametric method such as the histogram method, the input space is divided into many small hypercubes and then \( p(x) \) is estimated for each of them. Binning the data space subjects to the curse of dimensionality, besides not providing much useful insight in the statistical structure of the data. To model the probability distribution of the data properly, we need models with “in-between” flexibility and this is where semi-parametric models come about. The mixture of Gaussians (MOG) (Titterington et al., 1985) has been popular for its simplicity.

Adopted to our classification problem, the probability densities for the positive and negative classes are each first modeled as a mixture of multivariate normal densities.

\[
p(x) = \sum_{m} p(x|m)P(m) \tag{12}
\]

where for each cluster \( m \),

\[
p(x|m) = \frac{1}{\sqrt{2\pi}^{p}|\Sigma_m|} \exp \left[ -\frac{1}{2}(x - \mu_m)^T \Sigma_m^{-1}(x - \mu_m) \right] \tag{13}
\]

\( x^T \) denotes the transpose of \( x \) and \( |\Sigma| \) is \( \text{det}(\Sigma) \). The Expectation Maximization (EM) algorithm (Dempster et al., 1977) is used to find the parameters \( \mu_m \) and \( \Sigma_m \). With \( p(x|C_\pm) \) modeled in turns, Bayes’ rule can used to obtain the posterior probability for each class, given a new example \( x \),

\[
P(C_\pm|x) = \frac{p(x|C_\pm)P(C_\pm)}{p(x)} \tag{14}
\]

The architecture of the MOG classifier is shown in figure 3.
Figure 3. Architecture of the MOG used in a binary classification setting. $p(x|\pm)$ are the generative models for the two classes. Each is composed of a mixture of Gaussians ($p_+(x|m_+)$ or $p_-(x|m_-)$). Output $P(+|x)$ is obtained by applying Bayes' rule (equation 14) on $p(x|\pm)$.

3.4. **MIXTURE OF GENERALIZED GAUSSIANS**

Although the mixture of Gaussians provides a more flexible model to fit the density of the data, it would be undesirable to fit a density of long tails with two Gaussians. Besides adequately fitting the data density, user may want to understand the structure of the data in terms of number of real clusters and their deviation from normality. With the development of generalized Gaussian mixture model (Lee and Lewicki, 2000), we are able to model the class-conditional densities $p(x|C_{\pm})$ with higher flexibility, while preserving a comprehension of the statistical properties of the data in terms of means, variances and kurtosis etc. The mixture of generalized Gaussians (MGG) uses the same mixture model (equation 12) as the MOG. However, each cluster is now described by a linear combination of non-Gaussian random variables $s_m$,

\[
p(x|m) = \delta[x - (A_ms_m + b_m)]
\]

(15)
i.e. $s_m$’s are considered to be the hidden source responsible for generating the observation $x$’s given $A_m$ and $b_m$, and a noiseless case is assumed here. $s_m$’s are assumed to be generated from a generalized Gaussian density of zero mean, unit variance and parameter $\beta_m$. 
Variance of \( x \) in each cluster will be absorbed into \( A \).

\[
p(s_m | \beta_m) = p(s_{m1}, \ldots, s_{mD} | \beta_{m1}, \ldots, \beta_{mD}) = \prod_d \omega(\beta_{md}) \exp \left[ -c(\beta_{md}) |s_{md}|^{2/(1+\beta_{md})} \right]
\]

where

\[
c(\beta) = \frac{\Gamma[\frac{3}{2}(1+\beta)]}{\Gamma[\frac{1}{2}(1+\beta)]} \frac{1}{(1+\beta)}
\]

is to maintain unit variance and

\[
\omega(\beta) = \frac{\Gamma[\frac{3}{2}(1+\beta)]^{1/2}}{(1+\beta) \Gamma[\frac{1}{2}(1+\beta)]^{3/2}}
\]

is the normalization constant. \( \beta \) is a measure of kurtosis of the source (Box and Tiao, 1973).

\[
\text{kurtosis} = \frac{\Gamma[\frac{3}{2}(1+\beta)] \Gamma[\frac{3}{2}(1+\beta)]}{\Gamma[\frac{5}{2}(1+\beta)]^2} - 3
\]

\( \beta_m \) will be adapted together with \( A_m, b_m \) and \( P(m) \) from \( x \) during training, which could be done by gradient ascent on the data likelihood (Lee and Lewicki, 2000).

