

به نام خدا

تحلیل هندسی اطلاعات حالات شناختی در سیگنال‌های EEG

چکیده

سیگنال‌های الکتروانسفالوگرافی (EEG) نمایی ارزشمند از فعالیت مغز فراهم می‌کنند، اما استخراج شاخص‌های پایدار برای تمایز حالات شناختی از این سیگنال‌ها همچنان دشوار است. در این پژوهش، چارچوبی مبتنی بر هندسه اطلاعات برای کمی‌سازی تفاوت‌های احتمالی میان حالات (در اینجا: دو سطح valence) پیشنهاد می‌شود. با استفاده از مجموعه داده DEAP (۳۲ سوژه و ۱۲۸۰ آزمایش)، چند معیار فاصله/واگرایی میان توزیع‌های احتمالی پیاده‌سازی شد، تعمیم‌پذیری با اعتبارسنجی حذف یک سوژه (Leave-Subject-Out) ارزیابی شد، و معنی‌داری جدایی مشاهده‌شده با آزمون جایگزینی سنجیده شد. نتایج اعتبارسنجی میانگین دقت ۷۷.۷٪ با انحراف معیار ۱۱.۶٪ را نشان می‌دهد که از سطح تصادفی بالاتر است؛ با این حال، آزمون جایگزینی ($p = 0.003$) شواهد کافی برای معنی‌دار بودن جدایی حالات فراهم نمی‌کند. این یافته‌ها نشان می‌دهند که روش پیشنهادی بخشی از تنوع مرتبط را ثبت می‌کند، اما برای ادعای جدایی معنادار احتمالاً به ویژگی‌های غنی‌تر، مدل‌سازی توزیعی مناسب‌تر، یا داده بیشتر نیاز است.

کلیدواژه‌ها: الکتروانسفالوگرافی (EEG)، هندسه اطلاعات، تحلیل احتمالی، واگرایی Jensen-Shannon، تعمیم‌پذیری بین‌سوژه‌ای، آزمون جایگزینی

۱. مقدمه

۱.۱ انگیزه و اهمیت

درک رابطه بین فعالیت الکتریکی مغز و حالات شناختی انسان یکی از مسائل بنیادی در علوم مغز محسوب می‌شود. سیگنال‌های EEG به دلیل دستیابی‌پذیری بالا، هزینه کم، و وضوح زمانی خوب، وسیله‌ای رایج برای مطالعه چنین حالت‌هایی هستند. با این حال، سیگنال‌های EEG دارای نویز بالا، تنوع فردی قابل توجه، و مقیاس‌پذیری ضعیف بین سوژه‌ها هستند که تفکیک معنادار بین حالات را برای استفاده عملی دشوار می‌سازد. روش‌های سنتی برای تحلیل EEG و تفکیک حالات شناختی معمولاً بر روی ویژگی‌های متفرقه (مثل قدرت طیفی در باندهای مختلف) و طبقه‌بندی کنندگان ماشینی متکی‌اند. اگرچه این روش‌ها درجه‌ای از موفقیت داشته‌اند، اما اغلب تفسیرپذیری محدودی دارند و کمتر از آنچه درباره ساختار احتمالی زیرین داده‌ها فاش می‌کنند، استفاده می‌کنند.

۱.۲ نوآوری و سهم تحقیقاتی

هندسه اطلاعات، رشته‌ای برای مطالعه توزیع‌های احتمالی از منظر هندسی است. این روش امکان می‌دهد تا فاصله‌های بین توزیع‌ها را به روشی معنی‌دار و نظری‌مند اندازه‌گیری کنیم. سهم‌های کلیدی این مقاله عبارت‌اند از:

۱. اجرای جامع روش‌های هندسی اطلاعات: ما سه معیار فاصله مختلف را برای کمی‌سازی جدایی حالات EEG پیاده‌سازی کردیم:

- متقارن و محدود - Jensen-Shannon Divergence (JSD)
- مبتنی بر اطلاعات - Symmetric Kullback-Leibler (SKL)
- متریک سراسری - Hellinger Distance

2. ارزیابی کامل تعمیم‌پذیری: استفاده از Leave-Subject-Out Cross-Validation برای ارزیابی چگونگی عملکرد

