PHÂN TÍCH KHÁC BIỆT CHO TÍN HIỆU TỪ NÃO BỘ

Nguyễn Hoàng Huy*, Hoàng Thị Thanh Giang
Khoa Công nghệ Thông tin - Trường Đại học Nông nghiệp Hà nội
Email: nhhuy@hua.edu.vn*

TÓM TẮT

Chúng tôi đã đưa ra một hướng tiếp cận đa bước trong học máy và áp dụng nó để phân loại dữ liệu từ những giao diện máy tính-bộ não dựa trên nền tảng sóng điện não, tham khảo Huy, 2013. Hướng tiếp cận này rất hiệu quả cho những kiểu dữ liệu sóng điện não trên, tham khảo Huy và đồng nghiệp, 2012; Huy, 2013. Đầu tiên tất cả các thuộc tính được phân chia thành các nhóm con và phân tích khác biệt tuyến tính của Fisher (LDA) được sử dụng tính điểm cho mỗi nhóm thuộc tính. Tiếp theo LDA được áp dụng cho các nhóm con của các điểm vừa thu được. Quá trình này được lặp lại cho đến khi còn lại một điểm duy nhất, điểm này được sử dụng cho phân loại. Bằng cách này chúng tôi tránh phải ước lượng ma trận hiệp phương sai số chiều lớn. Chúng tôi gọi phương pháp trên là phân tích khác biệt tuyến tính đa bước (multi-step LDA). Đối với mô hình chuẩn, chúng tôi đã nghiên cứu tiệm cận của sai lầm phân loại khi số chiều d và dung lượng mẫu n tiến đến vô cùng, tham khảo Huy, 2013. Điều này chỉ ra cách xác định cỡ của các nhóm ở mỗi bước. Thêm nữa chúng tôi đưa ra cận trên của độ sai lầm phân loại lý thuyết đối với mô hình không gian-thời gian chuẩn với ma trận hiệp phương sai có thể tách được, kết quả này gợi ý cách nhóm các thuộc tính hoặc điểm cho loại dữ liệu này. Trong báo cáo này, chúng tôi kiểm tra hiệu suất phân loại của multi-step LDA với sự chú ý đặc biệt tới dữ liệu có dung lượng mẫu nhỏ. Đối với tín hiệu não bộ của một thí nghiệm giao diện máy tính-bộ não bởi Frenzel và đồng nghiệp, 2011, multi-step LDA thể hiện tốt hơn phân tích khác biệt tuyến tính được chính quy hóa (regularized LDA), phương pháp phân loại tiên tiến nhất cho kiểu dữ liệu này, tham khảo Blankertz và đồng nghiệp, 2011.

Discriminant Analysis for Brain Signals

ABSTRACT

We introduced a multi-step machine learning approach and used it to classify data from EEG-based brain-computer interfaces, see Huy, 2013. This approach works very well for the above kinds of EEG data, see Huy et al., 2012; Huy, 2013. First all features are divided into subgroups and Fisher's linear discriminant analysis (LDA) is used to obtain a score for each subgroup. Then it is applied to subgroups of the resulting scores. This procedure is iterated until there is only one score remaining and this one is used for classification. In this way we avoid estimation of the high-dimensional covariance matrix of all features. We call the above method multi-step linear discriminant analysis (multi-step LDA). For the normal model, we studied the asymptotic error rate when dimension d and sample size n tend to infinity, see Huy, 2013. This indicates how to define the sizes of subgroups at each step. In addition we presented a theoretical error bound for the spatio-temporal normal model with separable covariance matrix, see Huy et al., 2012, which results in a recommendation on how subgroups should be formed

for this kind of data. In this report, we investigate the classification performance with special attention to the small sample size case. For brain signals of the brain-computer interface experiment by Frenzel et al., 2011, the multi-step LDA performs better than regularized linear discriminant analysis (regularized LDA), which is the state-of-the-art classification method for this kind of data, see Blankertz et al., 2011.

