Bayesian neural network approaches to ovarian cancer identification from high-resolution mass spectrometry data

Jiangsheng Yu\textsuperscript{1,2} and Xue-Wen Chen\textsuperscript{2,3,}\textsuperscript{*}  
\textsuperscript{1}School of Electronics Engineering and Computer Science, Peking University, China, \textsuperscript{2}Information and Telecommunication Technology Center and \textsuperscript{3}Department of Electrical Engineering and Computer Science, University of Kansas, KS USA

Received on January 15, 2005; accepted on March 27, 2005

ABSTRACT

Motivation: The classification of high-dimensional data is always a challenge to statistical machine learning. We propose a novel method named shallow feature selection that assigns each feature a probability of being selected based on the structure of training data itself. Independent of particular classifiers, the high dimension of biodata can be fleetly reduced to an applicable case for consequential processing. Moreover, to improve both efficiency and performance of classification, these prior probabilities are further used to specify the distributions of top-level hyperparameters in hierarchical models of Bayesian neural network (BNN), as well as the parameters in Gaussian process models.

Results: Three BNN approaches were derived and then applied to identify ovarian cancer from NCI’s high-resolution mass spectrometry data, which yielded an excellent performance in 1000 independent \( k \)-fold cross validations (\( k = 2, \ldots, 10 \)). For instance, indices of average sensitivity and specificity of 98.56 and 98.42\%, respectively, were achieved in the 2-fold cross validations. Furthermore, only one control and one cancer were misclassified in the leave-one-out cross validation. Some other popular classifiers were also tested for comparison.

Availability: The programs implemented in Matlab, R and Neal’s fbm.2004-11-10.

Contact: xwchen@ku.edu

1 INTRODUCTION

Due to the high death rate in advanced stage diseases, the diagnosis of early-stage cancer is critical for public health. Of recent, the novel biotechnology of high-throughput and high-resolution MALDI-TOF mass spectrometry (MS) in the low-molecular-weight region of the blood proteome has made such diagnosis possible (Lilien \textit{et al}., 2003; Liotta \textit{et al}., 2003; Petricoin and Liotta, 2003, 2004; Wulfkuhle \textit{et al}., 2003). Many methods have been tried on such MS data for pattern recognition and identification of diseases, for instance, generic algorithm and self-organizing clustering (Petricoin \textit{et al}., 2002), classification and regression tree (CART, Vlahou \textit{et al}., 2003), random forest (Wu \textit{et al}., 2003), etc. Considering the high dimensionality of MS data and relatively scarce samples, more efficient statistical approaches are still urgently needed. In this paper, we propose three Bayesian neural network (BNN) models based on some data reduction techniques, that achieved a promising performance on the SELDI-TOF high-resolution ovarian MS data\textsuperscript{1} provided by National Cancer Institute (NCI).

Originally, the merged ovarian MS data are described by 373 401 \( m/z \) ratios. After binning, the dimension is reduced to 11 301. However, a further reduction is still necessary for various applications due to the computational complexity. In some ways, feature selection (FS) can be partially determined by the dataset itself, once some standards are given. Such FS, independent of particular classifiers, is fair to model comparisons and, sometimes can provide more information for the proposed classification. In Section 2, we utilized the bootstrap technique (Efron and Tibshirani, 1993) to assign each feature a probability of being selected in the two-sample Kolmogorov–Smirnov goodness-of-fit test (KS-test), which will be used to specify the distributions of top-level hyperparameters in Bayesian models.

The neural network (NN) approach, extensively used in classification and regression, is attractive because of its ability to model the complex non-linear system (Bishop, 1995; Ripley, 1996). Theoretically, NN can simulate any continuous function in a compact range if the number of hidden units tends to infinity (Cybenko, 1989). In recent years, BNN with infinite hidden units has shown some interesting relationship with the kernel method (Schölkopf and Smola, 2002), and with the help of hierarchical models the overfitting problem

\textsuperscript{1}There are 121 cancer cases and 95 non-cancer cases in the dataset, performed by QC/QA analysis (Conrads \textit{et al}., 2004) and available at http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp
has largely eased (Gelman et al., 2004). Especially for a small
training sample, prior knowledge, whether from the historical
data or domain experts, seems helpful to hasten learning and
yield good performance in prediction.