مدل‌ها روی سوژه‌های دیدنی نشده

3. آزمایش آماری دقیق: استفاده از آزمایش جایگشتی برای تشخیص اینکه آیا جدایی مشاهده‌شده احتمالاً واقعی است یا صرفاً تصادفی

4. شفافیت و بازتولیدپذیری: مستندات کامل و کد قابل بازتولیدپذیری برای تسهیل کار محققان آتی

۱.۳ ساختار مقاله

سازماندهی مقاله به شرح زیر است: بخش ۲ کار مرتبط را بررسی می‌کند. بخش ۳ روش‌شناسی را به تفصیل شرح می‌دهد. بخش ۴ نتایج تجربی را ارائه می‌دهد. بخش ۵ یافته‌ها را مورد بحث قرار می‌دهد و محدودیت‌ها را شناسایی می‌کند. سرانجام، بخش ۶ نتیجه‌گیری و راهنمایی‌های آتی را ارائه می‌دهد.

۲. کار مرتبط

۲.۱ تحلیل سیگنال‌های EEG

تحلیل سیگنال‌های EEG برای تفکیک حالات شناختی پزشکی، علوم مغز و کاربردهای تعاملی انسان-کامپیوتر در تاریخ را داراست. روش‌های کلاسیک برای استخراج ویژگی شامل تحلیل طیفی (بر پایه تبدیل فوریه)، تحلیل فرکانس بیشتر و موجک است. توزیع قدرت طیفی در باندهای شناخته‌شده (delta، theta، alpha، beta، gamma) معمولاً منعکس کننده حالات فعال‌سازی مختلف مغز تلقی می‌شود.

تحقیقات بسیاری نشان داده‌اند که ویژگی‌های EEG می‌توانند حالات احساسی یا شناختی مختلف را تفریق کنند، اگرچه دقت معمولاً به مجموعه داده و روش‌های استخراج ویژگی انتخاب‌شده بستگی دارد.

۲.۲ روش‌های احتمالی و آماری در EEG

رویکردهای احتمالی برای تحلیل EEG درجه‌ای از عدم قطعیت و تنوع طبیعی در سیگنال‌ها را دریافت می‌کنند. مدل‌های مخلوط گاوسی (Gaussian Mixture Models) و مدل‌های مارکوف پنهان (Hidden Markov Models) برای مدل‌سازی پویای حالات EEG استفاده شده‌اند. با این حال، اکثر رویکردهای احتمالی عمومی روی مسائل تصنیفی (مثل تعریف مرزهای کلاس) متمرکز بوده‌اند تا جای اینکه بر روی ساختار هندسی توزیع‌ها.

۲.۳ هندسه اطلاعات و معیارهای فاصله

هندسه اطلاعات زمینه‌ای مطالعه خواص هندسی فضاها احتمالی است. این رشته ابزارهایی برای اندازه‌گیری "فاصله" بین توزیع‌های احتمالی بر پایه اطلاعات فیشر یا واگرایی نسبی (KL divergence) فراهم می‌کند.

Jensen-Shannon Divergence (JSD)

است: KL divergence متقارن از JSD

$$D_{JS}(P||Q) = \frac{1}{2}D_{KL}(P||M) + \frac{1}{2}D_{KL}(Q||M)$$

جایی که $M = \frac{P+Q}{2}$ توزیع میانگین است. JSD متقارن و محدود ($[0, \ln 2]$) برای توزیع‌های گسسته است.

Kullback-Leibler (KL) Divergence

معیاری برای اینکه یک توزیع Q چقدر از توزیع واقعی P فاصله دارد KL divergence:

$$D_{KL}(P||Q) = \int P(x) \log \frac{P(x)}{Q(x)} dx$$

Hellinger Distance

فاصله Hellinger معیاری متریک است:

$$H(P, Q) = \sqrt{1 - \int \sqrt{P(x)Q(x)} dx}$$

این معیار کران‌دار است و خواص هندسی مناسبی دارد.