1 ĐẶT VẤN ĐỀ

"Ngày nay dữ liệu số chiều lớn xuất hiện trong hầu khắp các lĩnh vực như công nghệ thông tin, tin sinh, thiên văn học, ...", tham khảo Bühlmann và van de Geer, 2011. Đối với những loại dữ liệu này, số thuộc tính thường lớn hơn rất nhiều dung lượng mẫu. Trong báo cáo này chúng tôi tập trung vào dữ liệu sóng điện não số chiều lớn từ những giao diện máy tính-bộ não. Đích đến thật sự của những giao diện máy tính-bộ não là phân loại sóng điện não tương ứng với trạng thái của não bộ. Tuy nhiên "the curse of dimensionality" làm vấn đề này trở nên rất phức tạp, tham khảo Lotte và đồng nghiệp, 2007.

Krusienski và đồng nghiệp, 2008 đã chỉ ra rằng các phương pháp phân loại tuyến tính là đủ và sự thêm vào của những phương pháp phi tuyến là không cần thiết. Xu hướng chung cũng đề cao những phương pháp phân loại đơn giản như phân tích khác biệt tuyến tính cổ điển của Fisher hơn những phương pháp phức tạp, tham khảo Nicolas-Alonso và Gomez-Gil, 2012. Trong thực tế, LDA vẫn là một trong những phương pháp phân loại dữ liệu được sử dụng rộng rãi nhất. Đối với hai phân phối chuẩn có cùng ma trận hiệp phương sai $\mathbf{\Sigma}$ và vector trung bình khác nhau $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$, LDA cực tiểu hóa độ sai lầm phân loại, tham khảo Mclachlan, 1992; Hastie và đồng nghiệp, 2009. Điểm hay giá trị hàm phân loại δ_F của LDA đối với quan sát $\boldsymbol{X}$ được xác định bởi

$$\delta_F(\boldsymbol{X}) = (\boldsymbol{X} - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} \boldsymbol{\alpha} \text{ với } \boldsymbol{\alpha} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \text{ và } \boldsymbol{\mu} = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2).$$

Trong thực tiễn chúng ta thường không biết $\mathbf{\Sigma}$ và $\boldsymbol{\mu}_i$, do đó phải ước lượng chúng từ dữ liệu huấn luyện. Phương pháp này rất hiệu quả cho dữ liệu số chiều nhỏ, nhưng sự ước lượng $\mathbf{\Sigma}$ trở nên thực sự khó khăn cho dữ liệu số chiều lớn, tham khảo Ledoit và Wolf, 2002. Khi dung lượng mẫu n của dữ liệu huấn luyện nhỏ hơn số thuộc tính d thì ma trận hiệp phương sai mẫu cấp $d \times d$, $\hat{\mathbf{\Sigma}}$ không khả nghịch. Giả nghịch đảo Moore-Penrose có thể được sử dụng nhưng điều này sẽ làm suy yếu phân loại. Thậm chí khi n lớn hơn, nhưng có cùng độ lớn như d, sai số cộng dồn khi ước lượng rất nhiều phần tử của ma trận hiệp phương sai sẽ làm tăng đáng kể độ sai lầm của LDA. Đối với dữ liệu sóng điện não trong một thí nghiệm về giao diện máy tính-bộ não bởi Frenzel và đồng nghiệp, 2011, ở đó $100 \leq n \leq 250$ và $d = 1024$, những điều này sẽ được thể hiện rõ trong mục 3.

Một phương pháp có thể khắc phục vấn đề trên là regularized LDA, ở đó bội số của ma trận đơn vị $\boldsymbol{I}$ được cộng thêm vào ma trận hiệp phương sai mẫu, tham khảo Friedman, 1989. $\hat{\mathbf{\Sigma}} + r\boldsymbol{I}$ khả nghịch đối với mọi $r > 0$. Tuy nhiên để xác định tham số r tốt nhất, chúng ta có thể phải áp dụng những thuật toán tối ưu hóa phức tạp.