The output of BNN for a new sample, \(x_{\text{new}}\), is a distribution
of probability for \(x_{\text{new}}\) being in class 1, and the estimated mean
determines its class. In Section 3, two hierarchical models and
a Gaussian process (GP) model, based on maximum likeli-
hood estimation (MLE), whose (hyper)parameters are defined
by means of shallow feature selection (SFS), are derived
and discussed. This novel method improves the efficiency of
BNN while retaining good performance in classification. In
the section on experiments and analysis, we compare BNN
with some other machine learning methods, such as CART,
support vector machine (SVM), bagging (bootstrap aggreg-
ing) of multilayer perceptron (MLP) (Bauer and Kohavi,
1999), \(k\)-nearest neighbor (\(k\)-NN), etc. on NCI’s ovarian data.
Lastly, some further investigations proposed are mentioned in
the conclusion.

2 SFS

Given the (normalized) training set \(D = \{(x^i, t^i) | x^i \in \mathbb{R}^m, t^i \in \{0, 1\}\} \) and \(i = 1, 2, \ldots, n\), where \(0, 1\) are class
labels and samples are recorded in column vectors.

\[
\begin{array}{cccccccc}
1 & \cdots & x^k & x^{k+1} & \cdots & x^n \\
\hline
f_1 & x_{1,1} & \cdots & x_{1,k} & x_{1,k+1} & \cdots & x_{1,n} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
f_m & x_{m,1} & \cdots & x_{m,k} & x_{m,k+1} & \cdots & x_{m,n} \\
\end{array}
\]

FS is always a hard core of pattern recognition, which often
gets entangled badly in classification problems. Even so, there
are still some characteristics of the dataset itself that could
be used to filter out these significantly useless features. For
instance, \(f_i\) could be more useful if it could be known at which
of the two classes the empirical distributions are statistically
distinct. Moreover, to make sense for a reasonable represen-
tation of the training data, each feature should be associated
with a prior probability of being selected for classification,
independent of particular classifiers. We call it SFS.

Without any assumption of population distribution, the empirical
distributions of values in \(X_i = (x_{i,1}, \ldots, x_{i,k})\) and \(X_i' = (x_{i,k+1}, \ldots, x_{i,n})\) are compared by a two-sided KS-test
(i.e. the null hypothesis \(H_0\) is that \(X_i\) and \(X_i'\) are drawn from the same continuous distribution) with a given significance
level \(\alpha\). After repeating the procedure \((N - 1)\) more times
on the resampled training sets, simple voting will result in
\(
\Pr(f_i | \alpha) \approx P(H_0 \text{ rejected})/N,
\)
the probability of choosing \(f_i\) as a feature (Fig. 1).

A more flexible way can be based on the reasonably accurate
\(P\)-values guaranteed by \(n_i n'_i / (n_i + n'_i) \geq 4\), where \(n_i, n'_i\) are
the sample sizes of \(X_i\) and \(X_i'\) respectively (Lehmann, 1975).

In addition, the bias and standard error calculated from the \(N\)
bootstrap replicates of \(P\)-value will provide us more infor-
mation about SFS (Table 1), especially for the ambibulous
case of \(\Pr(f_i | \alpha) \approx 0.5\). Although the SFS is independent of
specific classifiers, it is not isolated from classification. In our
BNN models, \(\Pr(f_i | \alpha)\) is used to specify the covariance matrix
of GP and the prior distributions of input-to-hidden weights.

Unlike in some other classifier-independent approaches to
data reduction, such as principle component analysis (PCA),
wavelet analysis, etc. (Yu et al., 2005), SFS is capable of
selecting a set of features without losing its biological mean-
ing. Furthermore, the probability assigned to each feature is
actually a kind of prior knowledge about its contribution to
classification. In contrast, neither PCA nor wavelet utilizes the
class information in data reduction.