۲.۴ تقاطع ارزیابی و اعتبارسنجی

اعتبارسنجی متقاطع روش استاندارد برای ارزیابی توانایی تعمیم مدل‌هایی است. Leave-Subject-Out (LOO) variant. EEG داده‌های مرتبط به انسان مناسب است، زیرا انتظار داریم کاربرد واقعی روی سوژه‌های جدید صورت گیرد. یک نکته کلیدی در LOSOCV توجه به تفاوت میان «نشت داده» (data leakage) و «تعمیم واقعی» (proper generalization) است - وقتی تقلیل ابعاد برای تمام داده‌ها قبل از تقسیم به تقاطع انجام می‌شود، نتایج اعتبارسنجی متحیز خواهد بود.

۲.۵ آزمایش‌های جایگشتی برای اعتبارسنجی

آزمایش‌های جایگشتی (یا آزمایش‌های permutation test) روشی غیرپارامتری برای تعیین معنی‌داری آماری یک نتیجه هستند. برای آزمایش اینکه آیا توزیع‌های دو حالت واقعاً متفاوت هستند:

۱. مشاهده‌شده: فاصله محاسبه‌شده بین توزیع‌های حالات واقعی
 ۲. جایگشت (Permutation)
 ۳. مقدار p-value: نسبت permutations که فاصله بیشتر از مشاهده‌شده دارند
- اگر p-value کوچک است (مثلاً $0.05 >$)، جدایی احتمالاً معنی‌دار است. اگر p-value بزرگ است (مثلاً $0.10 <$)، نمی‌توانیم جدایی معنی‌دار را ادعا کنیم.

۲.۶ مجموعه داده DEAP

مجموعه داده DEAP توسط Koelstra و دیگران (۲۰۱۲) در University of Twente توسعه داده‌شد. این مجموعه شامل EEG های ۳۲ فرد هنگام تماشای ۴۰ ویدئوی متنوع و ارزیابی سطح احساسات خود بر حسب چهار بعد (valence، arousal، dominance، liking) است. داده‌های EEG پیش‌پردازش‌شده‌اند و در دامنه فرکانسی ۴-۴۵ هرتز صادر می‌شوند.

۳. روش‌شناسی

۳.۱ مفهوم و فرض‌های بنیادی

فرضیه بنیادی ما این است که اگر حالات شناختی مختلف تفاوت‌های الگویی در فعالیت EEG ایجاد می‌کنند، آنگاه توزیع‌های احتمالی ویژگی‌های EEG برای حالات مختلف باید متفاوت باشند. هندسه اطلاعات ابزاری فراهم می‌کند برای کمی‌سازی این تفاوت‌های احتمالی.

۳.۲ بیان ریاضی

بگذارید $X = \{x_i \in \mathbb{R}^d : i = 1, \dots, n\}$ مجموعه‌ای از ویژگی‌های استخراج‌شده از EEG باشد، و $y_i \in \{1, 2\}$ برچسب حالت باشد (۱ برای valence پایین، ۲ برای valence بالا). ما داده‌ها را به دو مجموعه تقسیم می‌کنیم:

$$X_1 = \{x_i : y_i = 1\}, \quad X_2 = \{x_i : y_i = 2\}$$

برای هر حالت، توزیع احتمالی را با یک توزیع گاوسی چندمتغیره تقریب می‌زنیم:

$$P_j \sim \mathcal{N}(\mu_j, \Sigma_j)$$

جایی که μ_j بردار میانگین و Σ_j ماتریس کواریانس برای حالت j است. سپس، یکی از معیارهای فاصله را محاسبه می‌کنیم:

JSD:

$$D_{JS}(P_1||P_2) = \frac{1}{2}D_{KL}(P_1||M) + \frac{1}{2}D_{KL}(P_2||M)$$

SKL:

$$D_{SKL}(P_1||P_2) = D_{KL}(P_1||P_2) + D_{KL}(P_2||P_1)$$

Hellinger:

$$H(P_1, P_2) = \sqrt{1 - \int \sqrt{P_1(x)P_2(x)} dx}$$

اگر این فاصله‌ها به‌طور آماری معنادار هستند (از طریق آزمایش جایگشتی تأیید می‌شوند)، نتیجه می‌گیریم که حالات واقعاً قابل تفریق‌اند.

