دانشگاه تهران
پردیس دانشکده‌های فنی
دانشکده مهندسی برق و کامپیوتر

یادگیری ماشین

پروژه اصلی فاز شماره ۲

نام و نام خانوادگی
محمد ستار امامی میبدی

شماره دانشجویی
۸۱۰۱۰۴۰۷۲

۲۶ بهمن ۱۴۰۴

فهرست مطالب

۱ خوشه‌بندی زبان از روی ویژگی‌های صوتی ۱
۱.۱ تعریف مسئله و هدف ۱
۲.۱ داده، ویژگی‌ها و آمار توصیفی ۱
۳.۱ چارچوب ارزیابی و معیارها ۲
۴.۱ پیش‌پردازش و ساخت دو نسخه‌ی داده (Step1) ۳
۵.۱ خوشه‌بندی با K-Means (Step2) ۴
۶.۱ خوشه‌بندی چگالی‌محور با DBSCAN (Step3) ۶
۷.۱ خوشه‌بندی با OPTICS (Step4) ۸
۸.۱ جمع‌بندی مقایسه‌ای تا پایان Step4 . . . . . ۱۰

فهرست تصاویر

۱ مقایسه‌ی silhouette بر حسب k برای K-Means در RAW و OPT (حالت no_augmentation). ۵

۲ نمودار elbow (بر اساس inertia) برای K-Means در RAW و OPT (حالت augmented). ۶

۳ نمای PCA2: برچسب واقعی زبان در کنار برچسب خوشه‌ی K-Means (نسخه‌ی RAW). ۶

۴ مقایسه‌ی نمای PCA2 خوشه‌های بهترین DBSCAN برای RAW و OPT. ۸

۵ منحنی k-distance برای تخمین eps (نمونه، حالت augmented-OPT). ۹

۶ نمای PCA2: برچسب واقعی زبان در کنار برچسب خوشه‌ی DBSCAN (نسخه‌ی RAW). ۱۰

۷ مقایسه‌ی reachability plot برای RAW و OPT (حالت no_augmentation). ۱۰

۸ مقایسه‌ی reachability plot برای RAW و OPT (حالت augmented). ۱۱

۹ نمای PCA2: برچسب واقعی زبان در کنار برچسب خوشه‌ی OPTICS (نسخه‌ی OPT). ۱۱

فهرست جداول

۱ نتایج بهترین تنظیم K-Means بر اساس بیشینه‌سازی silhouette (Step2) ۵
۲ بهترین تنظیم‌های DBSCAN و کیفیت (Step3) ۷
۳ Purity روی نقاط غیرنویز و دامنه‌ی اندازه‌ی خوشه‌ها در DBSCAN (Step3) ۷
۴ بهترین تنظیم‌های OPTICS و کیفیت (Step4) ۹
۵ Purity روی نقاط غیرنویز در OPTICS (Step4) ۱۰
۶ مقایسه‌ی خلاصه‌ی بهترین خروجی‌ها در سه الگوریتم (تا 4Step) ۱۱

خوشه‌بندی زبان از روی ویژگی‌های صوتی

۱.۱ تعریف مسئله و هدف

هدف ، خوشه‌بندی نمونه‌های صوتی بر اساس زبان گفتار است؛ به این معنا که با استفاده از بردار ویژگی استخراج‌شده از هر فایل صوتی، بدون استفاده از برچسب‌ها (یادگیری بدون‌نظارت)، نمونه‌ها به چند خوشه تقسیم شوند. سپس برای تحلیل ، کیفیت خوشه‌بندی هم با معیارهای بدون‌نظارت و هم با معیارهای وابسته به برچسب واقعی (صرفاً برای ارزیابی) بررسی می‌شود.

۲.۱ داده، ویژگی‌ها و آمار توصیفی

۱.۲.۱ ساختار داده و برچسب‌ها

بر اساس خروجی‌های ارسال‌شده در فایل outputs.zip:

- تعداد کل نمونه‌ها 720 است.
- تعداد زبان‌ها 4 است و توزیع داده در هر زبان یکنواخت است: برای هر زبان 180 نمونه.
- داده به دو حالت اصلی تقسیم شده است:
  - حالت no_augmentation: X_{train} با ابعاد $(576, 86)$ و X_{test} با ابعاد $(144, 86)$.
  - حالت augmented: X_{train} با ابعاد $(3456, 86)$ و X_{test} با ابعاد $(144, 86)$.