Table 1. Bootstrapping of \(P\)-value for SFS of \(m/z\) intervals from NCI’s
high-resolution ovarian MS data, where the training set is resampled 999
times

<table>
<thead>
<tr>
<th>(m/z) Interval</th>
<th>([710,711])</th>
<th>([8000,8001])</th>
<th>([9054,9055])</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original (P)-value</td>
<td>0.9579</td>
<td>5.1125 \times 10^{-9}</td>
<td>0.0675</td>
</tr>
<tr>
<td>Bias</td>
<td>(-1.9318 \times 10^{-14})</td>
<td>0</td>
<td>(-3.0531 \times 10^{-16})</td>
</tr>
<tr>
<td>Standard error</td>
<td>1.9328 \times 10^{-14}</td>
<td>0</td>
<td>3.0546 \times 10^{-16}</td>
</tr>
<tr>
<td>(\Pr(f_{[710,711]}</td>
<td>\alpha = 5%)</td>
<td>0.094</td>
<td>1</td>
</tr>
<tr>
<td>(\Pr(f_{[8000,8001]}</td>
<td>\alpha = 1%)</td>
<td>0.029</td>
<td>1</td>
</tr>
</tbody>
</table>

At the significance level \(\alpha = 5\%, \) the \(m/z\) interval \([8000,8001]\) will be more preferred
compared to \([710, 711]\) and \([9054, 9055]\) by the hypothesis
testing.

Fig. 1. The scatter plots of ‘probability’ of selecting an \(m/z\) ratio
interval from NCI’s high-resolution ovarian data as a feature at the
given significance levels of \(\alpha = 5\%\) and \(\alpha = 1\%\).
3 BINARY CLASSIFICATION BY BNN

Unlike the conventional frequentist training methods for MLP that are based on MLE, the Bayesian approach to NNs (Buntine and Weigend, 1991; MacKay, 1992a,b; Neal, 1996) utilizes the prior distribution of network weights before training and revises it after observing the data. Therefore, BNN can provide us not only the means of predictive weights but also their uncertainties. Another advantage of BNN is the appropriate choices of a number of hidden layers and their dimensions (Müller and Rios Insua, 1998; Rios Insua and Müller, 1998), which are almost impossible to solve by traditional NNs.

The inference of underlying Bayesian binary classifier from $D$, parameterized by the weight vector $w$, is to find $f(x, w|D, M)$ for the probability of $x$ being in class 1 in model $M$. As is usual let $M$ be a logistic model $f(x, w) = [1 + \exp[-h(x, w)]]^{-1}$ satisfying $\int f(x, w)dw < \infty$, that makes $h(x, w)$ meaningful by

$$h(x, w) = \log \frac{\Pr(t = 1|x, w)}{\Pr(t = 0|x, w)} \tag{1}$$

Considering the approximation ability, in this paper $h$ is assumed to be the adaptive basis function of one hidden layer (with $u$ units) perceptron$^2$ $h(x, w) = b + W^T \psi(b + M^T x)$. Since the biases $b$ and $b$, if necessary, can be absorbed in $M^T$ and $W^T$, respectively (Rios Insua and Müller, 1998), we simply consider

$$h(x, w) = W^T \psi(M^T x) = W^T y \tag{2}$$

where $w = (M, W)$ and $\psi$ is a bounded activation function, such as hyperbolic tangent $\tanh(z) = [\exp(z) - \exp(-z)]/[\exp(z) + \exp(-z)]$, or error function $\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z \exp(-t^2)dt$. The binary classification turns out to be a noise-free regression problem of $\log(\Pr(x \in \text{class 1})) = h(x, w)$, where $\log(p) = \log(p/(1-p))$. Obviously, the likelihood is

$$\Pr(D|w) = \prod_{i=1}^{N} f(x^{\mu}, w)^{y}(1 - f(x^{\mu}, w))^{1-y} \tag{3}$$

Once given $p(w)$, the posterior distribution of parameters $p(w|D) \propto p(w)p(D|w)$ is clear and the probability of testing sample $x$ being in class 1 is

$$\Pr(t_{\text{new}} = 1|x_{\text{new}}, D) = \int f(x_{\text{new}}, w)p(w|D)dw \approx \frac{1}{N} \sum_{i=1}^{N} f(x_{\text{new}}, w_i) \tag{4}$$

where $w_i$ are drawn from population of $p(w|D)$. The intuitive assumption of $p(w) \propto \exp[-\xi w^T w/2]$ relies on the belief that smaller weights lead to less certain predictions (Bishop, 1995), where the hyperparameter $\xi$ is defined in a hierarchical model, for instance the automatic relevance determination (MacKay, 1995). The integral in Equation (4) can be evaluated simply by straightforward Metropolis algorithm (Metropolis et al., 1953; Gelman et al., 2004), but is inefficient for applications (Kass et al., 1998). MacKay proposed Gaussian approximations centered on the posterior modes (MacKay, 1992a) implemented on the software of BUGS using Gibbs sampling (Spiegelhalter et al., 1995)$^3$, while Neal used hybrid Monte Carlo (HMC) (Duane et al., 1987; Liu, 2001) with simulated annealing (Neal, 1992). The calculation of Equation (4) becomes extremely difficult when the hidden units are too many.

