ERROR PROBABILITY ANALYSIS FOR LDA-BAYESIAN BASED CLASSIFICATION OF ALZHEIMER'S DISEASE AND NORMAL CONTROL SUBJECTS

Zhe Wang, Tianlong Song, Yuan Liang and Tongtong Li

Department of Electrical & Computer Engineering, Michigan State University

ABSTRACT

This paper provides a theoretical analysis on the classification accuracy of LDA-Bayesian based method with respect to the data sample size in brain connectivity analysis. More specifically, we show that when the sample size increases, both the classification error probability and its upper bound decreases monotonically. However, we also show that due to the model limitation of the Bayesian classifier, the classification error probability is actually lower bounded as well, which implies that the error probability actually converges to a non-zero constant even if the data sample size tends to infinity. Our analysis is demonstrated through fMRI based numerical results.

***Index Terms*—fMRI, Alzheimer's Disease Bayesian, Error Probability**

I. INTRODUCTION

Accurate distinction of Alzheimer's Disease (AD) and normal control (NC) subjects is critical for early diagnosis and treatment of brain disorders. Recently, functional magnetic resonance imaging (fMRI) data, which maps brain activities to metabolic changes in cerebral blood flow, has been used to classify AD and NC subjects [1], [2]. In [1], Wang et al. extracted two intrinsically anti-correlated networks using resting state fMRI data from 14 AD patients and 14 NC subjects, and applied a Pseudo-Fisher Linear Discriminative Analysis (pFLDA) on the high dimensional feature vectors. Their *two-category* classification accuracy was 83%. In [2], Chen et al. applied the same technique to larger datasets. Similarly, the accuracy of the *two-category* classification of AD patients and NC subjects was 82%.

Compared with other methods like EEG, fMRI data can display active brain areas more directly, and has much better spatial resolution throughout the brain. Unlike structural MRI which mainly reflects the anatomical information of brain tissues and structure, fMRI focuses on functional brain activities, and can provide more direct measurement on how different brain regions are involved in particular brain activities.

While structural MRI has been widely applied to clinical diagnosis of brain disorders, fMRI has mainly been used for research purposes. As a result, the size of fMRI data samples is generally quite limited, which has become a major bottleneck in fMRI based AD and NC classification. The underlying reason is that, when the sample size is small, most existing classifiers could potentially suffer from noise effects, due to both biological variability and measurement noise.

Motivated by this observation, in this paper, we provide a theoretical analysis on the influences of *size limited* fMRI data samples on the classification accuracy, based on the naive Bayesian classifier. More specifically, we show that as the number of data samples increases, the bound of error probability will decrease exponentially.

The major contributions of this paper can be summarized as:

- We construct the feature vectors out of real fMRI dataset using Linear Discriminant Analysis (LDA) [3], and carry out classification of AD patients and NC subjects based on the naive Bayesian classifier.
- We provide a theoretical analysis on how classification accuracy is influenced by sample size. It is shown that: due to the noise effect including both biological variability and measurement noise, when the naive Bayesian classifier is used, the upper bound of the error probability decreases exponentially as the sample size increases. This provides an estimation on the expected classification error probability for a given data sample size.

The rest of this paper is organized as follows. The framework of brain network connectivity pattern analysis is discussed in Section II. The Linear Discriminant Analysis, which is a prerequisite step for classifications, is briefed in Section III. In Section IV, we present the naive Bayesian approach and the proposed theoretical analysis on error probabilities of the Bayesian classifier. In Section V, we present the numerical results based on real fMRI data, and we conclude in Section VI.

Zhe Wang, Tianlong Song and, Yuan Liang and Tongtong Li are with the Department of Electrical and Computer Engineering, Michigan State University, East Lansing, MI, 48824, USA (E-mail: wangzh34@msu.edu, songtia6@msu.edu, liangy11@msu.edu, tongli@egr.msu.edu).

II. CLASSIFICATION BASED ON BRAIN CONNECTIVITY PATTERN

In this section, we present the problem of classification based on brain network connectivity analysis.