۲.۲.۱ ویژگی‌ها

برای هر نمونه یک بردار ویژگی با تعداد 86 ویژگی ساخته شده است. شمارش نوع ویژگی‌ها (از روی نام ویژگی‌ها) به صورت زیر است:

- ویژگی‌های MFCC و مشتقات آن: 78 ویژگی
- ویژگی‌های طیفی spectral: 4 ویژگی
- ویژگی‌های zero-crossing: 2 ویژگی

- ویژگی‌های RMS: 2 ویژگی

۳.۲.۱ نرمال‌سازی داده (استانداردسازی)

داده‌ی خام ویژگی‌ها از قبل استانداردسازی شده است (میانگین نزدیک به صفر و انحراف معیار نزدیک به یک). برای مثال در حالت no_augmentation روی داده‌ی آموزش:

- بازه‌ی میانگین ویژگی‌ها تقریباً بین -5.80×10^{-15} تا 3.55×10^{-15}.
- بازه‌ی انحراف معیار ویژگی‌ها تقریباً بین 1.00 تا 1.00 (با خطای عددی بسیار کوچک).

این نکته مهم است، چون بسیاری از روش‌های فاصله‌محور (مثل K-Means و روش‌های چگالی‌محور) نسبت به مقیاس ویژگی‌ها حساس هستند.

۴.۲.۱ مدت زمان فایل‌های صوتی

ستون duration نشان می‌دهد:

- کمینه‌ی مدت: 39.09 ثانیه
- میانگین مدت: 63.20 ثانیه
- بیشینه‌ی مدت: 1537.77 ثانیه

(این ناهمگنی می‌تواند در کیفیت استخراج ویژگی و میزان نویز اثرگذار باشد.)

۳.۱ چارچوب ارزیابی و معیارها

۱.۳.۱ معیار Silhouette (بدون‌نظارت)

برای ارزیابی کیفیت خوشه‌بندی بدون استفاده از برچسب‌ها، از silhouette score استفاده شده است. برای هر نمونه i:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}},$$

که در آن $a(i)$ میانگین فاصله‌ی نمونه‌ی i تا اعضای خوشه‌ی خودش و $b(i)$ کمترین میانگین فاصله‌ی i تا نزدیک‌ترین خوشه‌ی دیگر است. مقدار نهایی، میانگین $s(i)$ روی همه‌ی نمونه‌هاست. تعریف رسمی در مستندات scikit-learn آمده است.[۱]

[۱]

۲.۳.۱ معیار Purity (ارزیابی با برچسب واقعی)

هرچند خوشه‌بندی بدون‌نظارت است، برای تحلیل هم‌خوانی خوشه‌ها با زبان واقعی از purity استفاده می‌کنیم:

$$\text{Purity} = \frac{1}{n} \sum_{j=1}^{k} \max_{\ell} |C_j \cap L_\ell| ,$$

که C_j مجموعه‌ی نقاط خوشه‌ی j و L_ℓ مجموعه‌ی نقاط با برچسب واقعی ℓ است. (در روش‌های دارای نویز، purity معمولاً روی نقاط غیرنویز گزارش می‌شود.)

۳.۳.۱ نسبت نویز در روش‌های چگالی‌محور

در DBSCAN و OPTICS نقاطی که عضو هیچ خوشه‌ای نشوند با برچسب -1 (نویز) مشخص می‌شوند. بنابراین نسبت نویز:

$$\text{NoiseRatio} = \frac{\#\{i : \text{label}(i) = -1\}}{n}.$$

در OPTICS با استخراج Xi نیز نقاط خارج از خوشه‌ها با -1 برچسب می‌خورند.[۲]

۴.۱ پیش‌پردازش و ساخت دو نسخه‌ی داده (Step1)

در Step1 برای هر حالت داده (no_augmentation و augmented) دو نسخه ساخته شده است:

- نسخه‌ی RAW: همان فضای اصلی با 86 ویژگی استانداردشده.
- نسخه‌ی OPTIMIZED: داده‌ی تبدیل‌یافته با کاهش‌بُعد به کمک PCA.

۱.۴.۱ بررسی هم‌بستگی و حذف ویژگی‌های با هم‌بستگی بالا

در این نسخه از خروجی‌ها، آستانه‌ی حذف هم‌بستگی 0.99 بوده و در هر دو حالت:

scikit-learn silhouette_score documentation: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html

۲ scikit-learn cluster_optics_xi documentation: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.cluster_optics_xi.html

- تعداد ویژگی‌های حذف‌شده 0 و تعداد ویژگی‌های نگه‌داشته‌شده 86 است.