۳.۳ خطنوله پردازش

مرحله ۱: استخراج ویژگی‌ها
برای هر تجربه EEG:

- سیگنال را برای باندهای alpha (8-13 Hz)، theta (4-8 Hz)، و beta (13-30 Hz) فیلتر کنید
- برای هر کانال و هر باند، قدرت طیفی را محاسبه کنید (Welch method)
- این نتیجه‌اش ۳۲ کانال $3 \times$ باند = ۹۶ ویژگی برای هر نمونه است

مرحله ۲: پیش‌پردازش

- تصحیح خط پایه: هر ویژگی را با میانگین خط پایه تفریق کنید
- نرمال‌سازی درون‌سوژه‌ای: برای هر سوژه، ویژگی‌ها را z-score کنید

مرحله ۳: مرحله اکتشافی

- را برای تقلیل ابعاد اعمال کنید (۹۶ \rightarrow ۱۰ مؤلفه) PCA
- داده‌ها را بر اساس valence به دو حالت تقسیم کنید

- توزیع‌های گاوسی را برازش کنید
- فواصل را محاسبه کنید و MDS را برای بصری‌سازی اعمال کنید
- مرحله ۴: اعتبارسنجی بین‌سوژه‌ای
- For each fold (سوژه‌ای حذف شود):
 - Train: تمام سوژه‌های دیگر
 - PCA آموزش برازش کنید fold را برای
 - Test: سوژه حذف‌شده
 - آموزش تبدیل کنید PCA داده‌ها را با Test
- مدل را بر روی Train برازش کنید و روی Test ارزیابی کنید
- مرحله ۵: آزمایش جایگشتی
- اصلی: فاصله محاسبه‌شده برای حالات واقعی
- Null: 1000 times:
 - برچسب‌های حالات را تصادفی کنید
 - PCA را برای جایگشت برازش کنید
 - فاصله‌ی جدید را محاسبه کنید
- p-value: proportion of null distances بزرگتر یا مساوی اصلی

۳.۴ پارامترهای تنظیم

پارامتر	مقدار
PCA components	10
Artifact voltage threshold	100 μ V
JSD Monte Carlo samples	5000
Permutation test iterations	1000
Global random seed	42

۳.۵ توضیح انتخاب روش

- چرا گاوسی؟ گاوسی‌ها انتخاب محاسباتی دقیق و در بسیاری موارد تقریب معقول هستند. ما سپس انحراف‌ها را آزمایش کردیم.
- چرا PCA؟ تقلیل ابعاد به بهبود پایداری عددی و کاهش سر و صدا کمک می‌کند.
- چرا Leave-Subject-Out؟ این تنها راه برای ارزیابی واقعی تعمیم‌پذیری است (سوژه‌های دیدنی نشده).
- چرا Permutation Test؟ این روش غیرپارامتری و نسبتاً مقاوم، راهی معتبر برای آزمون معنی‌داری است.

۴. نتایج تجربی

۴.۱ شاخص‌های توصیفی

- سوژه‌ها: 32

- **Total trials:** 1,280 (40 per subject)
- **Original feature dimension:** 96
- **Reduced dimension (PCA):** 10
- **Explained variance (PCA):** 78.0%

۴.۲ توزیع حالات

State	Samples	%
Low Valence	995	77.7
High Valence	285	22.3

نسبت imbalance: 3.5:1. این عدم توازن به انتخاب آستانه (threshold = 5.0) برمی گردد.

۴.۳ فواصل اطلاعات هندسی (Exploratory Phase)

Metric	Distance	MDS Stress
JSD	0.0399	0.0000
SKL	0.1697	0.0000
Hellinger	0.2023	0.0000

یافته: معیارهای مختلف فاصله‌های مختلفی نشان می‌دهند. JSD کمترین است، Hellinger بیشترین. این تنوع ناشی از خواص ریاضی متفاوت معیارها است.