In fMRI based studies, it is a common practice to study multiple regions of interest (ROIs) instead of only one region. Regions within the ROI formulate a sub-network, and the network connectivity pattern analysis is then carried out by evaluating the correlation between all ROI pairs within the sub-network. The underlying argument is that: due to variability in the brain connectivity of each individual, the connectivity between two brain regions alone may not be sufficient to distinguish NC subjects from patients with cognitive impairments; brain network connectivity pattern analysis, which looks for subtle changes in the pattern of connectivity among multiple or all regions in the sub-network, may reveal more in-depth information.

The default mode network (DMN) is one of the well studied networks at the resting state [4]. Prior resting-state fMRI studies have demonstrated that the DMN is affected by AD [5]–[9]. Both hippocampus and ICC are part of the DMN, and can be well defined anatomically through the FreeSurfer software [9], even in brains with abnormal anatomy [8]. The paper by Zhu et al. [8] specifically demonstrated that the functional connection between hippocampus and ICC was decreased in AD.

Motivated by the observations above, in this paper, we select the right and left hippocampi and ICCs (4 regions) as our ROI sub-network. Our connectivity pattern analysis is carried out following the procedure below.

First, we calculate the Pearson correlation coefficients between all possible pairs of the ROIs within the group to formulate the feature vectors. As we now have 4 regions in the ROI sub-network, for each subject i, we can obtain a $d-$dimensional ($d = 6$) vector $\mathbf{v}_i$, consisting of the Pearson correlation coefficients for each pair of ROIs. When we have n_1 AD patients and n_2 normal control subjects, we get the feature vector set $V = \{\mathbf{v}_1, \cdots, \mathbf{v}_{n_1}, \mathbf{v}_{n_1+1}, \cdots, \mathbf{v}_{n_1+n_2}\}$.

Second, using the LDA, we map V to a one-dimensional subspace or axis, where the differences between AD and NC subjects are maximized, and denote the projected vectors as $\{x\} = \{x_1, \cdots, x_{n_1}, \cdots, x_{n_1+1}, \cdots, x_{n_1+n_2}\}$.

Finally, we carry out the classification using the naive Bayesian classifier based on the obtained $\{x\}$.

III. LINEAR DISCRIMINANT ANALYSIS

Linear Discriminant Analysis aims to separate two classes by projecting them into a subspace or direction where different classes show most significant differences [10]. Note that we have obtained a set of $d-$dimensional vector samples $V = \{\mathbf{v}_1, \cdots, \mathbf{v}_{n_1}, \mathbf{v}_{n_1+1}, \cdots, \mathbf{v}_{n_1+n_2}\}$, where n_1 of them are from the first class, denoted as C_1, and n_2 of them are from the second class, denoted as C_2. For $k = 1, 2$, the mean and scatter matrix (i.e., the scaled covariance matrix) of each of the two classes are defined as:

$$\boldsymbol{\mu}_k = \frac{1}{n_k} \sum_{\mathbf{v} \in C_k} \mathbf{v}, \tag{1}$$

$$S_k = \sum_{\mathbf{v} \in C_k} (\mathbf{v} - \boldsymbol{\mu}_k)(\mathbf{v} - \boldsymbol{\mu}_k)^t. \tag{2}$$

Consider the projection of vectors in V to a new $d-$dimensional space:

$$\mathbf{x} = W\mathbf{v}, \ \mathbf{v} \in V, \tag{3}$$

where W is a $d \times d$ matrix to be determined by the LDA algorithm. In this paper, we only utilize the first dimension x, of projected vector $\mathbf{x}$, where the differences among two classes are maximized. As a result, Equation (3) can be rewritten as:

$$x = \mathbf{w}^t \mathbf{v}, \tag{4}$$

Define $\boldsymbol{\mu} = \frac{1}{n_1+n_2} \sum_{i=1}^{n_1+n_2} \mathbf{v}_i$ as the overall mean, $S_W = \sum_{k=1}^{2} S_k$ as the within-class scatter matrix, and the between-class scatter matrix S_B as:

$$S_B = \sum_{k=1}^{2} n_k (\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})^t. \tag{5}$$