نکته‌ی مهم: با وجود این، در محاسبه‌ی هم‌بستگی روی X_train مشاهده می‌شود که در حالت no_augmentation تعداد 4 جفت ویژگی با $|r| > 0.95$ وجود دارد (ولی هیچ جفتی به 0.99 نمی‌رسد)، لذا حذف انجام نشده است.

۲.۴.۱ کاهش‌بُعد با PCA

در Step1، PCA با هدف حفظ حدود 90% واریانس اعمال شده است:

- در حالت no_augmentation: تعداد مؤلفه‌ها 30 و مجموع نسبت واریانس توضیح‌داده‌شده 0.9052.
- در حالت augmented: تعداد مؤلفه‌ها 32 و مجموع نسبت واریانس توضیح‌داده‌شده 0.9007.

در این خروجی‌ها گزینه‌ی whiten برابر False بوده است؛ بنابراین PCA صرفاً یک نگاشت خطی برای بیشینه‌سازی واریانس روی مؤلفه‌هاست و مقیاس‌دهی سفیدسازی انجام نشده است.

۵.۱ خوشه‌بندی با K-Means (Step2)

۱.۵.۱ مدل و تابع هدف

الگوریتم K-Means داده‌ها را به k خوشه تقسیم می‌کند و مراکز μ_j را طوری می‌یابد که مجموع مربعات فاصله‌ها کمینه شود:

$$J = \sum_{i=1}^{n} \left\| x_i - \mu_{c(i)} \right\|_2^2 .$$

در scikit-learn این مقدار با نام inertia گزارش می‌شود.[۳]

۲.۵.۱ روش انتخاب k

در خروجی‌های Step2، مقدار k در بازه‌ی $\{2, \ldots, 9\}$ پیمایش شده و برای انتخاب بهترین k از بیشینه‌سازی silhouette استفاده شده است. همچنین نمودار elbow (بر اساس inertia) برای

[۳] scikit-learn KMeans documentation: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

مشاهده‌ی تغییرات تابع هدف ذخیره شده است.

۳.۵.۱ نتایج عددی (بهترین k)

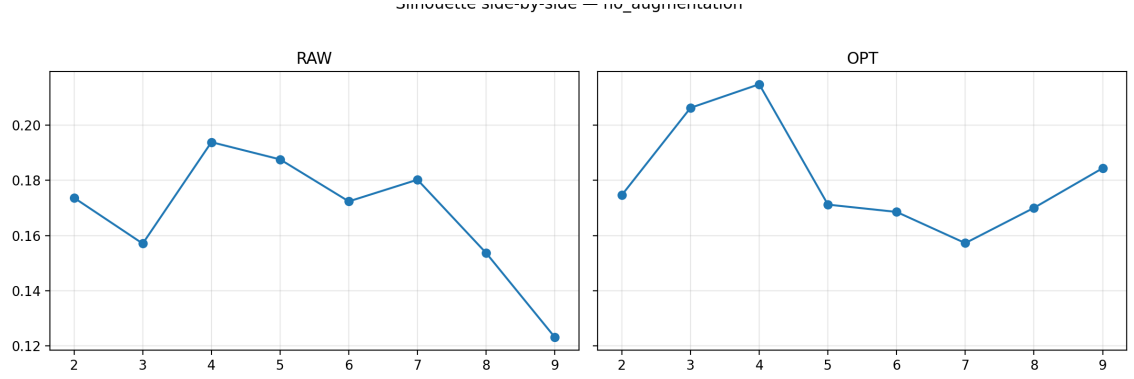
جدول زیر از فایل‌های summary.json در Step2 استخراج شده است:

جدول ۱: نتایج بهترین تنظیم K-Means بر اساس بیشینه‌سازی silhouette (Step2)