Bickel và Levina, 2004 đề nghị một giải pháp đơn giản hơn: bỏ qua sự tương quan giữa các thuộc tính và sử dụng ma trận chéo $\boldsymbol{D}_{\boldsymbol{\Sigma}} = \text{diag}(\boldsymbol{\Sigma})$ của $\boldsymbol{\Sigma}$ thay cho ma trận hiệp phương sai $\boldsymbol{\Sigma}$. Phương pháp này được gọi là Independence Rule. Hàm phân loại δ_I của nó được cho bởi

$$\delta_I(\boldsymbol{X}) = (\boldsymbol{X} - \boldsymbol{\mu})^T \boldsymbol{D}_{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\alpha}.$$

Tuy nhiên ngay cả với Independence Rule sử dụng tất cả các thuộc tính có thể tồi như dự đoán ngẫu nhiên do sự cộng dồn sai số khi ước lượng nhiều phần tử của vector trung bình, tham khảo Fan và Fan, 2008. Vì vậy Fan và Lv, 2010; Guo, 2010 đã đưa ra một vài phương pháp phân loại trước hết lựa chọn những thuộc tính có ảnh hưởng qua giá trị trung bình rồi áp dụng Independence Rule cho những thuộc tính vừa lựa chọn đó. Nhưng phương pháp này không cực tiểu hóa độ sai lầm phân loại do bỏ qua sự tương quan của các thuộc tính. Khi lựa chọn những thuộc tính cho phân loại, chúng ta không chỉ phải xác định những thuộc tính có ảnh hưởng qua giá trị trung bình mà còn những thuộc tính có ảnh hưởng nhờ sự tương quan của chúng đối với thuộc tính khác, tham khảo Zhang và Wang, 2011. Đây là vấn đề rất khó giải quyết khi d lớn hơn rất nhiều so với n.

Huy và đồng nghiệp, 2012; Huy, 2013 đã giới thiệu một giải pháp khác. Giải pháp này sử dụng một vài chứ không phải tất cả các hệ số tương quan của các thuộc tính và nó rất hiệu quả cho dữ liệu sóng điện não số chiều lớn từ những giao diện máy tính-bộ não. Trong báo cáo này chúng tôi thực hiện một số kiểm định thống kê cho nhận định trên.

2 PHƯƠNG PHÁP MULTI-STEP LDA

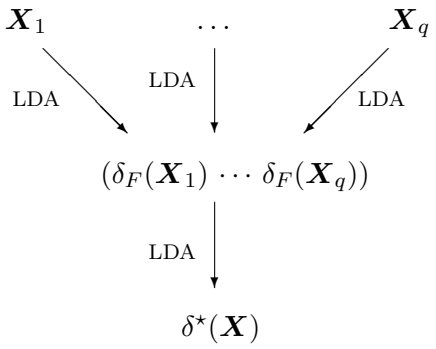
Chúng tôi đã đưa ra phân tích khác biệt tuyến tính đa bước (multi-step LDA), phương pháp này áp dụng LDA trong vài bước thay cho áp dụng nó một lần cho tất cả các thuộc tính. Để trình bày phương pháp rõ ràng nhưng không mất tính tổng quát chúng ta bắt đầu với phân tích khác biệt tuyến tính hai bước (two-step LDA). Tất cả d thuộc tính của một quan sát $\boldsymbol{X} \in \mathbb{R}^d$ được chia thành các nhóm con rời nhau

$$\boldsymbol{X} = \left[\boldsymbol{X}_1^T, \cdots, \boldsymbol{X}_q^T\right]^T,$$

ở đó $\boldsymbol{X}_j \in \mathbb{R}^p$, và $pq = d$. LDA được áp dụng để tính điểm cho mỗi nhóm con các thuộc tính. Trong bước thứ hai, LDA lại được áp dụng cho những điểm này để đưa ra điểm toàn bộ sử dụng cho phân loại. Do đó hàm phân biệt của two-step LDA là

$$\delta^\star(\boldsymbol{X}) = \delta_F(\delta_F(\boldsymbol{X}_1), \cdots, \delta_F(\boldsymbol{X}_q)),$$

ở đó δ_F biểu diễn hàm LDA. Hình 1 minh họa thủ tục two-step LDA. Giả định về tính chuẩn cần cho LDA sẽ thỏa mãn trong bước thứ hai. Phân phối của điểm có thể tính được bằng cách áp dụng đại số tuyến tính và tính chất của phân phối chuẩn nhiều chiều.



Hình 1: Giản đồ của two-step LDA.