LDA seeks a transform vector $\mathbf{w}$ that maximizes the following objective function:

$$J(\mathbf{w}) = \frac{\mathbf{w}^t S_B \mathbf{w}}{\mathbf{w}^t S_W \mathbf{w}}. \tag{6}$$

It can be proved [3], [10] that to maximize Equation (6), $\mathbf{w}$ should satisfy

$$S_W^{-1} S_B \mathbf{w} = \lambda \mathbf{w}, \tag{7}$$

for some constant λ. Performing eigenvalue decomposition to matrix $S_W^{-1} S_B$, LDA then chooses the eigenvector corresponding to the largest eigenvalues of the matrix $S_W^{-1} S_B$ as $\mathbf{w}$. As will be shown in Section III, various classifiers, such as the Bayesian classifier can then be applied to the projected vectors $\{x_i = \mathbf{w}^t \mathbf{v}_i\}_{i=1}^{n_1+n_2}$ for further classification.

IV. INFLUENCE OF SAMPLE SIZE ON THE ACCURACY OF BAYESIAN CLASSIFICATION

In this section, we theoretically analyze the influence of sample size on the classification error probability. Suppose we have a set of normally distributed data samples $\{x\}$, where n_1 of them are from the first class, denoted as C_1, and n_2 of them are from the second class, denoted as C_2. For $i = 1, 2$, the mean of each class is denoted as μ_i. Without loss of generality, assume $\mu_1 < \mu_2$, and two classes of data samples have the same sample size and variance, i.e., $n_1 = n_2 = n$ and $\sigma_1^2 = \sigma_2^2 = \sigma_0^2$.

Consider the basic Bayesian classifier, which aims to find the decision regions by calculating the boundary points b. More specifically, for any given $x_i, i = 1, 2, ..., 2n$: If $x_i < b$, then $x_i \in C_1$; otherwise $x_i \in C_2$. To calculate b, suppose y is a random variable, C_y the corresponding class, and $\tilde{C}_y$ the class assigned by the classifier. Set

$$P(\tilde{C}_y = C_1) = P(\tilde{C}_y = C_2). \tag{8}$$

That is,

$$\int_{-\infty}^{b} \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{(y-\mu_1)^2}{2\sigma_0^2}} dy = \int_{b}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{(y-\mu_2)^2}{2\sigma_0^2}} dy. \tag{9}$$

Because the Gaussian probability density function is symmetric, Equation (9) can be rewritten as:

$$\int_{-\infty}^{b} \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{(y-\mu_1)^2}{2\sigma_0^2}} dy = \int_{-\infty}^{2\mu_2 - b} \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{(y-\mu_2)^2}{2\sigma_0^2}} dy. \tag{10}$$

Let $u = (y - \mu_1)/\sigma_0$, and $v = (y - \mu_2)/\sigma_0$, Equation (10) can be further simplified as:

$$\int_{-\infty}^{\frac{b-\mu_1}{\sigma_0}} \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{u^2}{2}} du = \int_{-\infty}^{\frac{\mu_2 - b}{\sigma_0}} \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{v^2}{2}} dv. \tag{11}$$

Letting $(b - \mu_1)/\sigma_0 = (\mu_2 - b)/\sigma_0$, the final solution can be derived as $b = (\mu_1 + \mu_2)/2$. That is, the classifier simply compares the Euclidean distances of a data point to the center of two classes and assigns it to the nearest neighbor.

The probability of the error that the random variable y is incorrectly classified by the Bayesian classifier is:

$$\begin{aligned} P_{err} &= P(C_y = C_1)P(\tilde{C}_y \neq C_1 | C_y = C_1) \\ &\quad + P(C_y = C_2)P(\tilde{C}_y \neq C_2 | C_y = C_2) \\ &= \frac{1}{2} \int_{b}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{(y-\mu_1)^2}{2\sigma_0^2}} dy \\ &\quad + \frac{1}{2} \int_{-\infty}^{b} \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{(y-\mu_2)^2}{2\sigma_0^2}} dy. \end{aligned} \tag{12}$$