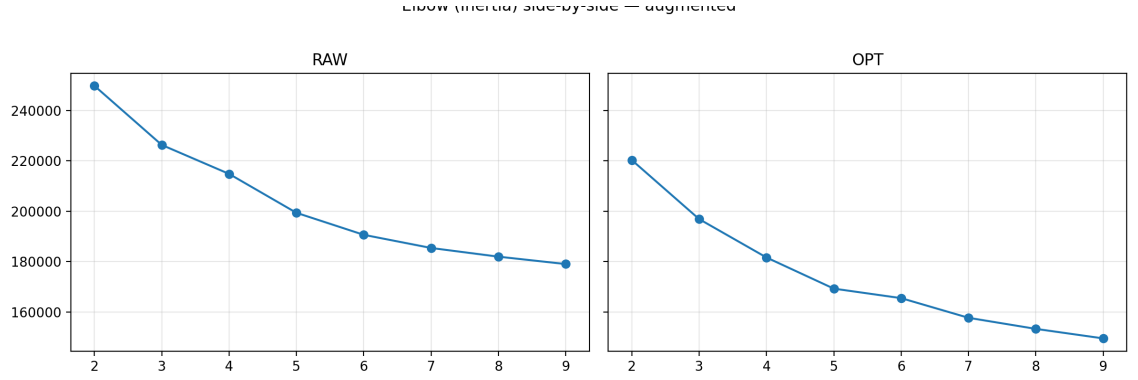
Purity	Silhouette	$k^\star$	نسخه	حالت داده
0.6181	0.1938	4	RAW	no_augmentation
0.6181	0.2148	4	OPT	no_augmentation
0.5483	0.1460	5	RAW	augmented
0.4627	0.1583	4	OPT	augmented

تفسیر.

- در حالت no_augmentation، هر دو نسخه‌ی RAW و OPT بهترین k را برابر 4 برمی‌گردانند (هم‌خوان با تعداد زبان‌ها) و silhouette در نسخه‌ی OPT بهتر شده است.
- در حالت augmented، نسخه‌ی RAW بهترین k را 5 می‌دهد، اما نسخه‌ی OPT بهترین k را 4 بازمی‌گرداند و silhouette نیز بهتر می‌شود؛ با این حال purity در نسخه‌ی OPT کاهش یافته است. این می‌تواند نشان دهد که کاهش‌بُعد با PCA هندسه‌ی خوشه‌ها را از نظر فاصله‌ای بهتر کرده، اما هم‌راستایی خوشه‌ها با زبان واقعی را (به علت حذف مؤلفه‌های کم‌واریانس ولی شاید تمایزبخش) کاهش داده است.



شکل ۱: مقایسه‌ی silhouette بر حسب k برای K-Means در RAW و OPT (حالت no_augmentation).



شکل ۲: نمودار elbow (بر اساس inertia) برای K-Means در RAW و OPT (حالت augmented).

شکل ۳: نمای PCA2: برچسب واقعی زبان در کنار برچسب خوشه‌ی K-Means (نسخه‌ی RAW).

۶.۱ خوشه‌بندی چگالی‌محور با DBSCAN (Step3)

۱.۶.۱ ایده‌ی اصلی و تعریف نقاط هسته/مرزی/نویز

الگوریتم DBSCAN خوشه‌ها را بر اساس چگالی تعریف می‌کند. برای یک نقطه x، همسایگی ε به صورت

$$\mathcal{N}_\varepsilon(x) = \{y : \ d(x,y) \le \varepsilon\}$$

تعریف می‌شود. اگر $|\mathcal{N}_\varepsilon(x)| \ge \text{min\_samples}$، نقطه **هسته** است و خوشه با اتصال نقاط هسته توسعه می‌یابد. نقاطی که به هیچ خوشه‌ای نپیوندند **نویز** بوده و با برچسب -1 مشخص می‌شوند.[۴]

[۴] scikit-learn DBSCAN documentation:
https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html

۲.۶.۱ روش جست‌وجوی پارامترها

در Step3 :

- روی min_samples یک شبکه از مقادیر گسسته بررسی شده است.
- برای هر min_samples، با استفاده از منحنی k-distance و چند percentile، مقادیر کاندید برای eps تولید شده است.
- silhouette روی نقاط غیرنویز محاسبه شده و در انتخاب تنظیم بهتر، علاوه بر نزدیک بودن تعداد خوشه‌ها به 4، محدودیت روی نسبت نویز نیز در نظر گرفته شده است.

۳.۶.۱ نتایج عددی بهترین تنظیم‌ها

(مقادیر از فایل‌های best_params.json در Step3 استخراج شده‌اند.)

