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

Presenter: Yuan Liang

Department of Electrical & Computer Engineering

Michigan State University

IEEE GlobalSIP 2016, Greater Washington, D.C., USA

Introduction

- fMRI based classification of Alzheimer's Disease (AD) and normal control (NC) subjects is beneficial for early diagnosis and treatment of brain disorders [1, 2].
- The size of fMRI data samples is generally quite limited, which has become a major bottleneck. Most existing classifiers could potentially suffer from noise effects, due to both biological variability and measurement noise.
- 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.

Brain Connectivity Pattern Classification

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.

Major Procedure

- 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.

  – Pearson correlation coefficients between all possible pairs of the ROIs within the group to formulate the feature vectors.

  – Dimensionality reduction using the Linear Discriminant Analysis.

  – Classification using the naive Bayesian classifier.

Linear Discriminant Analysis

- Linear Discriminant Analysis aims to separate two classes by projecting them into a subspace where different classes show most significant differences [3].

- Given a set of $d-$dimensional vector samples $V = \{\mathbf{v}_1, \cdots, \mathbf{v}_{n_1}, \mathbf{v}_{n_1+1}, \cdots, \mathbf{v}_{n_1+n_2}\}$, consider the projection of vectors in V to a new $1-$dimensional space:

$$x = \mathbf{w}^t \mathbf{v}, \tag{1}$$

  where w is a $d \times 1$ matrix to be determined by the LDA algorithm.

- After projection, 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.

Influence of Sample Size on The Accuracy of Bayesian Classification

- Suppose we have a set of normally distributed data samples $\{x\}$, where n of them are from the first class C_1, and n of them are from the second class C_2. Assume $\mu_1 < \mu_2$ and $\sigma_1^2 = \sigma_2^2 = \sigma_0^2$.
- The basic Bayesian classifier aims to find the decision regions by calculating the boundary points $b = (\mu_1 + \mu_2)/2$. The probability of the error that the random variable y is incorrectly classified by the Bayesian classifier is:

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

- In real applications, μ_i and b will be replaced with the estimated values $\hat{\mu}_i$ and $\hat{b}$. Hence an extra error probability will be introduced:

$$P_{oe} = \int_b^{\hat{b}} \frac{1}{\sqrt{2\pi}\sigma_0}[e^{-\frac{(y-\mu_2)^2}{2\sigma_0^2}} - e^{-\frac{(y-\mu_1)^2}{2\sigma_0^2}}]dy = \int_0^e g(z)dz, \quad (3)$$

where $z = y-b, e = \hat{b}-b$, $d' = (\mu_2-\mu_1)/2$, $g(z) = \frac{1}{\sqrt{2\pi}\sigma_0}[e^{-\frac{(z-d')^2}{2\sigma_0^2}} - e^{-\frac{(z+d')^2}{2\sigma_0^2}}]$.

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

$$P(n) = P_{err} + P_e(n), \quad (4)$$

where $P_e(n)$ is the mean of the extra error probability P_{oe}.

Monotonic Analysis

- 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 $P_e(n)$ can be calculated as:

$$P_e(n) \quad = \quad \int_0^\infty P_{oe} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{e^2}{2\sigma^2}} de = \int_0^\infty g(z) Q(\frac{\sqrt{n}z}{\sigma_0}) dz, \tag{5}$$

where $e' = e/\sigma$, and Q function is the tail probability of the standard normal distribution.

- 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.

Upper Bound of Error Probability

- The error probability P_{err} is upper bounded by [4]:

$$P_{err} \leq \frac{1}{2} e^{-\frac{(\mu_2-\mu_1)^2}{8\sigma_0^2}} = \frac{1}{2} e^{-\frac{\Delta^2}{8\sigma_0^2}}. \tag{6}$$

- When μ_i is replaced with $\hat{\mu}_i$, Δ will be replaced by $\hat{\Delta}$:

$$\hat{\Delta} \;=\; \hat{\mu}_2 - \hat{\mu}_1 = \mu_2 - \mu_1 - [(\hat{\mu}_1 - \mu_1) - (\hat{\mu}_2 - \mu_2)] = \Delta - s, \tag{7}$$

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 upper bound $B(s)$ can be roughly approximated as:

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

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

$$B = \int_{-\infty}^{+\infty} B(s) \frac{1}{\sqrt{2\pi}\sigma_s} e^{-\frac{s^2}{2\sigma_s^2}} ds = \frac{1}{2}\sqrt{\frac{2n}{2n+1}} e^{-\frac{\Delta^2}{8\sigma_0^2}\sqrt{\frac{2n}{2n+1}}}. \tag{9}$$

- It can be seen from Equation (9) that the bound of the average estimated error probability will decrease monotonically as sample size n increases.