When $b = (\mu_1 + \mu_2)/2$, the error probability P_{err} can be minimized. In real applications, however, μ_1 and μ_2 are not known, and the means in Equation (12) will be replaced with the estimated averages $\hat{\mu}_i = \frac{1}{n} \sum_{x \in C_i} x, i = 1, 2$. In this case, the calculated boundary $\hat{b} = (\hat{\mu}_1 + \hat{\mu}_2)/2$. Recall that the training data set $\{x\}$ are assumed to be normally distributed with variance σ_0^2, we can know that for $i = 1, 2$, $\hat{\mu}_i$ is a Gaussian random variable with mean μ_i and variance $\sigma^2 = \sigma_0^2/n$. As a result, an extra error probability P_{oe} will be introduced into P_{err} because of inaccurate estimation of $\hat{b}$. Without losing generality, assume $\hat{b} > b$, we have:

$$\begin{aligned} P_{oe} &= \int_{b}^{\hat{b}} \frac{1}{\sqrt{2\pi}\sigma_0} [e^{-\frac{(y-\mu_2)^2}{2\sigma_0^2}} - e^{-\frac{(y-\mu_1)^2}{2\sigma_0^2}}] dy \\ &= \int_{0}^{e} \frac{1}{\sqrt{2\pi}\sigma_0} [e^{-\frac{(z-d')^2}{2\sigma_0^2}} - e^{-\frac{(z+d')^2}{2\sigma_0^2}}] dz \\ &= \int_{0}^{e} g(z) dz, \end{aligned} \tag{13}$$

where $z = y - b, e = \hat{b} - b$, $d' = (\mu_2 - \mu_1)/2$ and $g(z) = \frac{1}{\sqrt{2\pi}\sigma_0} [e^{-\frac{(z-d')^2}{2\sigma_0^2}} - e^{-\frac{(z+d')^2}{2\sigma_0^2}}]$. Since $\hat{\mu}_i, i = 1, 2$ are normally distributed with variance σ^2, e will also be normally distributed with mean 0 and variance $\sigma^2 = \sigma_0^2/n$. Hence the mean of the extra error probability $P_e(n)$ can be calculated as:

$$\begin{aligned} P_e(n) &= \int_{0}^{\infty} P_{oe} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{e^2}{2\sigma^2}} de \\ &= \int_{0}^{\infty} \int_{0}^{e} g(z) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{e^2}{2\sigma^2}} dz de \\ &= \int_{0}^{\infty} \int_{z}^{\infty} g(z) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{e^2}{2\sigma^2}} de dz \\ &= \int_{0}^{\infty} g(z) \int_{\frac{z}{\sigma}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{e'^2}{2}} de' dz \\ &= \int_{0}^{\infty} g(z) Q(\frac{z}{\sigma}) dz \\ &= \int_{0}^{\infty} g(z) Q(\frac{\sqrt{n}z}{\sigma_0}) dz, \end{aligned} \tag{14}$$

where $e' = e/\sigma$, and Q function is the tail probability of the standard normal distribution. It can be seen from Equation (14) that, because the Q function is always monotonically decreasing withe respect to $\frac{\sqrt{n}z}{\sigma_0}$, for every z, when the sample size n increases, $Q(\frac{\sqrt{n}z}{\sigma_0})$ will decrease, and so is P_e as well.

The final classification error probability $P(n)$ is then the sum of P_{err} and $P_e(n)$, i.e.,

$$P(n) = P_{err} + P_e(n). \tag{15}$$

With Equation (15) and (14), we have the following results:

Proposition 1 Given a data sample size n, the corresponding error probability $P(n)$ decreases monotonically with n, and is bounded by P_{err}, i.e., $P(n) \geq P_{err}$. That is, $\lim_{n \to \infty} P(n) = P_c$, where P_c is a constant.

Upper Bound of Error Probability The error probability P_{err} in Equation (12) is upper bounded by the Bhattacharyya Bound [3]:

$$P_{err} \leq \frac{1}{2} e^{-\frac{(\mu_2 - \mu_1)^2}{8\sigma_0^2}} = \frac{1}{2} e^{-\frac{\Delta^2}{8\sigma_0^2}}, \tag{16}$$

in which $\Delta = \mu_2 - \mu_1$ is a constant number.