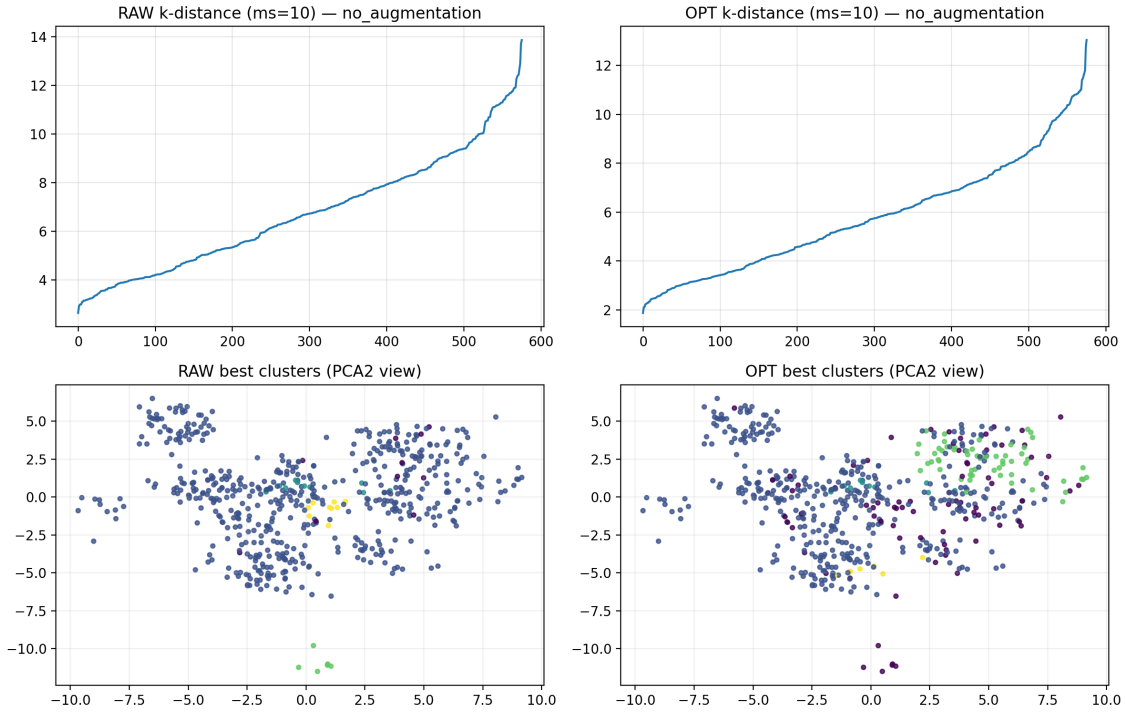
جدول ۲: بهترین تنظیم‌های DBSCAN و کیفیت (Step3)

$\text{Silhouette}_{\neg\text{noise}}$	NoiseRatio	#clusters	min_samples	ε	نسخه	حالت
0.1755	0.0174	4	5	10.0007	RAW	no_augmentation
0.2138	0.1406	4	5	6.9179	OPT	no_augmentation
0.1617	0.2338	4	15	6.5405	RAW	augmented
0.1571	0.2338	4	15	5.6557	OPT	augmented

جدول ۳: Purity روی نقاط غیرنویز و دامنه‌ی اندازه‌ی خوشه‌ها در DBSCAN (Step3)

$\max \lvert C_j \rvert$	$\min \lvert C_j \rvert$	$\text{Purity}_{\neg\text{noise}}$	نسخه	حالت
541	6	0.2968	RAW	no_augmentation
420	7	0.4263	OPT	no_augmentation
2341	30	0.4060	RAW	augmented
2340	30	0.4052	OPT	augmented

تفسیر. در همه‌ی حالت‌ها DBSCAN توانسته تعداد خوشه‌ها را 4 تولید کند؛ اما purity نسبتاً پایین است. همچنین در حالت augmented یک خوشه‌ی بسیار بزرگ (بیش از 2000 نقطه) دیده می‌شود که می‌تواند نشانه‌ی ادغام چند زبان/گوینده در یک ناحیه‌ی چگال یا انتخاب eps نسبتاً بزرگ باشد.



شکل ۴: مقایسه‌ی نمای PCA2 خوشه‌های بهترین DBSCAN برای RAW و OPT.

۷.۱ خوشه‌بندی با OPTICS (Step4)

۱.۷.۱ ایده‌ی اصلی و نقش reachability

الگوریتم OPTICS نیز چگالی‌محور است، اما به جای تعیین یک eps ثابت، ساختار چگالی را در مقیاس‌های مختلف ثبت می‌کند و خروجی کلیدی آن reachability distance است. با رسم reachability plot معمولاً «دره‌ها» متناظر با خوشه‌ها تفسیر می‌شوند. نمونه و توضیح رسمی در مثال‌های scikit-learn موجود است.[۵]