Mệnh đề 1. *Giả sử* $\boldsymbol{X}$ *có phân phối chuẩn với tham số* $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ *và* $\boldsymbol{\Sigma}$. *Để* $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1 = (\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_q^T)^T$ *và* $\boldsymbol{\Sigma}_{ij} \in \mathbb{R}^{p\times p}$ *là ma trận con của* $\boldsymbol{\Sigma}$ *tương ứng với hai nhóm con* i *và* j *sao cho* $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_{ij})_{i,j=1}^q$. *Khi đó các điểm* $(\delta_F(\boldsymbol{X}_1), \ldots, \delta_F(\boldsymbol{X}_q))^T$ *cũng có phân phối chuẩn với cùng ma trận hiệp phương sai* $\boldsymbol{\Theta}$ *và vector trung bình* $\pm(1/2)\boldsymbol{m}$ *cho bởi*

$$\Theta_{ij} = \boldsymbol{\alpha}_i^T \boldsymbol{\Sigma}_{ii}^{-1} \boldsymbol{\Sigma}_{ij} \boldsymbol{\Sigma}_{jj}^{-1} \boldsymbol{\alpha}_j, \;\; m_i = \Theta_{ii}, \;\; \textit{với } i, j = 1, \ldots, q.$$

Multi-step LDA sẽ lặp lại mỗi bước như đã được miêu tả trong two-step LDA cho đến khi còn một điểm duy nhất và điểm này sẽ được dùng phân loại cho tất cả các thuộc tính. Trong khuôn khổ báo cáo này, những thủ tục multi-step LDA chia tất cả các thuộc tính hay điểm thành những nhóm con liên tiếp, rời nhau ở mỗi bước. Những nhóm con này có cùng dung lượng. Dung lượng của các nhóm con ở bước thứ i, $i = 1, \ldots, l$ được xác định là phần tử thứ i của vector $\boldsymbol{t} = (p_1, \ldots p_l)$ và chúng ta giả sử $\prod_{i=1}^l p_i = d$ được thỏa mãn. Vector $\boldsymbol{t}$ sẽ được gọi là kiểu của multi-step LDA. Trong báo cáo này chúng tôi sẽ gọi thủ tục two-step LDA, multi-step LDA, LDA, hay regularized LDA mẫu cho những thủ tục vừa kể trên với các tham số được ước lượng từ dữ liệu huấn luyện.

3 KẾT QUẢ VÀ THẢO LUẬN

Ở mục này chúng tôi áp dụng multi-step LDA và các phương pháp phân loại tiên tiến khác để phân loại sóng điện não thành hai lớp tương ứng với trạng thái của não bộ. Dữ liệu sóng điện não được chúng tôi sử dụng trong báo cáo này là của một thí nghiệm giao diện máy tính-bộ não bởi Frenzel và đồng nghiệp, 2011. Một giao diện máy tính-bộ não có thể giúp người sử dụng điều khiển các thiết bị điện tử hoặc máy tính bằng cách chỉ sử dụng hoạt động của não bộ. Hệ thống này nhận ra ý định của người sử dụng dựa trên dữ liệu sóng điện não quan sát được. Do tiến bộ của khoa học kỹ thuật, số giá trị sóng điện não (hiệu điện thế đo được trên da đầu) hay số chiều của mỗi quan sát $\boldsymbol{X}$ thường rất lớn ($d \approx 30000$). Sau khi áp dụng những phương pháp tiền xử lý số liệu tiêu chuẩn cho kiểu dữ liệu này số chiều d vẫn còn rất lớn ($d = 1024$).

