AI_A3 سیلان کوشکی حیری A1o1o1496

پیش پردازش و استخراج ویژگی ها:

۱- استخراج ویژگی ها از تصاویر به جای استفاده مستقیم از پیکسل ها، مزایای متعددی دارد. این کار به ویژه در پروژه های مانند clustering یا Machine learning اهمیت دارد.

① کاهش پیچیدگی داده ها: تصاویر خام شامل مقادیر زیادی از پیکسل هستند که بسیاری از آن ها ممکن است برای تحلیل های سطح بالا غیر ضروری یا غیر مفید باشند. با استخراج ویژگی ها، پیچیدگی داده ها کاهش می یابد و به طور معمول مجموعه ای از ویژگی های مختصر و متمرکز تر به دست می آید.

② تاکید بر ویژگی های مهم بصری: مدل های عمیق مانند VGG16 با استفاده از لایه های کانولوشن، الگوها و ساختارهای مهم بصری را شناسایی می کنند. این ویژگی ها اغلب مواردی مانند لبه ها، بافت ها، اشکال و سایر اطلاعات معنی دار را نشان می دهند. این فرآیند به الگوریتم های machine learning اجازه می دهد تا به جای پیکسل های خام بر ویژگی های برجسته تر و معنادارتر تمرکز کنند.

③ استفاده از ویژگی های استخراج شده برای اهداف مختلف: با استخراج ویژگی ها می توانید از آن ها در کاربردهای مختلف مانند clustering، طبقه بندی، کاهش ابعاد یا بازشناسی استفاده کنید.

④ پایداری نسبت به تغییرات و تبدیل ها: ویژگی های استخراج شده اغلب در برابر تغییرات جزئی در تصاویر مقاومت بیشتری دارند. این به مدل ها کمک می کند تا از ویژگی های استوار برای تصمیم گیری استفاده کنند.

⑤ ارتباط با دانش قبلی: استخراج ویژگی ها بر اساس مدل هایی مانند VGG16 که با مقادیر عظیمی از داده های آموزشی تمرین شده اند، به شما این امکان را می دهد از دانشی که این مدل از قبل دارند استفاده کنید. این یک راه کارآمد برای استفاده از ویژگی های استخراج شده است.

در کل، استخراج ویژگی ها از تصاویر به شما اجازه می دهد تا با داده های متمرکزتر، تحلیل های دقیق تر و مناسب تری کار کنید.

۲- استخراج ویژگی از تصاویر یک فرآیند کلیدی در پردازش تصویر و بینایی ماشین است که طی آن اطلاعات معنادار و قابل استفاده از تصاویر خام استخراج می شود. این فرآیند به تحلیلگران و مدل های یادگیری ماشین امکان می دهد تا به جای کار با داده های خام پیکسلی، با مجموعه ای از ویژگی های مفیدتر و معنادارتر کار کنند. ویژگی ها می توانند شامل اشکال، بافت ها، لبه ها، رنگ ها و سایر اطلاعات بصری مهم باشند.

۳- تکنیک رایج برای استخراج ویژگی ها از تصاویر

① استخراج ویژگی های کانولوشن Convolutional Feature Extraction: این روش بر پایه شبکه های عمیق کانولوشن (CNN)

Subject: Year: Month: Date: ( )

است CNN ها با استفاده از لایه‌های کانولوشن ویژگی‌های مهمی از تصاویر را استخراج می‌کنند این ویژگی‌ها ممکن است شامل لبه‌ها، بافت‌ها، اشکال و ساختارهای سطح مختلف باشند.

یکی از مدل‌های معروف برای استخراج ویژگی‌های کانولوشن مدل VGG16 است که با لایه‌های متراکم کانولوشن و ماکس پولینگ، ویژگی‌های مهم را استخراج می‌کند. سایر مدل‌های مشهور شامل Inception, ResNet, MobileNet هستند

② روش‌های مبتنی بر بافت (Texture_Based Feature Extraction): این روش‌ها برای استخراج اگاه‌ها و ساختار‌های بافتی در تصاویر استفاده می‌شوند. بافت‌ها اطلاعات بصری راجع به سطح و هزینات سطوح ارائه می‌دهند. تکنیک‌های مانند Gray-level Co-occurrence Matrix، Local Binary patterns، Histogram of Oriented Gdm از جمله روش‌های هستند که برای استخراج ویژگی‌های بافتی استفاده می‌شوند.

③ استخراج ویژگی‌های مبتنی بر رنگ (Color_Based_Feature extraction): این تکنیک‌ها براساس اطلاعات رنگی در تصاویر کار می‌کنند. ویژگی‌های رنگی می‌توانند به صورت مقادیر میانگین یا توزیع‌های رنگی مرکز اصلی مدل RGB یا سایر فضاهای رنگی مانند HSV یا هاما استخراج شوند. یکی از تکنیک‌های رایج در این حوزه استفاده از هیستوگرام رنگ است که توزیع رنگ‌ها را در تصویر مشخص می‌کند. این روش‌ها اغلب برای تحلیل تصاویر رنگی و شناسایی اشیا یا بافت‌های مختلف در تصاویر استفاده می‌شوند.