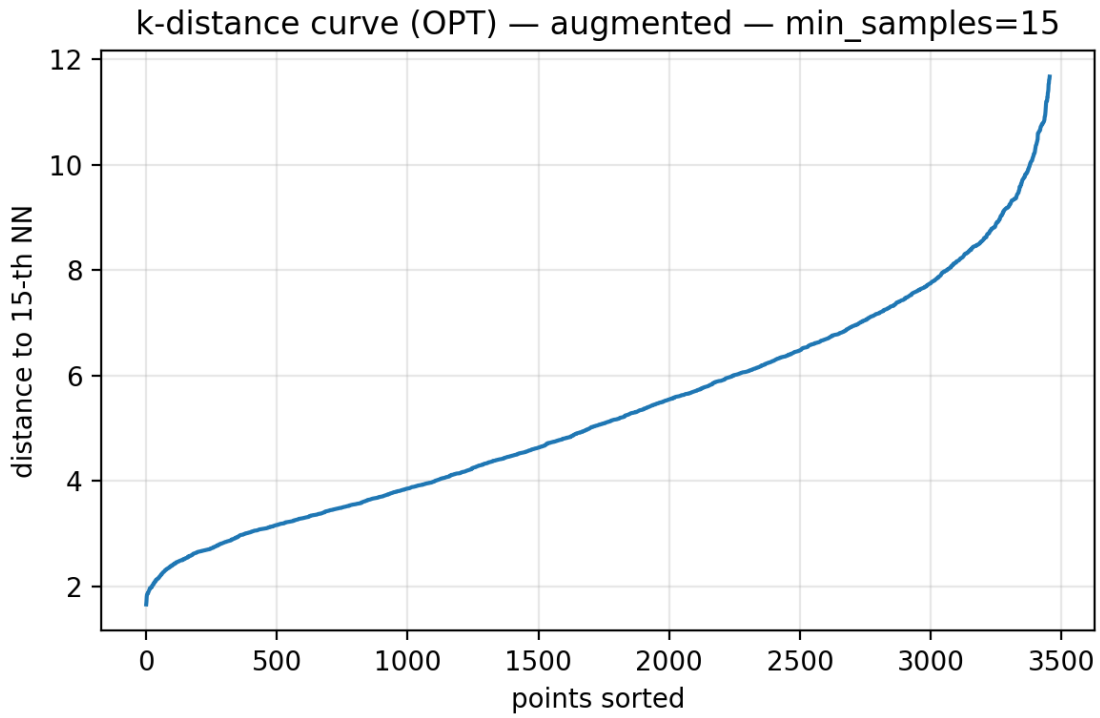
۲.۷.۱ استخراج خوشه‌ها با روش Xi

در این پروژه خوشه‌ها با روش Xi استخراج شده‌اند که پارامتر xi حداقل تندیِ لازم روی reacha-bility plot برای مرز خوشه را تعیین می‌کند و نقاط خارج از خوشه‌ها با -1 برچسب می‌خورند.[۶]

---

۵ scikit-learn OPTICS example (reachability plot): https://scikit-learn.org/stable/auto_examples/cluster/plot_optics.html

۶ scikit-learn cluster_optics_xi documentation: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.cluster_optics_xi.html



شکل ۵: منحنی k-distance برای تخمین eps (نمونه، حالت augmented-OPT).

۳.۷.۱ نتایج عددی بهترین تنظیم‌ها

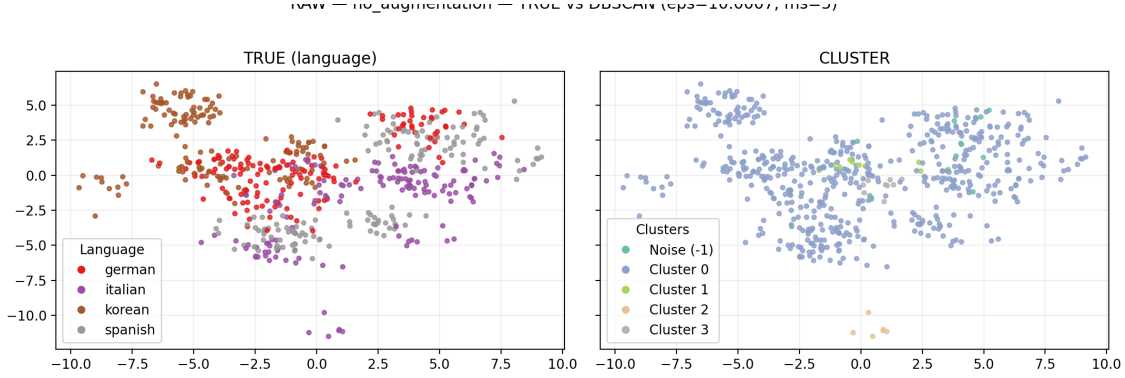
(مقادیر از فایل‌های best_params.json در Step4 استخراج شده‌اند.)