3.5. STATPAC

The Humphrey Field Analyzer (HFA) comes with a statistical analysis package (STATPAC) that provides both the raw data and several specialized statistical analyses related to diagnosing glaucoma. The purpose of these analyses is to aid the clinician in interpretation of the visual field. The global indices included in STATPAC are mean deviation (MD), pattern standard deviation (PSD), short-term fluctuation (SF), corrected pattern standard deviation (CPSD) and the glaucoma hemifield test (GHT). MD is the depression of the patient’s overall field (all test locations averaged) as compared to the age-corrected normative database within the HFA. PSD is a measurement of the degree to which the shape of the field departs from the age-corrected reference field. SF is an index of the consistency of the patient’s answers during the field test and is obtained by testing twice at ten predetermined points. CPSD is the PSD corrected for SF to attempt to remove the effects of patient variability during the test and to reveal only irregularities caused by actual field loss.
The glaucoma hemifield test (GHT) divides the superior hemifield into five zones and compares locations within each zone to those within a mirror image zone in the inferior hemifield. The five pairs of mirroring sectors are compared and a difference score for each is determined. If this difference is outside the 99.5% limits in any one pair compared to the difference score found in age-corrected normal eyes, the field is flagged "outside normal limits (ONL)". If it is outside the 97% limit the field is flagged "borderline (BL)". Lesser differences are considered "within normal limits (WNL)". These analyses are currently used to help the clinician in interpretation of the visual field.

We will evaluate the efficiency of GHT, indices PSD and CPSD in glaucoma diagnosis. Their results will be used as the baseline against which our classifiers' performance are measured.

4. Experimental Settings

4.1. The Machine Classifiers

Our glaucoma dataset contains a collection of 156 eyes with glaucomatous optic neuropathy (GON) and 189 eyes without GON. Visual fields were not used in labeling the data. Since the amount of data available is limited, separation into training and test sets is not preferred. We used a ten folds cross validation scheme instead to evaluate the classifiers. The dataset was divided uniformly into ten subsets. Each subset was in turn held aside as the test set when the other nine were used to train the classifiers. The results on the ten subsets were combined into one single ROC plot (see section 5.1) for each classification method.

To facilitate training, we first normalized each of the 52 locations raw sensitivity threshold values and age to have zero mean and unit variance. For the MLP, Matlab Neural Network Toolbox 3.0 was used. The network contained a hidden layer of 10 tanh units and a logistic output unit. The network was trained using the Levenberg-Marquardt method (Levenberg, 1944). Early stopping was used to prevent overfitting. In each fold of the cross validation, 20 networks were trained and their output were averaged to return a single \( P(C_+ | \mathbf{x}) \) value for each testing data point \( \mathbf{x} \).

In the SVM, we used the radial basis function (RBF) kernel (equation (9)) with \( \sigma = \sqrt{33} \approx 7.3 \) and \( A=1 \). These values were chosen to give approximately optimal performance of the SVM on the validation set. We implemented the Sequential Minimal Optimization (SMO) (Platt, 1998; Keerthi et al., 1999) in MATLAB code to train the SVM.

For the mixture of Gaussian (MOG) and mixture of generalized Gaussian (MGG) classifier, due to the limited availability of data, we
first performed principle component analysis (PCA) on the normalized data to reduce the dimension. The data was projected onto the first eight components subspace which accounted for more than 80% of variance. Class conditional densities \( p(x|C_\pm) \) were modeled separately on the two classes. The optimal number of clusters could have been found using the Bayesian Information Criterion (BIC) (Schwarz, 1978) or the Akaike Information criterion (AIC) (Akaike, 1974). However, we simply chose the number of clusters for each class to produce the best performance on the testing set. For a fair comparison of the classifiers, the MLP and SVM were also trained in the eight dimension subspace, to single out the gain or loss from dimension reduction.