۴.۴ نتایج اعتبارسنجی (Cross-Validation)

خلاصه دقت:

```
| Metric | Value |
|-----|-----|
| Mean | 0.777 |
| Std Dev | 0.116 |
| Median | 0.788 |
| Min | 0.475 |
| Max | 0.950 |
| 95% CI | [0.567, 0.987] |
```

توزیع برحسب سوژه:

- بالای 90%: s04, s12, s13, s21 (4 سوژه)
- 80-90%: s01, s09, s14, s16, s19, s24, s27, s29 (8 سوژه)
- 70-80%: s02, s08, s10, s14, s17, s18, s22, s25, s26, s30, s32 (11 سوژه)
- 60-70%: s05, s11 (2 سوژه)

- (سوژه 4) s03, s06, s23, s31: <60%

تفسیر: میانگین دقت ۷۷.۷٪ بالاتر از سطح تصادفی (۵۰٪) است، اما واریانس بالا تفاوت‌های بین‌فردی معنادار را نشان می‌دهد.

۴.۵ نتایج آزمایش جایگشتی (Permutation Test)

Metric	Value
Observed distance	0.1697
Null mean	0.2064
Null std	0.0447
p-value	0.8030
Significant ($\alpha=0.05$)?	No

توزیع صفر: میانگین (0.2064) null distribution بیشتر از فاصله مشاهده‌شده (0.1697) است. از 1000 جایگشت، در 803 مورد فاصله بیشتر یا مساوی فاصله مشاهده‌شده بود ($p = 0.803$).
تفسیر: این نشان می‌دهد که جدایی مشاهده‌شده احتمالاً تصادفی است و معنی‌دار نیست.

۴.۶ پارامترهای توزیع برآزش‌شده

میانگین‌های PCA (۱۰ مؤلفه):

```
| PC | Low Valence | High Valence |
|----|-----|-----|
| 0 | -0.1046 | 0.3651 |
| 1 | 0.0238 | -0.0832 |
| 2 | 0.0440 | -0.1535 |
| ... | ... | ... |
```

نکته: بزرگ‌ترین تفاوت در PC0 است (-0.10 vs 0.37).
واریانس‌ها (Diagonal):

```
| PC | Low | High |
|----|-----|-----|
| 0 | 47.48 | 51.74 |
| 1 | 13.49 | 13.61 |
| 2 | 5.56 | 5.02 |
```

۴.۷ آزمایش نرمالیتی

Shapiro-Wilk Test ($p < 0.01$):

- Low Valence: 9/10 مؤلفه غیرنرمال
 - High Valence: 5/10 مؤلفه غیرنرمال
- نتیجه: فرض گاوسی تقریب است. نتایج باید با احتیاط تفسیر شوند.

۵. بحث

۵.۱ تفسیر نتایج

۵.۱.۱ موفقیت تقاطع و شکست جایگشتی

دقت تقاطع ۷۷.۷٪ نشان می‌دهد که فاصله معیار (distance metric) توانایی دریافت بخشی از تفاوت‌های حالات را دارد. اما آزمایش جایگشتی با $p = 0.803$ نشان می‌دهد که این تفاوت احتمالاً معنی‌دار نیست. چگونه می‌توان این تناقض را توضیح داد؟

۱. دقت بالا می‌تواند صرفاً تصادفی باشد: اگر حالات بسیار متفاوت هستند (آستانه شدید) اما بر اساس اثر نویز، مدل می‌تواند نویز را بیاموزد.
۲. تأثیر عدم توازن کلاس: نسبت 3.5:1 می‌تواند الگوریتم‌های طبقه‌بندی را بر سمت کلاس بزرگتر کج کند، دقت apparent را بیشتر می‌نماید.
۳. درون فرض‌ها: توزیع‌های برازش‌شده ممکن است بیش از حد برای داده‌های مشاهده‌شده برازش باشند **Overfitting**.