جدول ۴: بهترین تنظیم‌های OPTICS و کیفیت (Step4)

حالت	نسخه	min_samples	ξ	#clusters	NoiseRatio	$\text{Silhouette}_{\neg\text{noise}}$
no_augmentation	RAW	15	0.08	4	0.8142	0.4298
no_augmentation	OPT	20	0.05	4	0.7569	0.5578
augmented	RAW	5	0.03	2	0.9248	0.4869
augmented	OPT	5	0.03	2	0.8843	0.4704

تفسیر.

- در حالت no_augmentation، OPTICS تعداد خوشه را 4 بازمی‌گرداند و روی نقاط غیرنویز، purity تقریباً کامل است؛ اما نسبت نویز بسیار بالا است (بیش از 0.75). این یعنی الگوریتم فقط هسته‌های بسیار متراکم را به عنوان خوشه تشخیص داده و بخش بزرگی از داده را نامطمئن (نویز) فرض کرده است.
- در حالت augmented، بهترین تنظیم‌ها فقط 2 خوشه را استخراج کرده‌اند و نسبت

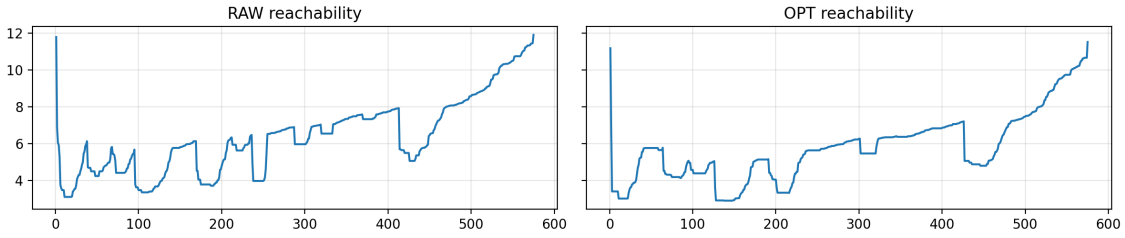


شکل ۶: نمای PCA2: برچسب واقعی زبان در کنار برچسب خوشه‌ی DBSCAN (نسخه‌ی RAW).

جدول ۵: Purity روی نقاط غیرنویز در OPTICS (Step4)

حالت	نسخه	$\text{Purity}_{\neg\text{noise}}$	توضیح کلیدی
no_augmentation	RAW	1.0000	نویز بسیار زیاد، خوشه‌های غیرنویز بسیار خالص
no_augmentation	OPT	0.9929	نویز زیاد، بهبود silhouette در OPT
augmented	RAW	1.0000	فقط 2 خوشه + نویز بسیار زیاد
augmented	OPT	0.4125	هم تعداد خوشه کم، هم هم‌خوانی پایین‌تر

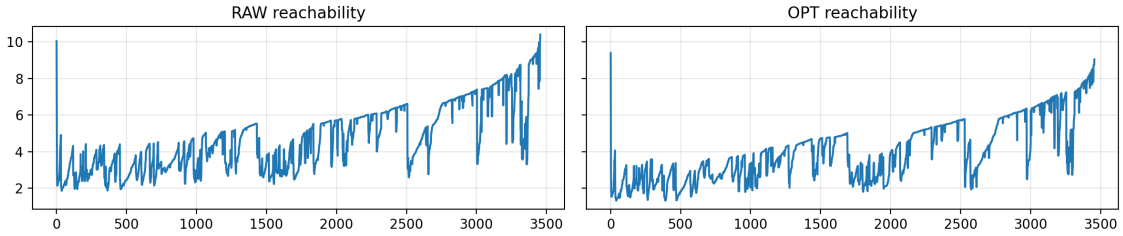
نویز حتی بالاتر است. این می‌تواند نشان دهد که افزایش تنوع داده، ساختار چگالی را طوری تغییر داده که روش استخراج Xi «چهار دره‌ی پایدار» روی reachability plot پیدا نمی‌کند و خوشه‌بندی محافظه‌کارانه می‌شود.