The Linear Discriminant Function (LDF) is a special form of the MOG in which only one Gaussian is fitted to each of the glaucoma and non-glaucoma classes. Moreover, the two Gaussians are restricted to have equal covariance. The LDF can work in both full and reduced dimensional space. Its performance will be reported as a base line comparison.

4.2. Feature Selection

It is always useful and interesting to identify a subset of input variables that contribute more in the classification. Besides, getting rid of irrelevant input variables which function as noise usually would improve classification. In the simplest setting, input variables can be ranked by their classification rates when used alone to perform class prediction. This ignores the dependence between the variables which could be important in predicting the output variable. Or in a linear classifier, we can order the input by the absolute values of the weights associated with the input variables. For non-linear classifiers, exhaustive search over all possible combinations of input variables to identify the best subset is prohibitively expensive, especially when number of input variables to be included is unknown. Here we would use forward selection and backward elimination (Ripley, 1996) to rank the variables and identify the subset that would give best classification. Forward selection is sequentially adding variables one at a time, choosing the next variable that most increases or least decrease classification. Backward elimination starts with all input variables and sequentially delete the next variable that most decreases or least increases classification.
Figure 4. Sample distribution of classifier output $u(x)$ on a two classes dataset

5. Results

5.1. Receiver Operating Characteristic (ROC) curve

A simple way to assess the performance of classifiers is to compare their average misclassification rate. In biomedical data, the dataset usually comes with a much higher proportion of normal (–ve class) data. Classifiers will hence tend to achieve an overall low misclassification rate by sacrificing the +ve class data, resulting in a higher misclassification rate on the +ve class than the –ve class. However, the misclassification cost associated with the +ve class is usually much higher than that of –ve class. For a classifier which outputs a scalar value showing the degree of a given data point $x$ belonging to the +ve class, such as $P(C_+|x)$ or $u(x)$ as in equation (6), we would like to pick a decision threshold other than 0.5 for $P(C_+|x)$ (or 0 for $u(x)$), in favor of the +ve class.

In figure 4, data points of two classes are placed next to the x-axis according to the value of $u(x)$ given by a classifier. Together plotted are the factional densities (smoothed histogram) showing the distribution of $u(x)$ for the two classes. Different classifiers would have different density distribution for $u(x)$. True positive rate, also known as sensitivity, is the fraction (or %) of positively labeled test data classified as +ve. This is also the area under the positive class density curve, to the
right of the decision threshold (0.35 in figure 4). Specificity is the true negative rate, fraction of negatively labeled examples classified as −ve.

By varying the threshold level \( \theta \) such that a given example \( \mathbf{x} \) will be classified as positive if \( u(\mathbf{x}) > \theta \), we get a trade off between sensitivity and specificity. Receiver operating characteristic (ROC) curve is a plot of sensitivity v.s. 1–specificity (or true positive rate v.s. false positive rate). Area under the ROC curve summarizes the quality of classification over a wide range of differences in misclassification cost. We estimate the variance of the ROC area by a non-parametric method as described by (DeLong et al., 1988). There is an interesting interpretation of the ROC area. It is equal to the probability of a random sample from positive class \( \mathbf{x}_+ \) begin assigned, by the classifier, a \( u(\mathbf{x}_+) \) (or \( P(C_+|\mathbf{x}_+) \)) value greater than that of a random negative class sample, i.e. \( \text{ROC area} = \text{Prob} (u(\mathbf{x}_+) > u(\mathbf{x}_-)) \), for any random sample pair \((\mathbf{x}_+, \mathbf{x}_-)\) drawn from the two classes.