Again, the desired $p(w)$ can be viewed as the prior distribution over all the functions described by Equation (2). Motivated by the GP approach to nonparametric Bayesian regression (O’Hagan, 1978), Neal showed that a large class of NN models converge to GPs within the limit of an infinite number of hidden units, and argued compellingly that the overfitting problem could be alleviated by some hierarchical model of hyperparameters (Neal, 1996). Since then, there is growing interest in applications of GP to BNN (Rasmussen, 1996; Gibbs, 1997; MacKay, 1998; Williams, 1998; Neal, 1999; Schönloft and Smola, 2002; Williams, 2002). Theoretically, it is guaranteed by Kolmogorov’s consistency theorem that the finite-dimensional distributions of GP are uniquely determined by its finite mean and covariance functions.

The advantage of GP is that we do not need to make complicated assumptions on the distribution of $w$. Nevertheless there are no free lunches. We have to argue for the covariance function—it seems worth it because the distribution of $h(x_{\text{new}}, w|D)$ will be achieved. The following diagram is a comparison between the traditional way and GP method.

\[ \begin{array}{c}
  w \rightarrow h(\cdot, w) \rightarrow \text{Gaussian process} \\
  \downarrow \rightarrow \text{Parameterized covariance matrix of r.v.} \\
  p(\psi) \downarrow \text{h(\cdot, w) with E(h) = 0, which specifies} \\
  \downarrow \text{Estimates of hyperparameters in the} \\
  \downarrow \text{covariance matrix (MLE or MCMC methods)} \\
  \downarrow \text{Gaussian posterior distribution of h at x_{new}} \\
  \end{array} \]

$^3$See MacKay’s homepage http://wol.ra.phy.cam.ac.uk/mackay and the website of WinBUGS http://www.mrc-bsu.cam.ac.uk/bugs for the user manual of BUGS (version 0.6) or WinBUGS (version 1.4).
Although the logistic model in terms of latent values $h = h(x, u)$ can be specified by a wide variety of covariance functions, mainly it is in the form of

$$\text{Cov}(h, h') = \delta_{hh'}\epsilon + c^2 + x^T \text{diag}(\gamma_1^2, \ldots, \gamma_m^2) x' + \beta^2 \exp \left\{ -(x - x')^T \text{diag}(\lambda_1^2, \ldots, \lambda_m^2)(x - x') \right\}$$

where $c^2, \gamma_i^2, \beta^2, \lambda_i^2$ are hyperparameters. Due to the computational issue, in practice the covariance function is often added a ‘jitter’ $\delta_{hh'}\epsilon$, where $\epsilon > 0$ and $\delta_{hh'}$ is the Kronecker delta. The term of $c^2 + x^T \text{diag}(\gamma_1^2, \ldots, \gamma_m^2)x'$ is for the linear regression, and the last term is based on the assumption that nearby inputs will have highly correlated outputs, allowing a different distance measure for each input dimension, that is, the smaller the $\lambda_i^2$, the lesser the importance of $i$-th feature (Williams and Rasmussen, 1996).

The joint distribution of GP $h_{n+1} = (h_1, \ldots, h^n, h_{\text{new}})^T$ given $X_{n+1} = (x_1, \ldots, x^n, x_{\text{new}})$ and parameterized covariance function matrix $K_{n+1}$ is

$$h_{n+1}|X_{n+1}, K_{n+1} \sim N_{n+1}(E h_{n+1}, K_{n+1})$$

where $K_{n+1} = \begin{pmatrix} K_n & k_{\text{new}} \\ k_{\text{new}}^T & \kappa \end{pmatrix}$, $K_n = \text{Cov}(h^{i\mu}, h^{j\nu})|_{n\times n}$.