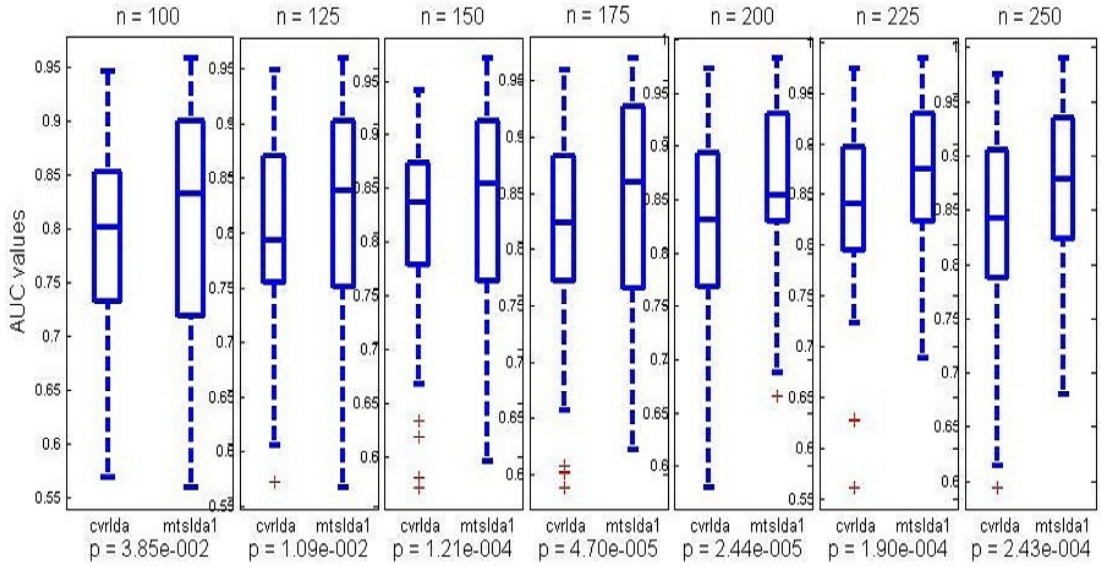
n	100	125	150	175	200	225	250
lda	0.7704	0.7723	0.7824	0.7918	0.8010	0.8132	0.8157
cvrlda	0.7893	0.8018	0.8104	0.8130	0.8218	0.8347	0.8386
oprlda	0.7823	0.7923	0.8030	0.8094	0.8162	0.8260	0.8296
tslda	0.7467	0.7767	0.7959	0.8100	0.8223	0.8367	0.8471
mtslda1	0.8064	0.8223	0.8356	0.8434	0.8556	0.8620	0.8647
mtslda2	0.8079	0.8190	0.8281	0.8367	0.8414	0.8440	0.8500
mtslda3	0.8076	0.8192	0.8326	0.8414	0.8461	0.8516	0.8557
mtslda4	0.8207	0.8309	0.8440	0.8489	0.8580	0.8646	0.8686
mtslda5	0.7872	0.8079	0.8237	0.8306	0.8418	0.8503	0.8531

Bảng 1: Trung bình giá trị AUC của LDA, regularized LDA, two-step LDA và multi-step LDA qua 42 tập dữ liệu.

Trong khi đó do yêu cầu của việc phân loại trực tuyến, dung lượng của dữ liệu huấn luyện n càng nhỏ càng tốt. Đối với dữ liệu trên, n chỉ cỡ hàng trăm.

Khi ứng dụng multi-step LDA chúng ta sẽ gặp vấn đề chia các thuộc tính ở mỗi bước như thế nào để nhận được kết quả tốt. Huy, 2013 đã chỉ ra rằng độ sai lầm của two-step LDA, multi-step LDA mẫu sẽ xấp xỉ độ sai lầm lý thuyết khi cỡ của mỗi nhóm thuộc tính $p_i, i = 1, \ldots, l$ là nhỏ so với dung lượng mẫu của dữ liệu huấn luyện n. Cụ thể bậc độ nhỏ của p_i so với n như thế nào được thảo luận tỉ mỉ trong Huy, 2013. Hơn nữa kiểu dữ liệu trên được cho là tuân theo mô hình không gian-thời gian với ma trận hiệp phương sai có thể tách được, tham khảo Mitchell và đồng nghiệp, 2006; Genton, 2007. Huy và đồng nghiệp, 2012 chỉ ra rằng, đối với kiểu dữ liệu đó, chúng ta nên nhóm các thuộc tính tương ứng với thời điểm khác nhau vào các nhóm. Còn trong trường hợp tổng quát chúng ta nên nhóm các thuộc tính sao cho giữa các nhóm càng độc lập càng tốt.