5.2. Classification Results

The ROC areas for the classifiers are summarized in table I, grouped into the following categories: STATPAC, Discriminative (in full and reduced dimension), Generative (reduced dimension) and Feature Selection. We listed also the sensitivities of the classifiers at specificities of 0.90 and 0.75. The index PSD is very competitive to our classifiers. The CPSD, which is derived from PSD by correcting for short term fluctuation (SF) in visual field sensitivity, does not exhibit improvement over PSD. The GHT identified 66.7\% of the glaucoma eyes as “outside normal limits” and all of the normals as “within normal limits”. Probably this is because GHT is designed to have a specificity value of 99.5\%.

In general, all the machine classifiers perform better than the STATPAC indices. For example, at 0.75 specificity, the STATPAC indices (excluding SF) gave an average of 0.786 in sensitivity, while the machine classifiers (excluding feature selection) had an average of 0.842. Performance on the glaucoma dataset of both the discriminative and generative classifiers are very similar, consider that they all are given the best parameters for the results reported here. The ROC areas vary around 84\% to 92\% while the error estimated is approximately 2\%. The ROC curves of the best classifier from each category are plotted together in figure 5 for comparison.

5.3. Feature Selection

From table I, SVM with Gaussian kernel is the best classifier that can work directly on the full dimensional input. In addition, its fast
Table I. A comparison of performance of classifiers by their ROC areas and sensitivities at selected specificities.

<table>
<thead>
<tr>
<th></th>
<th>ROC area (std. err.)</th>
<th>Sensitivities at 0.90 Spec</th>
<th>Sensitivities at 0.75 Spec</th>
</tr>
</thead>
<tbody>
<tr>
<td>STATPAC:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GHT(^1)</td>
<td></td>
<td>0.667</td>
<td></td>
</tr>
<tr>
<td>MD</td>
<td>0.838 (0.022)</td>
<td>0.654</td>
<td>0.731</td>
</tr>
<tr>
<td>SF</td>
<td>0.694 (0.029)</td>
<td>0.365</td>
<td>0.532</td>
</tr>
<tr>
<td>PSD</td>
<td>0.883 (0.020)</td>
<td>0.756</td>
<td>0.846</td>
</tr>
<tr>
<td>CPSD</td>
<td>0.844 (0.025)</td>
<td>0.737</td>
<td>0.782</td>
</tr>
<tr>
<td>Discriminative:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(full dim.)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLP</td>
<td>0.898 (0.017)</td>
<td>0.737</td>
<td>0.878</td>
</tr>
<tr>
<td>gaussian SVM</td>
<td>0.903 (0.017)</td>
<td>0.712</td>
<td>0.878</td>
</tr>
<tr>
<td>linear SVM</td>
<td>0.894 (0.018)</td>
<td>0.686</td>
<td>0.859</td>
</tr>
<tr>
<td>(reduced dim.)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLP</td>
<td>0.897 (0.017)</td>
<td>0.724</td>
<td>0.833</td>
</tr>
<tr>
<td>gaussian SVM</td>
<td>0.899 (0.017)</td>
<td>0.750</td>
<td>0.853</td>
</tr>
<tr>
<td>linear SVM</td>
<td>0.887 (0.018)</td>
<td>0.673</td>
<td>0.846</td>
</tr>
<tr>
<td>Generative:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(full dim.)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LDF</td>
<td>0.832 (0.023)</td>
<td>0.603</td>
<td>0.763</td>
</tr>
<tr>
<td>(reduced dim.)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LDF</td>
<td>0.879 (0.018)</td>
<td>0.635</td>
<td>0.814</td>
</tr>
<tr>
<td>MOG</td>
<td>0.922 (0.015)</td>
<td>0.788</td>
<td>0.840</td>
</tr>
<tr>
<td>MGG</td>
<td>0.906 (0.017)</td>
<td>0.782</td>
<td>0.853</td>
</tr>
<tr>
<td>Features Selection (gaussian SVM)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>forward selection</td>
<td>0.921 (0.015)</td>
<td>0.756</td>
<td>0.910</td>
</tr>
<tr>
<td>backward elimination</td>
<td>0.923 (0.015)</td>
<td>0.769</td>
<td>0.929</td>
</tr>
</tbody>
</table>