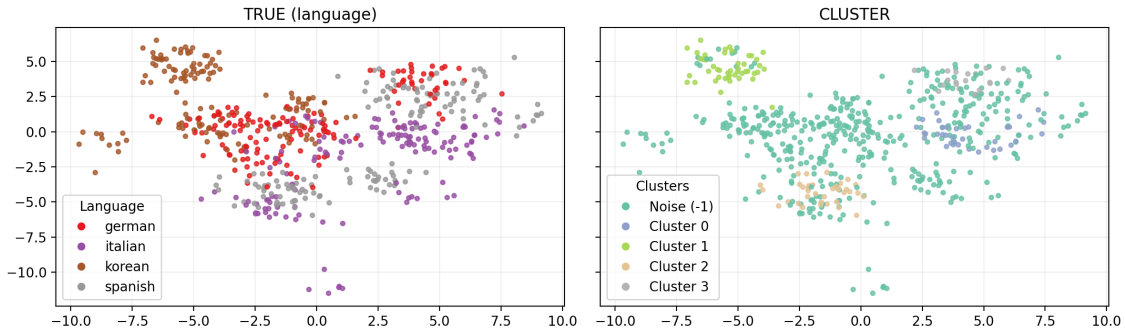
شکل ۷: مقایسه‌ی reachability plot برای RAW و OPT (حالت no_augmentation).

۸.۱ جمع‌بندی مقایسه‌ای تا پایان Step4

برای نتیجه‌گیری، یک دید خلاصه از خروجی هر الگوریتم ارائه می‌کنیم. در این جدول، برای K-Means نسبت نویز تعریف نمی‌شود (علامت —)، اما برای DBSCAN و OPTICS گزارش شده است.



شکل ۸: مقایسه‌ی reachability plot برای RAW و OPT (حالت augmented).



شکل ۹: نمای PCA2: برچسب واقعی زبان در کنار برچسب خوشه‌ی OPTICS (نسخه‌ی OPT).

جدول ۶: مقایسه‌ی خلاصه‌ی بهترین خروجی‌ها در سه الگوریتم (تا 4Step)

NoiseRatio	Purity	Silhouette	معیار انتخاب	#clusters	نسخه	حالت	الگوریتم
—	0.6181	0.1938	silhouette max	4	RAW	no_augmentation	K-Means
—	0.6181	0.2148	silhouette max	4	OPT	no_augmentation	K-Means
—	0.5483	0.1460	silhouette max	5	RAW	augmented	K-Means
—	0.4627	0.1583	silhouette max	4	OPT	augmented	K-Means
0.0174	0.2968	0.1755	نزدیک به 4 + نویز کم	4	RAW	no_augmentation	DBSCAN
0.1406	0.4263	0.2138	نزدیک به 4 + نویز کم	4	OPT	no_augmentation	DBSCAN
0.2338	0.4060	0.1617	نزدیک به 4 + نویز کنترل‌شده	4	RAW	augmented	DBSCAN
0.2338	0.4052	0.1571	نزدیک به 4 + نویز کنترل‌شده	4	OPT	augmented	DBSCAN
0.8142	1.0000	0.4298	silhouette + نزدیک به 4	4	RAW	no_augmentation	OPTICS
0.7569	0.9929	0.5578	silhouette + نزدیک به 4	4	OPT	no_augmentation	OPTICS
0.9248	1.0000	0.4869	بهترین در شبکه‌ی Xi	2	RAW	augmented	OPTICS
0.8843	0.4125	0.4704	بهترین در شبکه‌ی Xi	2	OPT	augmented	OPTICS

۱.۸.۱ نتیجه‌گیری نهایی

- اگر هدف اصلی «بازگرداندن تعداد خوشه‌ی 4» باشد، در این خروجی‌ها K-Means (به‌ویژه در no_augmentation و نیز augmented-OPT) و همچنین DBSCAN (در

همه‌ی حالت‌ها) تعداد خوشه‌ی 4 را به دست داده‌اند.

- OPTICS در no_augmentation تعداد خوشه‌ی 4 را می‌دهد، اما با هزینه‌ی نویز بسیار زیاد. بنابراین باید در تفسیر، همزمان به NoiseRatio توجه کرد (نویز زیاد می‌تواند باعث «خالص شدن» خوشه‌های باقی‌مانده و افزایش ظاهری purity روی نقاط غیرنویز شود).
- در حالت augmented، OPTICS (با استخراج Xi) به 2 خوشه رسیده است؛ که نشان می‌دهد ساختار چگالیِ داده‌ی افزایش‌یافته برای استخراج Xi به شکل فعلی مناسب نبوده یا نیازمند شبکه‌ی پارامتری گسترده‌تر است.