۳- برای استفاده از مدل VGG16 برای استخراج ویژگی‌های تصاویر گل، چند مرحله پیش‌پردازش نیاز است. ابتدا تصاویر باید به اندازه و فرمت مناسب تبدیل شوند و سپس مدل ساخته و آماده‌سازی شود تا بتوانید به مدل VGG16 وارد شوند.

① بارگذاری تصاویر: تصاویر را از پوشه باید بخوانیم

② تبدیل اندازه تصاویر: تصاویر VGG16 معمولاً به اندازه 224×224 نیاز دارند

③ نرمال‌سازی داده‌ها: مدل VGG16 نیاز به داده‌هایی با دامنه مقادیر مشخص (۰ تا ۲۵۵) دارند که معمولاً به شکل فلوات نرمال‌سازی می‌شوند. داده‌ها را به آرایه‌های numpy تبدیل کنیم تا آماده ورود به مدل شوند

④ استخراج ویژگی‌ها با استفاده از مدل VGG16: مدل VGG16 را با حذف لایه‌های بالا (Fully connected) بارگذاری کنیم تا فقط ویژگی‌ها را استخراج کنیم.

★ در مدل VGG16 بخش پایینی (کانولوشن) وظیفه استخراج ویژگی‌های کلان و بصری را دارد، در حالی که لایه‌های Fully connected وظیفه طبقه‌بندی نهایی را به عهده دارند. با حذف لایه‌های Fully connected می‌توانید از ویژگی‌های بصری استخراج شده برای clustering استفاده کنیم.

Subject: Year: Month: Date: ( )

«یادداشت خوشه‌بندی»

۱- clustering یک روش یادگیری بدون نظارت است که هدف آن گروه‌بندی اشیاء داده‌ها بر اساس ویژگی‌های مشابه آن‌هاست. دو روش معروف در clustering، K-Means و DBSCAN است. در ادامه توضیح هر کدام را می‌گویم.

«K-Means»

K-means یکی از روش‌های پرکاربرد خوشه‌بندی است که هدف آن این است که داده‌ها به K خوشه تقسیم شوند. الگوریتم به طور معمول به صورت زیر عمل می‌کند:

۱- انتخاب اولیه مراکز خوشه‌ها: الگوریتم با انتخاب K نقطه به عنوان مراکز خوشه شروع می‌شود.

2- اختصاص داده‌ها: نزدیک‌ترین مرکز خوشه: داده‌ها بر اساس فاصله به نزدیک‌ترین مرکز خوشه تخصیص داده می‌شوند.

3- تغییر مراکز خوشه‌ها: مراکز خوشه‌ها بر اساس میانگین نقاط تخصیص یافته به هر خوشه به روزرسانی می‌شوند.

4- تکرار تا زمانی که خوشه‌ها تثبیت شوند: این مراحل تکرار می‌شوند تا زمانی که تغییرات مراکز خوشه‌ها قابل اغماض باشد.

* مزایا: ① سادگی: K-Means یک الگوریتم نسبتاً ساده و پیاده‌سازی است که به راحتی قابل اجرا است.

② کارایی: به ویژه برای داده‌های بزرگ، K-Means به دلیل سادگی و کارایی‌اش محبوب است.

③ قابلیت تفسیر: خوشه‌های ایجاد شده توسط K-Means به راحتی قابل تفسیر و تحلیل هستند.

* معایب: ① نیاز به تعیین تعداد خوشه‌ها (K): نیاز به مشخص کردن تعداد خوشه‌ها دارد که این ممکن است از پیش شناخته شده نباشد.

② حساسیت به مقدار اولیه: نتایج K-Means می‌تواند به انتخاب اولیه مراکز اولیه خوشه‌ها حساس باشد.

③ فرضیه ساختار کروی: K-Means فرض می‌کند که خوشه‌ها به صورت کروی و با اندازه‌های مشابه هستند، که ممکن است در همه مقادیر صادق نباشد.

④ حساسیت به داده‌های Noise: داده‌های Noise می‌تواند به طور قابل توجهی روی نتایج K-means تأثیر بگذارد.

«DBSCAN»

DBSCAN یک الگوریتم خوشه‌بندی مبتنی بر چگالی است. در این روش، cluster ها بر اساس مناطق با چگالی بالا تشکیل می‌گیرند. الگوریتم به طور معمول به صورت زیر عمل می‌کند:

۱- شناسایی نقاط پرتراکم: نقاطی که تعداد معینی از همسایگان در یک شعاع مشخص (ε) هستند به عنوان نقاط پرتراکم شناخته می‌شوند.