When the ensemble means in Equation (16) are replaced with the estimated averages, the constant number Δ becomes

a random variable $\hat{\Delta}$:

$$\begin{aligned}\hat{\Delta} &= \hat{\mu}_2 - \hat{\mu}_1 \\ &= \mu_2 - \mu_1 - [(\hat{\mu}_1 - \mu_1) - (\hat{\mu}_2 - \mu_2)] \\ &= \Delta - s, \end{aligned} \tag{17}$$

where $s = (\hat{\mu}_1 - \mu_1) - (\hat{\mu}_2 - \mu_2)$ is the skew introduced by the estimated averages. In this case, the corresponding Bhattacharyya Bound $B(s)$ can be roughly approximated as:

$$B(s) = \frac{1}{2} e^{-\frac{(\Delta - s)^2}{8\sigma_0^2}}. \tag{18}$$

Since for $i = 1, 2$, $\hat{\mu}_i$ is a Gaussian random variable with mean μ_i and variance $\sigma^2 = \sigma_0^2/n$, based on the properties of mean and variance, we can know that s is also a Gaussian random variable with mean 0 and variance $\sigma_s^2 = 2\sigma^2 = 2\sigma_0^2/n$. As a result, the expectation of the Bhattacharyya Bound B can be roughly approximated as:

$$\begin{aligned} B &= \int_{-\infty}^{+\infty} B(s) \frac{1}{\sqrt{2\pi}\sigma_s} e^{-\frac{s^2}{2\sigma_s^2}} ds \\ &= \int_{-\infty}^{+\infty} \frac{1}{2} e^{-\frac{(\Delta - s)^2}{8\sigma_0^2}} \frac{1}{\sqrt{2\pi}\sigma_s} e^{-\frac{s^2}{2\sigma_s^2}} ds \\ &= \frac{1}{2}\sqrt{\frac{4\sigma_0^2}{4\sigma_0^2 + \sigma_s^2}} e^{-\frac{\Delta^2}{8\sigma_0^2}\sqrt{\frac{4\sigma_0^2}{4\sigma_0^2+\sigma_s^2}}} \\ &= \frac{1}{2}\sqrt{\frac{2n}{2n+1}} e^{-\frac{\Delta^2}{8\sigma_0^2}\sqrt{\frac{2n}{2n+1}}}. \end{aligned} \tag{19}$$

It can be seen from Equation (19) that the bound of the average estimated error probability will decrease monotonically as sample size n increases. That means, to achieve a satisfying classification accuracy, the sample size should be as large as possible. This provides an estimation on the expected classification error probability for a given data sample size.

V. NUMERICAL ANALYSIS

In this section, we provide numerical results Bayesian classifications based on brain connectivity pattern analysis, which confirms the influence of sample size on the classification accuracy.

In our data collection process, 10 patients with mild-to-moderate probable Alzheimer's Disease and 12 age- and education-matched healthy NC subjects were recruited to participate in this study. The fMRI experiment was conducted on a GE 3T $Signa^{\textcircled{R}}$ HDx MR scanner (GE Healthcare, Waukesha, WI) with an 8-channel head coil. To study resting-state brain function, echo-planar images, starting from the most inferior regions of the brain, were acquired for 7 minutes with the following parameters: 38 contiguous $3mm$ axial slices in an interleaved order, time of echo = $27.7ms$, time of repetition = $2500ms$, flip angle = 80°, field of view = $22cm$, matrix size = 64×64, ramp sampling, and with the first four data points discarded. Each volume of slices was acquired 164 times. Common pre-processing procedures on resting state fMRI data were carried as detailed in [11].

