

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 مجموعه‌داده DEAP (۳۲۰ سوزه و ۱۲۸۰ آزمایش)، چند معیار فاصله/واگرایی میان توزیع‌های احتمالی پیاده‌سازی شد، تعمیم‌پذیری با اعتبارسنجی حذف یک سوزه (Leave-Subject-Out) ارزیابی شد، و معنی‌داری جدایی مشاهده شده با آزمون جایگشتی 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 جایگشتی ($p = 0.03$) 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 بین‌سوزه‌ای، آزمون جایگشتی

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 قابل‌توجه، و مقیاس‌پذیری ضعیف بین سوزه‌ها 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 درجه‌های از موفقیت داشته‌اند، اما غالب تفسیر‌پذیری محدودی دارند و کمتر از آنچه درباره ساختار احتمالی زیرین داده‌ها فاش می‌کنند، استفاده می‌کنند.

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 بهروشی 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:

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

۲. ارزیابی کامل تعمیم‌پذیری: استفاده از Leave-Subject-Out Cross-Validation برای ارزیابی چگونگی عملکرد مدل‌ها روی سوژه‌های دیدنی نشده

۳. آزمایش آماری دقیق: استفاده از آزمایش جایگشتی برای تشخیص اینکه آیا جدایی مشاهده شده احتمالاً واقعی است یا صرفاً تصادفی

۴. 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

۱.۲ تحلیل سیگنال‌های EEG

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

۲.۱ روش‌های احتمالی و آماری در EEG

رویکردهای احتمالی برای تحلیل EEG درجه‌ای از عدم‌قطعیت و تنوع طبیعی در سیگنال‌ها را دریافت می‌کنند. مدل‌های مخلوط گاوی (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.

۲.۲ هندسه اطلاعات و معیارهای فاصله

هندسه اطلاعات زمینه‌ای مطالعه خواص هندسی فضاهای 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 JSD تصحیح متقارن از KL divergence:

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

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

Kullback-Leibler (KL) Divergence

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

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

Hellinger Distance

فاصله Hellinger معیاری متريک است:

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

این معیار کراندار است و خواص هندسی مناسبی دارد.

۴. تقاطع ارزیابی و اعتبارسنجی

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

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:

1. مشاهده شده: فاصله محاسبه شده بین توزیع‌های حالات واقعی
2. جایگشت (Permutation)
3. مقدار p-value: نسبت permutations که فاصله بیشتر از مشاهده شده دارند
اگر p-value کوچک است ($p < 0.05$)، جدایی احتمالاً معنی‌دار است. اگر p-value بزرگ است ($p > 0.10$ ، نمی‌توانیم جدایی معنی‌دار را ادعا کنیم.

۶. DEAP مجموعه داده

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

۳. روش‌شناسی

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.

۲. ۳. بیان ریاضی

بگذارید $\{y_i \in \{1, 2\} : i = 1, \dots, n\}$ مجموعه‌ای از ویژگی‌های استخراج شده از EEG باشد، و $X = \{x_i \in \mathbb{R}^d : i = 1, \dots, n\}$ 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 گاووسی چندمتغیره تقریب می‌زنیم:

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

جایی که μ_j بردار میانگین و Σ_j ماتریس کوواریانس 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 به طور آماری معنادار هستند (از طریق آزمایش جایگشتی تأیید می‌شوند)، نتیجه می‌گیریم که حالات واقعاً قابل تفرقی‌اند.

۳. خطوله پردازش

مرحله ۱: استخراج ویژگی‌ها
برای هر تجربه EEG

- سیگنال را برای باندهای (beta (13-30 Hz), alpha (8-13 Hz)، theta (4-8 Hz) فیلتر کنید
- برای هر کanal و هر باند، قدرت طیفی را محاسبه کنید (Welch method)
- این نتیجه‌اش $32 \text{ کanal} \times 3 \text{ باند} = 96$ ویژگی برای هر نمونه است

مرحله ۲: پیش‌پردازش

- تصحیح خط پایه: هر ویژگی را با میانگین خط پایه تفرقی کنید

• نرمال‌سازی درون‌سوژه‌ای: برای هر سوژه، ویژگی‌ها را z-score کنید

مرحله ۳: مرحله اکتشافی

- PCA را برای تقلیل ابعاد اعمال کنید ($96 \rightarrow 10$ مؤلفه)
- داده‌ها را بر اساس valence به دو حالت تقسیم کنید

- توزیع‌های گاوی را برآورد 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
 - آموزش برآوردهای fold را برای PCA
 - Test: سوژه حذف شده
 - آموزش تبدیل کنید PCA داده‌ها را با
 - مدل را بر روی Train برآورده کنید و روی Test ارزیابی کنید
- Null: 1000 times:
 - برچسب‌های حالات را تصادفی کنید
 - را برای جایگشت برآورده کنید
 - فاصله‌ی جدید را محاسبه کنید
- p-value: proportion of null distances بزرگتر یا مساوی اصلی

۴. ۳. 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

۵. ۳. توضیح انتخاب روش

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

۴. نتایج تجربی

۱. ۴. شاخص‌های توصیفی

- **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 3.5:1 imbalance. این عدم توازن به انتخاب آستانه (threshold = 5.0) برمی‌گردد.