Two-step LDA, multi-step LDA được so sánh với LDA và regularized LDA. Nếu dung lượng mẫu của dữ liệu huấn luyện $n < p + 2$, hàm LDA mẫu $\hat{\delta}_F$ không xác định do $\hat{\mathbf{\Sigma}}^{-1}$ không tồn tại. Chúng tôi thay $\hat{\mathbf{\Sigma}}^{-1}$ bởi giả nghịch đảo Moore-Penrose của $\hat{\mathbf{\Sigma}}$. Tham số chính quy của regularized LDA được tính bởi công thức đưa ra bởi Schäfer và Strimmer (oprlda) hoặc xác nhận chéo (cvrlda), tham khảo Huy và đồng nghiệp, 2012. Ở đây chúng tôi sử dụng 42 tập dữ liệu với dung lượng mẫu từ 450 đến 477. Đối với mỗi tập dữ liệu, n quan sát đầu tiên, với $n = 100, 125, 150, 175, 200, 225, 250$ được dùng như dữ liệu huấn luyện, những quan sát còn lại như kiểm thử. Do một trong hai lớp có dung lượng nhỏ nên độ sai lầm phân loại không phải là đánh giá có ý nghĩa. Thay vào đó chúng tôi sử dụng giá trị AUC, một độ đo tiêu chuẩn khác trong đánh giá hiệu quả các phương pháp phân loại. Nếu giá trị AUC càng gần một thì phương pháp phân loại càng hoàn hảo. Trung bình của các giá trị AUC qua 42 tập dữ liệu được trình bày trong bảng 1. Đối với những tập dữ liệu này, cvrlda tốt hơn đôi chút so với oprlda. Trung bình giá trị AUC của LDA (lda), two-step LDA (tslda), và multi-step LDA với kiểu $(16, 2, 2, 2, 2, 2, 2)$ (mtslda1), $(2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2)$ (mtslda2),



Hình 2: So sánh hiệu suất của multi-step LDA kiểu $(16, 2, 2, 2, 2, 2, 2)$ và regularized LDA qua 42 tập dữ liệu. Mức ý nghĩa thống kê p được tính bởi một kiểm thử xếp hạng Wilcolxon.

$(4, 8, 2, 2, 2, 2, 2)$ (mtslda3), $(8, 4, 2, 2, 2, 2, 2)$ (mtslda4), $(32, 2, 2, 2, 2, 2)$ (mtslda5) cũng được đưa ra trong bảng này. Trừ LDA, two-step LDA và multi-step LDA với kiểu $(32, 2, 2, 2, 2, 2)$ hiệu suất phân loại của chúng đều tốt hơn regularized LDA, phương pháp tiên tiến nhất cho loại dữ liệu này cho tới thời điểm hiện nay.

Chúng ta thấy rằng hiệu suất của two-step LDA và multi-step LDA với kiểu $(32, 2, 2, 2, 2, 2)$ tồi đối với n nhỏ, như $n = 100$ nhưng tốt hơn đối với n lớn, như $n = 250$ so với regularized LDA. Điều đó có thể được giải thích bởi ảnh hưởng của số chiều lớn, khi trong two-step LDA, chúng ta thuần túy áp dụng LDA đối với mỗi nhóm thuộc tính với dung lượng $p_1 = 32$ ở bước đầu tiên. Khi n tăng tốc độ hội tụ của độ sai lầm two-step LDA mẫu tới giá trị lý thuyết nhanh hơn LDA, tham khảo Huy, 2013. Điều này lý giải tại sao trung bình giá trị AUC của two-step LDA tại $n = 250$ cao hơn của LDA và regularized LDA. Điều đó có thể là một lợi thế đáng kể trong thực tiễn của multi-step LDA, do dữ liệu huấn luyện từ các nghiên cứu về giao diện máy tính-bộ não thường nhỏ, đặc biệt từ các giao diện máy tính-bộ não trực tuyến. Hiện tượng trên cũng được thấy trong hình 2. Hình vẽ này trình bày biểu đồ hộp của những giá trị AUC của regularized LDA sử dụng xác nhận chéo (cvrlda) và multi-step LDA kiểu $(16, 2, 2, 2, 2, 2, 2)$ (mtslda1) qua 42 tập dữ liệu như trên. Mỗi biểu đồ con tương ứng với dung lượng mẫu n từ 100 đến 250. Như trung bình giá trị AUC trong bảng 1, trung vị của multi-step LDA kiểu $(16, 2, 2, 2, 2, 2, 2)$ lớn hơn regularized LDA. Mức ý nghĩa thống kê p được tính bởi một kiểm thử xếp hạng Wilcoxon. Đầu tiên giá trị p giảm cho đến $n = 175$ rồi tăng. Điều này có thể giải thích như sau. multi-step LDA có tốc độ hội tụ nhanh hơn regularized LDA. Khi n đủ lớn, độ sai lầm của multi-step