$$k_{\text{new}} = (\text{Cov}(h^{1}, h_{\text{new}}), \ldots, \text{Cov}(h^{n}, h_{\text{new}}))^T$$

and $\kappa = \text{Cov}(h_{\text{new}}, h_{\text{new}})$. With the assumption of $E h = 0$ for the equal prior probability of each class, we get the conditional distribution of $h_{\text{new}}$ with hyperparameters given $t_n = (t_1, \ldots, t^n)^T$ or equivalently $h_n = (h_1, \ldots, h^n)^T$ (Bickel and Doksum, 2001)

$$h_{\text{new}}|x_{\text{new}}, D, K_{n+1} \sim N_{\kappa} \left( K_{\kappa}^{-1} h_n, \kappa - K_{\kappa}^{-1} K_n k_{\text{new}} \right)$$

For any hyperparameter $\theta$ in $K_n$, it can be simply replaced by the MLE from $\delta L / \delta \theta = 0$, where $L$ is the log-likelihood of $h_n$ (Mardia and Marshall, 1984)

$$L = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log \text{det} K_n - \frac{1}{2} h_n^T K_n^{-1} h_n$$

$$\frac{\partial L}{\partial \theta} = -\frac{1}{2} \text{tr} \left( K_n^{-1} \frac{\partial K_n}{\partial \theta} \right) + \frac{1}{2} h_n^T K_n^{-1} \frac{\partial K_n}{\partial \theta} K_n^{-1} h_n$$

Figure 2 is a simulation of random $h$-surface for MLP with 2-D input $(x_1, x_2)^T$ by Equation (5), where $\lambda_1^2 = \Pr(f_1|x) = 0.9, \lambda_2^2 = \Pr(f_2|x) = 0.3$ are assumed for feature selecting probabilities and $\beta^2 = 1, \epsilon = 0.001, c^2 = 0, \gamma_i^2 = 0$. Less $\Pr(f_i|x)$ makes $\sigma^2$ more likely close to 0, restricting the input-to-hidden $M_{i,j}$ near 0 regularly.

One of the main tasks of statistical machine learning is to train the parameters, that may affect both the efficiency and performance of classification substantially. In practice, parameter tuning is not trivial for BNN models because of the hidden noisy features in high-dimensional data. Instead of being initialized randomly or trained under strong assumption of distributions, our MLE-based GP model is specified by $\lambda_1^2 = \gamma_1^2 = \Pr(f_1|x), \beta^2 = c^2 = 1, \epsilon = 0.001$, where $\Pr(f_i|x)$ under weakens the influence of ‘useless’ features in the covariance function. This strategy fairly simplifies the GP model and yields excellent performance on ovarian data analysis (Table 2). It may be argued that an FS is suitable for all classifiers, since the categorical characteristics explored and utilized by distinct methods could be quite different. That is why SFS is preferred—the prior knowledge on features can

---

3Since the input-to-hidden weight vectors [i.e. the columns of $M = (M_{1,1}, \ldots, M_{\lambda,\mu})$] are iid, surely $y_j$'s are also iid, nevertheless $y_j, y'_j$ are not independent.
Table 2. k-fold cross validations of MLE-based GP model on reduced data (k = 2, . . . , 10)

| k   | Data reduced by Pr(f|α = 0.1%) ≥ 1 | Data reduced by Pr(f|α = 0.1%) ≥ 0.9 | Data reduced by Pr(f|α = 0.1%) ≥ 0.6 |
|-----|-----------------------------------|-------------------------------------|-------------------------------------|
|     | Control Mean SD                   | Cancer Mean SD                       | Control Mean SD                     |
| 2   | 0.9855 0.0081                      | 0.9804 0.0094                        | 0.9831 0.0112                       |
| 3   | 0.9883 0.0092                      | 0.9813 0.0092                        | 0.9896 0.0071                       |
| 4   | 0.9874 0.0081                      | 0.9836 0.0110                        | 0.9895 0.0078                       |
| 5   | 0.9893 0.0088                      | 0.9827 0.0093                        | 0.9883 0.0076                       |
| 6   | 0.9892 0.0101                      | 0.9809 0.0099                        | 0.9886 0.0080                       |
| 7   | 0.9883 0.0104                      | 0.9821 0.0100                        | 0.9894 0.0080                       |
| 8   | 0.9910 0.0081                      | 0.9821 0.0110                        | 0.9900 0.0086                       |
| 9   | 0.9896 0.0095                      | 0.9801 0.0111                        | 0.9888 0.0075                       |
| 10  | 0.9885 0.0111                      | 0.9823 0.0105                        | 0.9891 0.0088                       |
| 216 | daf-0281                          | daf-0608, 0667, 0754                | daf-0281                           |