Numerical Results

- 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.
- 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.

- When the sample size $n = 4$, the classification accuracy is as low as 54%, which is slightly higher than that of random guess; and 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.

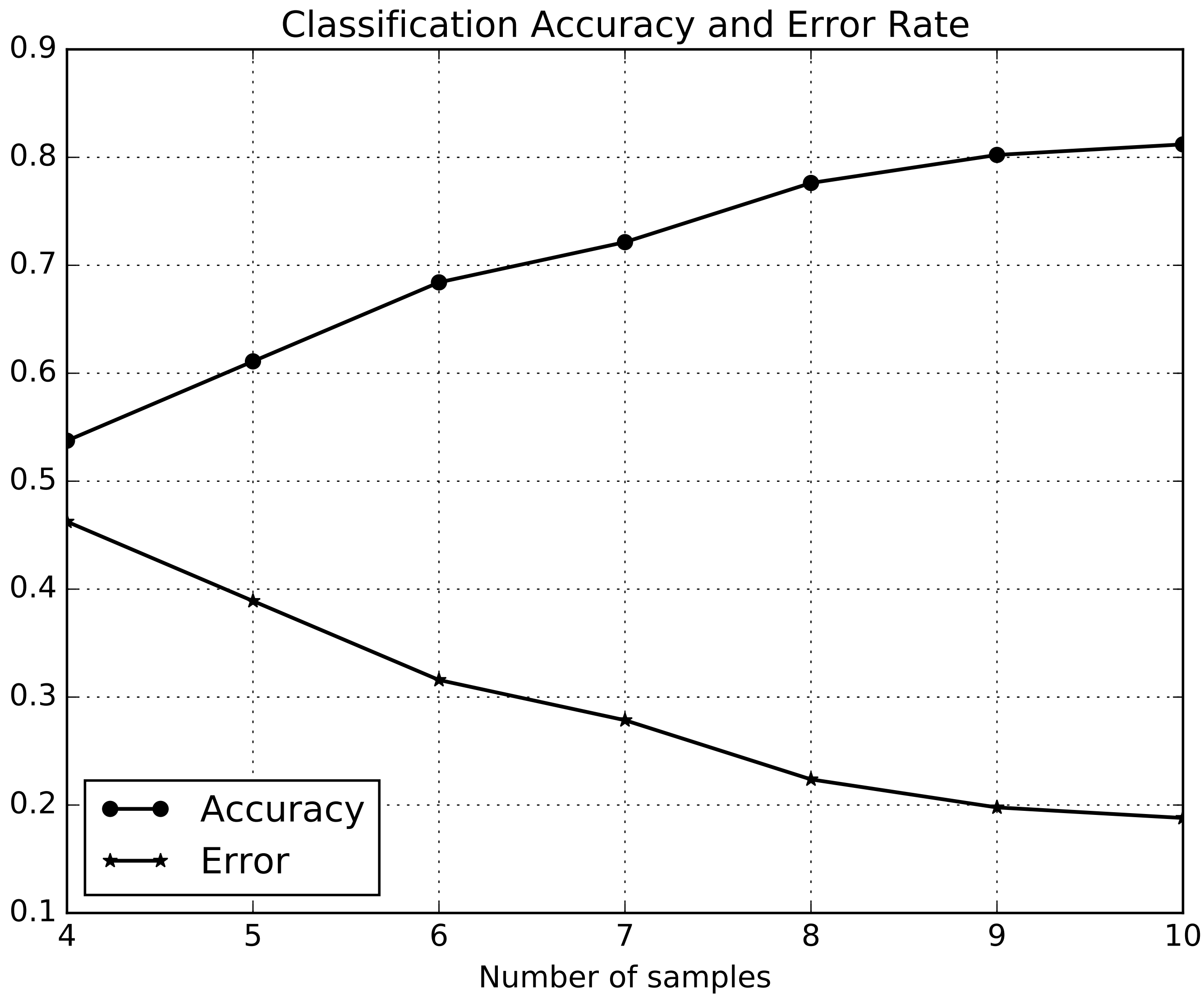

Figure 1: Classification accuracies and error probabilities with respect to the sample size.

Conclusion

- 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.

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

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

[5] 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.