LDA sẽ ổn định trong khi độ sai lầm của regularized LDA vẫn trong quá trình hội tụ.

Tài liệu

Bickel, P. J. và Levina, E. (2004): Some theory of Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli, 10(6):989-1010.*

Blankertz, B. và Lemm, S. và Treder, M. và Haufe, S và Müller K. R. (2011): Single-trial analysis and classification of ERP components – A tutorial. *NeuroImage, 56, 814—825.*

Bühlmann, P. và van de Geer, S. (2011): *Statisstics for High-Dimensional Data: Methods, Theory and Applications.* Springer Series in Statistics. Springer Verlag.

Fan, J và Fan, Y. (2008): High-Dimensional Classification Using Features Annealed Independence Rules. *The Annals of Statistics, 36(6):2605-2637.*

Fan, J. và Lv, J. (2010): A selective overview of variable selection in high dimensional feature space (invited review article). *Statistica Sinica, 20(1):101-148.*

Frenzel, S. và Neubert, E. và Bandt, C. (2011): Two communication lines in a 3×3 matrix speller. *Journal of Neural Engineering, 8, 036021.*

Friedman, J. H. (1989): Regularized Discriminant Analysis. *Journal of the American Statistical Association, 84(405):165-175.*

Genton, M. G. (2007): Separable approximation of space-time covariance matrices. *Environmetrics, 18, 681–695.*

Guo, J. (2010): Simultaneous Variable Selection and Class Fusion for High-Dimensional Linear Discriminant Analysis. *Biostatistics, 11(4):599-608.*

Hastie, T. và Tibshirani, R. và Friedman, J. (2009): *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Series in Statistics. Springer Verlag.

Huy, N. H. (2013): *Multi-Step Linear Discriminant Analysis and Its Applications*, Ph.D. Thesis, Department of Mathematics and Computer Science, University of Greifswald.

Huy, N. H. và Frenzel, S. và Bandt, C. (2012): Two-Step Linear Discriminant Analysis for Classification of EEG Data. *Data Analysis, Machine Learning and Knowledge Discovery In Studies in Classification, Data Analysis, and Knowledge Organization, 2012. Đã được chấp nhận.*

Krusienski, D. J. và Sellers, E. W. và McFarland, D. J. và Vaughan, T. M. và Wolpaw, J. R. (2008): Toward enhanced P300 speller performance. *Journal of Neuroscience Methods, 167(1):15-21.*

Ledoit, O. và Wolf, M. (2002): Some hypothesis test for the covariance matrix when the dimension is large compared to the sample size. *The Annals of Statistics, 30(4):1081-1102.*

Lotte, F. và Congedo, M. và Lécuyer, A. và Lamarche, F. và Arnaldi, B. (2007): A Review of Classification Algorithms for EEG-based Brain-Computer Interfaces. *Journal of Neural Engineering, 4(2):R1-R13*

Mclachlan, G. J. (1992): *Discriminant Analysis and Statistical Pattern Recognition.* Wiley series in probability and statistics. John Wiley & Sons, Hoboken, New Jersey.

Mitchell, M. W. và Genton, M. G. và Gumpertz, M. L. (2006): A likelihood ratio test for separability of covariances. *Journal of Multivariate Analysis, 97, 1025–1043.*

Nicolas-Alonso, L. F. và Gomez-Gil, J. (2012): Brain Computer Interfaces, a Review. *Sensors, 12(2):1211-1279.*

Zhang, Q. và Wang, H. (2011): On BIC's selection consistency for discriminant analysis. *Statistica Sinica, 21(2):731-740.*