The features are selected by Pr(f|α) ≥ threshold from the training sets, here we choose α = 0.1% and threshold = 1, 0.9, 0.6. The dimensions of feature space are brought down from 11 301 to ~1320, 4270 and 5830 respectively. There is a balance between accuracy and complexity. The last line records the misclassified samples in the leave-one-out cross validations.

be modified after the data are introduced, even for a small training sample.

Alternatively in a fully Bayesian way, each parameter is estimated as the mean of its posterior distribution by MCMC methods (Gelman et al., 2004; Liu, 2001) based on an assumption for its prior, which seems more robust in many complicated cases (Rasmussen, 1996). In the following, we will specify the distributions for mutually independent parameters in BNN model (4) and GP model (7), respectively. These two fully Bayesian methods, together with the MLE method for GP model, will be examined on NCI's ovarian data in the next section.

1. A mathematically convenient prior of unknown variance is inverse gamma distribution, the conjugate family for Gaussian weights.

\[ W | \sigma_0^2 \sim N_u(0, \sigma_0^2 I_u) \] (11)

\[ \sigma_0^2 \sim \text{Inv-Gamma}(a_0, b_0) \] (12)

\[ M_{., j} | \Sigma \sim N_m(0, \Sigma) \] (13)

\[ \sigma_i^2 \sim \text{Inv-Gamma}(a_i, b_i) \] (14)

where Inv-Gamma(\(\theta|a, b\) = \(\frac{b^a}{\Gamma(a)} \theta^{-(a+1)} \exp(-b/\theta)\), \(\theta > 0\) with shape \(a > 0\) and scale \(b > 0\). The full conditionals of \(\sigma_0^2\) and \(\sigma_i^2\) are

\[ \sigma_0^2 | W, M, \Sigma, D \sim \text{Inv-Gamma}(a_0 + u/2, b_0 + \|W\|^2/2) \] (15)

\[ \sigma_i^2 | \sigma_0^2, \ldots, \sigma_{i-1}^2, \sigma_{i+1}^2, \ldots, \sigma_n^2, W, M, D \sim \text{Inv-Gamma}(a_i + u/2, b_i + \|M_{., i}\|^2/2) \] (16)

Similar to Neal’s suggestion (Neal, 1996), here we set \(a_0 = 0.01, b_0 = 0.5, a_i = 0.01, b_i = Pr(f_i|\alpha)\) to make the density flatter for larger Pr(f_i|\alpha).

2. In the fully Bayesian model of GP, the priors of linear part \(y^2 = (y_{11}, \ldots, y_{n1})^T\) and relevance \(\lambda^2 = (\lambda_1^2, \ldots, \lambda_n^2)^T\) are similar to that of \(M_{., j}\), and the prior of scale \(\beta^2\) is similar to that of \(W_j\). The jitter \(\epsilon\) and constant \(c^2\) in Equation (5) are set to be 0.01 and 1 respectively.

4 EXPERIMENTS AND ANALYSIS

For each k-fold cross validation of MLE-based GP method (\(k = 2, \ldots, 10\)), independent experiments have been repeated 1000 times on the data reduced by SFS to get the average specificity and sensitivity with corresponding standard deviation (Table 2). The number of selected features is controlled by both \(\alpha\) and threshold, for instance on the original binned data,

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Significance level (\alpha)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1%</td>
</tr>
<tr>
<td>1</td>
<td>1328</td>
</tr>
<tr>
<td>0.9</td>
<td>4274</td>
</tr>
<tr>
<td>0.7</td>
<td>5423</td>
</tr>
<tr>
<td>0.5</td>
<td>6232</td>
</tr>
</tbody>
</table>

Also, MLE-based GP model (briefly, GP_{MLE}) behaves very stable in leave-one-out cross validations on distinct reduced datasets (see the last line of Table 2).