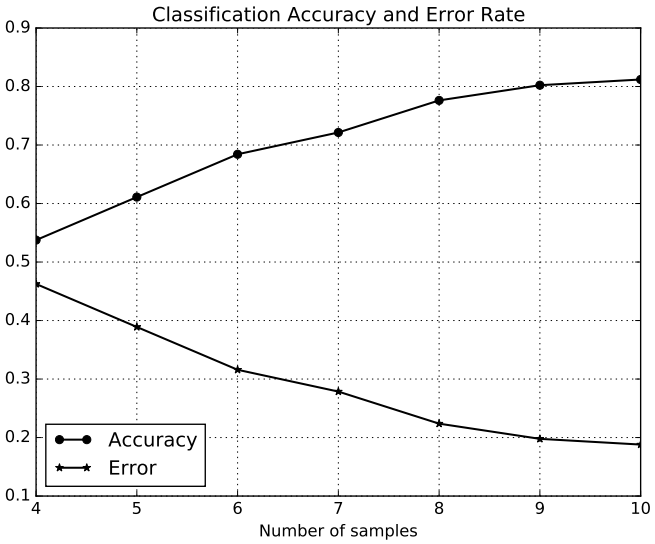

Fig. 1: Classification accuracies and error probabilities with respect to the sample size.

In the simulations, we vary the sample size of each subject group from 4 to 10. Since the size of data samples is small, the performance of the classifier is evaluated by the Leave-One-Out (LOO) cross-validation. Figure 1 shows the classification accuracies and error probabilities of the Bayesian classifier with respect to the sample size. It can be seen that when the sample size $n = 4$, the classification accuracy is as low as 54%, which is slightly higher than that of random guess. As n increases, the accuracy is increased as well. When the size $n = 10$, the accuracy is increased to be higher than 80%. This provides an estimation on the expected classification error probability for a given data sample size.

VI. CONCLUSIONS

In this paper, we analyzed the influence of sample sizes on the classification accuracies and error probabilities in the brain connectivity pattern analysis. Both theoretical and numerical analyses showed that: as the sample size increases, the errors caused by inaccurate estimation of optimal decision bound of the Bayesian classifier and the upper error bound will be reduced.

VII. REFERENCES

[1] K. Wang *et al.*, "Discriminative analysis of early Alzheimers disease based on two intrinsically anti-correlated networks with resting-state fMRI,"

Medical Image Computing and Computer-Assisted Intervention–MICCAI 2006, pp. 340–347, 2006.

[2] G. Chen *et al.*, "Classification of Alzheimer disease, mild cognitive impairment, and normal cognitive status with large-scale network analysis based on resting-state functional MR imaging," *Radiology*, vol. 259, no. 1, pp. 213–221, 2011.

[3] R. O. Duda *et al.*, *Pattern classification*. John Wiley & Sons, 2012.

[4] B. Yeo *et al.*, "The organization of the human cerebran cortex estimated by intrinsic functional connection," *J Neurophysiol*, vol. 106, pp. 1125–1165, June 2011.

[5] M. A. Binnewijzend *et al.*, "Resting-state fmri changes in alzheimer's disease and mild cognitive impairment," *Neurobiology of aging*, vol. 33, no. 9, pp. 2018–2028, 2012.

[6] M. D. Greicius *et al.*, "Default-mode network activity distinguishes alzheimer's disease from healthy aging: evidence from functional mri," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 13, pp. 4637–4642, 2004.

[7] H.-Y. Zhang *et al.*, "Resting brain connectivity: Changes during the progress of alzheimer disease 1," *Radiology*, vol. 256, no. 2, pp. 598–606, 2010.

[8] D. C. Zhu *et al.*, "Alzheimer's disease and amnestic mild cognitive impairment weaken connections within the default-mode network: a multi-modal imaging study," *Journal of Alzheimer's Disease*, vol. 34, no. 4, pp. 969–984, 2013.

[9] B. Fischl and thers, "Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain," *Neuron*, vol. 33, no. 3, pp. 341–355, 2002.

[10] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[11] D. C. Zhu *et al.*, "Alzheimer's disease and amnestic mild cognitive impairment weaken connections within the default-mode network: a multi-modal imaging study," *Journal of Alzheimer's Disease*, vol. 34, no. 4, pp. 969–984, 2013.