\(^1\)The specificity of GHT cannot be varied. Reported sensitivity corresponds to the specificity of 1.0 on the test data.

training allows lots of repeated training in a reasonable period of time. We performed forward selection and backward elimination using the Gaussian SVM to rank the input variables. ROC area from the ten fold cross-validation is used as the criteria for selecting subset of variables. In figure 6 plotted are the ROC areas as a function of number
of variables included from the list ranked by forward selection and backward elimination. Both methods peak at 18 variables and give a 20% increase in ROC area over full dimensional Gaussian SVM. Moreover, performance of the full dimension input has be achieved by using only the 4 most important input variables. In figure 7 we plot the ranks given by forward selection and backward elimination on the 53 input variables onto a two dimensional space. Visual field locations are labeled 1 to 54 as displayed in the upper left insert. Location 18 and 31 correspond to the blind spot on the retina and are omitted. Variables near the origin (e.g. location 5, 6 and 47 etc.) are considered by both forward selection and backward elimination most important towards glaucoma diagnosis.
Figure 6. ROC area as a function of number of variables used in a Gaussian SVM. Variables are added according to the rank given by forward selection or backward elimination.

6. Discussion

6.1. Compared to STATPAC

From table I and figure 5, the machine classifiers perform in par with the STATPAC indices PSD and CPSD at high specificities and better at medium to low specificities. Although the results of our classifiers are listed with best parameters used, they are at a relative disadvantage since the STATPAC indices are derived using age corrected reference visual fields from a much larger normative database within the HFA. Our classifiers are trained on the raw sensitivity threshold values (and age) from a relatively limited dataset. We would expect better performance from our classifiers if the age corrected reference visual fields are available.

6.2. Among the Machine Classifiers

Contrary to reported in other papers (Roth and Steinhag, 2000; Platt et al., 2000), the SVM produces no significant improvement over the MLP. Its ROC area is very close to that of MLP while its sensitivities at the selected specificities are only slightly better than that of MLP.
Figure 7. Correlation of rankings given by forward selection and backward elimination on the variables (52 visual field locations + Age). Variables close to origin (e.g. location 5, 6 and 47 etc.) are more informative in glaucoma diagnosis. Upper left insert displays the relative positions on the retina of the variables.

However, the SVM is 20 to 40 times faster to train than the MLP, and does not require repeated training with random initial conditions. Since the SVM by itself is a linearly constrained quadratic programming problem, it will not get stuck at local minimum during training. The performance of the SVM is insensitive to the exact choice of the parameters such as $A$ and $\sigma$. Their values found in cross validation can be used to train the whole dataset for prediction of future unseen examples.

Although the mixture of generalized Gaussian (MGG) is an enhancement over the MOG in modeling densities of continuous variables, the MGG + Bayes' rule did not improve over the MOG classifier. One explanation is that non-Gaussian nature of data has already been tackled by the mixture model, the extra flexibility in modeling non-Gaussian densities introduces additional free parameters. This may result in over-
fitting of the training data and poor generalization to unseen testing data.

6.3. Feature Selection

The sequential selective addition/elimination of features permits identifying subset of input variables that collectively contribute most in classification in a manageable amount of time. The number of variables that gives the peak performance can be interpreted as the intrinsic dimension of the data. From figure 6, the glaucoma SAP data has an intrinsic dimension around 18. This is reasonable since visual fields close together should have correlated sensitives. Besides giving better classification accuracy, this also allows an reduction in the time spent by the patients in standard automated perimetry (SAP) from 15 to 5 minutes per eye and hence more screening can be done. Moreover, real time classification is made possible when the visual locations are tested in the order ranked by feature selection.