۵.۱.۲ فواصل مختلف

است. این تفاوت انعکاسی از ریاضیات متفاوت است (Hellinger (0.2023 و SKL (0.1697 کمتر از JSD (0.0399

- میانگین‌سازی متقارن، مقاوم به توزیع‌های بسیار متفاوت: JSD
- جمع دو طرفه، بزرگتر هنگام اختلاف بیشتر: SKL
- انتگرال جذر، هندسی تفسیرپذیر: Hellinger

۵.۲ نقاط قوت

۱. ✓ روش سخت: استفاده از LOSOCV، آزمایش جایگشتی، و اعتبارسنجی آماری
۲. ✓ بازتولیدپذیری: کد کامل و قابل اجرا فراهم‌شده است
۳. ✓ شفافیت: فرض‌ها، محدودیت‌ها، و تنظیم‌ها به وضوح بیان‌شده‌اند
۴. ✓ تنوع معیار: چندین معیار فاصله برای مقایسه

۵.۳ نقاط ضعف

۱. X فرض گاوسی ضعیف: 50٪ تا 90٪ مؤلفه‌ها از نرمالیتی انحراف دارند
۲. X عدم معنی‌داری آماری: $p = 0.803$ نتیجه‌ای صریح است
۳. X عدم توازن شدید: نسبت 3.5:1 نگران‌کننده است
۴. X تنوع بین‌فردی بالا: برخی سوژه‌ها دقتی نزدیک به تصادفی دارند
۵. X هیچ انتخاب فیچر: تمام ۹۶ ویژگی استفاده می‌شود

۵.۴ تفسیرات جایگشتی

نتیجه جایگشتی شاید باید غافلگیری نیست. چرا؟

- شاید یک سازه‌ی پیوسته بیشتر است تا واقعی دسته بندی **valence**
- بسیار نویزی است و هر ۷۷.۷٪ دقت می‌تواند اساساً نویز مرتبط با نویز باشد EEG

- فیچرهای استخراج شده (قدرت طیفی) ممکن است برای تفکیک دقیق valence کافی نباشند

۵.۵ انحراف‌های از فرض

انحراف‌های در نرمالیتی پیشنهاد می‌کنند:

1. روش‌های غیرپارامتری ممکن است بهتر باشند
2. توزیع‌های مخلوط ممکن است مناسب‌تر باشند
3. تبدیل‌های غیرخطی ممکن است مفید باشند

۶. محدودیت‌ها

1. اندازه نمونه: ۱۲۸۰ نمونه برای تحلیل احتمالی با ۹۶ بعد کوچک است
2. **DEAP**: تک: تنها از یک مجموعه داده استفاده شد؛ نتایج شاید تعمیم‌ناپذیر باشند
3. **یک بعد**: تنها valence مورد بررسی قرار گرفت؛ ارتباط‌های بین‌بعدی نادیده گرفته شدند
4. **خطی PCA**: تقلیل ابعاد خطی ممکن است ساختارهای غیرخطی مهم را از دست دهد
5. **عدم بررسی ویژگی**: هیچ انتخاب یا تحلیل حساسیت ویژگی انجام نشد

۷. نتیجه‌گیری و کارهای آتی

۷.۱ خلاصه یافته‌ها

این مطالعه روش‌های هندسی اطلاعات را برای کمی‌سازی تفاوت‌های احتمالی بین حالات شناختی (valence) در سیگنال‌های EEG اعمال کرد. نتایج ما:

- **تقاطع**: میانگین دقت ۷۷.۷٪ (معقول، اما غیریقین)
- **جایگشتی**: $p = 0.803$ (معنی‌دار نیست)
- **نتیجه‌گیری**: روش می‌تواند بخشی از تنوع را دریافت کند، اما برای ادعای جدایی معنی‌دار احتمالاً نیاز به بهبود است

۷.۲ کارهای آتی

1. ویژگی‌های بهتر: Non-linear features (entropy, complexity) یا connectivity measures
2. روش‌های غیرپارامتری: Kernel density estimation یا mixture models
3. داده‌های بیشتر: مجموعه‌های داده دیگر (SEED، EmoDB) برای اعتبارسنجی
4. بهبود متن توازن: Re-sampling یا Cost-sensitive learning
5. تحلیل عمیق: Feature importance، source-level analysis

۷.۳ اثر بالقوه

اگرچه نتایج حاضر نشان‌دهنده چالش‌هایی هستند، اما این کار پایه‌هایی را برای:

- درک نقش فاصله‌های احتمالی در تحلیل EEG فراهم می‌کند
- روش‌های تصحیح برای مطالعات پسینی را پیشنهاد می‌کند
- اهمیت اعتبارسنجی آماری سخت را برجسته می‌کند