۴.۳ فوایل 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 متقابل 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 بر حسب سوژه:

• بالای 4 سوژه (s04, s12, s13, s21) : 90٪

- 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)

- <60%: s03, s06, s23, s31 (سوژه 4)

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

۴. 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 مشاهده شده (0.1697) است. از 1000 جایگشت، در

803 مورد فاصله بیشتر یا مساوی فاصله مشاهده شده بود ($p = 0.803$).

تفسیر: این نشان می‌دهد که جدایی مشاهده شده احتمالاً تصادفی است و معنی‌دار نیست.

۵. پارامترهای توزیع برآورده شده

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 vs 0.37 (0.10) است (-)

:واریانس‌ها (Diagonal)

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

۶. آزمایش نرمالیتی

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

- Low Valence: 9/10 مؤلفه غیرنرمال

- High Valence: 5/10 مؤلفه غیرنرمال

نتیجه: فرض گاوی تقریب است. نتایج باید با احتیاط تفسیر شوند.

۵. بحث

۱.۵ تفسیر نتایج

۱.۱.۵ موفقیت تقاطع و شکست جایگشتی

دقت تقاطع ۷۷.۷٪ نشان می‌دهد که فاصله معيار (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. دقتش 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 ۳.۵:۱ می‌تواند الگوریتم‌های طبقه‌بندی را بر سمت کلاس بزرگتر کج کند، دقتش apparent را بیشتر می‌نماید.
۳. **Overfitting**: توزیع‌های برآراش شده ممکن است بیش از حد برای داده‌های مشاهده شده برآراش باشند.

۲.۱.۵ فواصل مختلف

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

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

۲.۵ نقاط قوت

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

۳.۵ نقاط ضعف

۱. X فرض گاوی ضعیف: ۹۰٪ تا ۵۰٪ مؤلفه‌ها از نرمالیتی انحراف دارند
۲. X عدم معنی‌داری آماری: $p = 0.803$ نتیجه‌ای صریح است
۳. X عدم توازن شدید: نسبت ۳.۵:۱ نگران‌کننده است
۴. X تنوع بین‌فردی بالا: برخی سوژه‌ها دقیق نزدیک به تصادفی دارند
۵. X هیچ انتخاب فیچر: تمام ۹۶ ویژگی استفاده می‌شود

۴.۵ تفسیرات جایگشتی

نتیجه جایگشتی شاید باید غافلگیری نیست. چرا؟

- valence شاید یک سازه‌ی پیوسته بیشتر است تا واقعی دسته‌بندی
- EEG بسیار نویزی است و هر ۷۷.۷٪ دقتش می‌تواند اساساً نویز مرتبط با نویز باشد

- فیچر های استخراج شده (قدرت طبی) ممکن است برای تفکیک دقیق valence کافی نباشد

۵. انحراف های از فرض

انحراف های در نرمالیتی پیشنهاد می کنند:

1. روش های غیرپارامتری ممکن است بهتر باشند
2. توزیع های مخلوط ممکن است مناسب تر باشند
3. تبدیل های غیرخطی ممکن است مفید باشند

۶. محدودیت ها

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

۷. نتیجه گیری و کارهای آتی

۱. خلاصه یافته ها

این مطالعه روش های هندسی اطلاعات را برای کمی سازی تفاوت های احتمالی بین حالات شناختی (valence) در سیگنال های EEG اعمال کرد. نتایج ما:

- تقاطع: میانگین دقت ۷۷.۷٪ (معقول، اما غیریقین)
- جایگشتی: $p = 0.803$ (معنی دار نیست)
- نتیجه گیری: روش می تواند بخشی از تنوع را دریافت کند، اما برای ادعای جدایی معنی دار احتمالاً نیاز به بهبود است

۲. کارهای آتی

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

۳. اثر بالقوه

اگرچه نتایج حاضر نشان دهنده چالش هایی هستند، اما این کار پایه هایی 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