As an efficient simulation means, HMC method (Duane et al., 1987; Liu, 2001; Neal, 1992) plays a major practical role in computing the tractable Bayesian integrals by using dynamical methods and Metropolis acceptance rules to propose and accept transition states. We constructed an NN with
J.Yu and X.-W.Chen

Fig. 3. The left figure is the HMC simulation of hidden-to-output weights, where \( u = 15 \) and \( x \)-axis is the number of iterations. The right one is the simulation of the first five input-to-hidden weights of hidden unit 1.

Table 3. 2, 10-fold and leave-one-out cross validations of SVM (\( \gamma = 2, C = 1 \)), bagging of back-propagation MLP with one-hidden layer of 300 units, \( k \)-NN on NCI’s (reduced) binned ovarian MS data, each validation is repeated 100–1000 times independently

<table>
<thead>
<tr>
<th>Method</th>
<th>2-Fold cross validation</th>
<th>10-Fold cross validation</th>
<th>Leave-one-out (( #(\text{misclassifications}) ))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control</td>
<td>Cancer</td>
<td>Control</td>
</tr>
<tr>
<td>---------------------</td>
<td>---------</td>
<td>--------</td>
<td>---------</td>
</tr>
<tr>
<td>GP MLE, GP Bayes</td>
<td>0.9842</td>
<td>0.0069</td>
<td>0.9856</td>
</tr>
<tr>
<td>BNN</td>
<td>0.9723</td>
<td>0.0090</td>
<td>0.9862</td>
</tr>
<tr>
<td>SVM</td>
<td>0.9414</td>
<td>0.0214</td>
<td>0.8929</td>
</tr>
<tr>
<td>Bagging of MLP</td>
<td>0.9221</td>
<td>0.0180</td>
<td>0.9697</td>
</tr>
<tr>
<td>1-NN</td>
<td>0.8793</td>
<td>0.0299</td>
<td>0.8883</td>
</tr>
<tr>
<td>2-NN</td>
<td>0.8409</td>
<td>0.0377</td>
<td>0.8775</td>
</tr>
<tr>
<td>CART</td>
<td>0.7836</td>
<td>0.0478</td>
<td>0.8225</td>
</tr>
</tbody>
</table>

One hidden layer of 15 units, the hidden-to-output weights converge after 80 iterations (Figure 3). Compared to GP MLE and SVM, the fully Bayesian methods are much more time-consuming because of MC simulations. Nevertheless, there is not too much difference in performance between the GP model and Bayesian MLP (Lampinen and Vehtari, 2001), hence GP MLE may be an experiential way for both accuracy and efficiency.

Some other classifiers are also tested, but the precision is inferior to that obtained by BNN as far as it goes (Table 3). Using some sophisticated techniques of data reduction, such as restriction of coefficient of variation, wavelet analysis, SVM (Cortes and Vapnik, 1995; Cristianini and Shawe-Taylor, 2000; Vapnik, 1995) with soft margin can also achieve a good performance level in sensitivity and specificity on NCI’s ovarian data. Regardless of computational complexity, the bagging of MLP (Bauer and Kohavi, 1999) is able to yield another reliable \( \Pr(t_{new} | x_{new}, D) \), especially for a small training sample, from the viewpoint of ensemble learning. Several samples were found almost always misclassified in cross validations, for instance those explored by MLP bagging in Table 3.

5 CONCLUSION

We described the SFS method as the main technique of data reduction, playing an important role in specifying the top-level hyperparameters of hierarchical BNN. Compared with the published results in literature, this method shows to be a promising application for ovarian cancer identification from NCI’s high-resolution MS data. In 1000 independent 2-fold cross validations, BNN achieves average levels of sensitivity and specificity of 98.56 and 98.42%, respectively, and only misclassifies one control and one cancer in leave-one-out cross validation. Contrasted with low-resolution data, high-resolution can lead to superior classification (Conrads et al., 2004; Yu et al., 2005) and then to successful clinical applications (Petricoin and Liotta, 2004).

Further work will cover more applications to other common diseases, such as prostate cancer, colon cancer, breast cancer,
etc. Interesting topics that were not covered systematically in this paper are the accurate FS for biomarkers, and a method to figure out those abnormal sample points that are always misclassified by a high-quality ensemble, however well trained by resampled datasets or by distinct methods.

ACKNOWLEDGEMENTS

This investigation was supported by the National Science Foundation (Grant No. EPS-0236913). We also appreciate the support from ITTC and the University of Kansas.

REFERENCES


Springer-Verlag, New York.