6.4. The LDF

In this paper, the linear discriminant function (LDF) is categorized under generative classifiers despite the word “discriminant” in its name. This may look strange but in fact a LDF can be either generative or discriminative, depending on if its weight are obtained from first modeling $p(x|C_{\pm})$ or not. Logistic regression, a single layer MLP or a SVM with linear dot product kernel are all effectively linear classifiers. LDF resulted from them should be considered as discriminative since they work directly on the decision boundary during training. As seen from table I, the linear SVM performs better than the generative LDF both in the full and reduced dimensional space. Moreover, the generative LDF show markable improvement when working in the reduced dimensional space. This shows that discriminative classifiers are generally more robust against irrelevant input variables. This is because the number of parameters for discriminative LDF grows at $O(D)$ while that of generative LDF grows at $O(D^2)$ ($D$ is the dimension of the input space). The generative LDF is intrinsically more complex in structure and would be more variable and less robust against outliers.

6.5. Generative v.s. Discriminative

Although the posterior probabilities $P(C_{\pm}|x)$ are insensitive to minor variation in class conditional densities $P(x|C_{\pm})$, since $P(x|C_{\pm})$ is vulnerable to noise and outliers of the data, classifiers based on the generative model usually show a lower degree of robustness against
outliers. Opposite to this expectation, performance of the MOG in the subspace is better than the MLP and SVM in our glaucoma problem. This demonstrates the power of Bayes' rule in classification, provided that we can model the underlying statistical structure of the data accurately (compare LDF and MOG). Use of generalized Gaussian mixture enables modeling data of unknown kurtosis along different directions in the data space. The enhanced flexibility provides better fit to non-Gaussian real world data. Even in cases where the generative models together with Bayes' rule perform sub-optimally, they allow understanding of the intermediate steps of the decision, rather than working like a black box as in MLP.

Another advantage of the generative methods is besides outputting $P(C_+|x)$ and $P(C_-|x)$, the generative classifiers can alert the user when $\hat{p}(x) = p(x|C_+)P(C_+) + p(x|C_-)P(C_-)$ falls below some pre-chosen level. The new observation $x$ is probably an outlier, better classified as unknown and requires extra attention from user. Furthermore, when a new $x$ to be classified contains missing entries, instead of simply putting in the means from training data, we can infer the missing values $x_{\text{missing}}$ by marginalizing $p(x)$ over $x_{\text{observed}}$. It is anticipated to have a better classification than simply replacing missing entries with mean values.

While the discriminative classifiers MLP and SVM do not differentiate between discrete and continuous input data and hence can be directly applied to discrete data without modification, the extension of the mixture models to data vector containing both continuous and discrete entries need some attention. This is done by first dividing data vector $x$ into a continuous part $v$ plus a discrete part $z$, i.e. $x = (v, z) = (v, z_1, \ldots, z_n)$, then model $p(x)$ as

$$p(x) = p(v, z) = p(v|z)P(z) = p(v|z_1, \ldots, z_n)P(z_1, \ldots, z_n)$$  \hspace{1cm} (20)$$
i.e. we model the continuous data space separately for each possible value of the discrete component. When available data is limited or $N$ is large, we may employ the naive Bayes' classifier assumption and equation (20) reduces to

$$p(v, z) \approx p(v)P(z) \approx p(v) \prod_i^n P(z_i)$$  \hspace{1cm} (21)$$
which is feasible both computationally and practically. Domingos and Pazzani (Domingos and Pazzani, 1996) have shown that the naive Bayesian classifier is remarkably effective, even if the independent assumption seldom holds.
7. Conclusion

We have compared a variety of machines classifiers to the STATPAC indices in glaucoma diagnosis on standard automated perimetry (SAP) data. The machines classifiers in general give improvement over the STATPAC indices as measured by area under the ROC curve. They show promise for use in clinical setting together with the STATPAC indices for glaucoma diagnosis. Forward selection and backward elimination were used to rank and select the visual field locations for glaucoma diagnosis and improved classification is obtained. Properties, advantages and disadvantages of generative and discriminative machine classifiers have been also compared. The success of the machine classifiers in the SAP data shows promise for their applications in progression prediction and other visual function or structural tests for glaucoma, such as Short-wavelength Automated Perimetry (Sample and Weinreb, 1990) and optic nerve head topography (Mikelberg et al., 1995).

This research is supported by grant EY08208 (PAS) and EY13235 (MHG).
References