۲- گسترش خوشه‌ها: خوشه‌ها با گسترش از نقاط پرتراکم به همسایگان آن‌ها تشکیل می‌شوند.

۳- نقاط Noise: داده‌هایی که به هیچ خوشه‌ای تعلق ندارند به عنوان داده Noise ثبت می‌شوند.

Caspian

Subject: Year: Month: Date: ( )

* مزایا: ① عدم نیاز به تعداد خوشه‌ها از پیش: DBSCAN نیازی به تعیین تعداد خوشه‌ها ندارد.
② مدیریت داده‌های Noise: DBSCAN می‌تواند داده‌های Noise را شناسایی کند و از cluster ها جدا کند.
③ قابلیت خوشه‌بندی ساختارهای پیچیده: DBSCAN قادر به شناسایی خوشه‌های با شکل‌های غیر استاندارد است.

* معایب: ① نیاز به تنظیم پارامترها: DBSCAN به تنظیم ε (شعاع) و تعداد نقاط حداقل در یک خوشه نیاز دارد که ممکن است در ابتدا مشخص نباشند
② حساسیت به چگالی: اگر چگالی داده‌ها در مناطق مختلف متفاوت باشد، ممکن است DBSCAN نتواند خوشه‌های مناسب را شناسایی کند
③ کارایی برای داده‌های بسیار بزرگ: برای داده‌های بسیار بزرگ، DBSCAN ممکن است کندتر از K-Means یا DBSCAN باشد.

** انتخاب بین K-Means و DBSCAN:
انتخاب بین این دو روش بستگی به طبیعت داده‌ها و نیازهای پروژه دارد. K-means برای داده‌های با ساختار کروی و خوشه‌های با اندازه‌ی تقریباً مساوی مناسب است. در مقابل DBSCAN برای داده‌های با ساختارهای پیچیده‌تر و نیاز به شناسایی نقاط پرت انتخاب بهتری دارد.

ه- برای پیدا کردن تعداد مناسب خوشه‌ها (K) در روش K-Means روش‌های مختلفی وجود دارد. یکی از رایج‌ترین روش‌ها Elbow است. این روش شامل اجرای K-means با تعداد مختلفی از خوشه‌ها است و سپس بررسی می‌شود که کدام تعداد خوشه‌ها چگونه بر معیارهایی مانند جمع مربعات خطا تأثیر می‌گذارد. در روش Elbow شما مقادیر مختلف K الگوریتم K-means را اجرا می‌کنید و سپس مقدار SSE را برای هر مقدار K محاسبه می‌کنید و وقتی که کاهش SSE به طور قابل توجهی کمتر می‌شود این نقطه ممکن است مقدار مناسب K باشد.

* من در کنار استفاده از الگوریتم elbow، نمره سیلوئت را هم اضافه کردم که بهترین K مناسب را انتخاب کند. حال اصلاً نمره سیلوئت چیست؟

نمره سیلوئت یک معیار برای ارزیابی کیفیت خوشه‌بندی است که نشان می‌دهد چگونه داده‌های موجود در یک خوشه به هم نزدیک هستند و چقدر از خوشه‌های دیگر فاصله دارند. به عبارت دیگر این نمره به ما کمک می‌کند تا بفهمیم که داده‌ها چقدر به خوشه‌های خود وابسته هستند و چقدر خوشه‌ها از یکدیگر جدا شده‌اند.

نحوه محاسبه نمره سیلوئت ← a: میانگین فاصله‌ی بین خوشه‌ای: میانگین فاصله یک نمونه با سایر نمونه‌های موجود در خوشه‌ی خودش

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

b: میانگین فاصله‌ی نمونه با نزدیک‌ترین خوشه‌ی دیگر

این نمره بین ۱- و ۱ است. نمره سیلوئت بالا (نزدیک به ۱) نشان می‌دهد که نمونه به طور کامل در خوشه خود قرار گرفته و خوشه‌ها به خوبی

Subject:
Date

از یکدیگر جدا شده‌اند. نمره نزدیک به ۰ نشان دهنده این است که نمونه در مرز خوشه قرار گرفته و نمره منفی نشان دهنده این است که نمونه به خوشه دیگری نزدیک‌تر است.

کاربرد نمره سیلوئت ← ارزیابی کیفیت خوشه‌بندی: این نمره نشان می‌دهد که چقدر خوشه‌ها از یکدیگر متمایز هستند و چقدر داده‌ها در خوشه خود پایبند هستند.

انتخاب تعداد خوشه‌های مناسب: می‌توان از نمره سیلوئت برای پیدا کردن تعداد بهینه خوشه‌ها استفاده کرد. مقادیر K که بیشترین نمره سیلوئت را دارد، ممکن است تعداد مناسب خوشه‌ها باشد.

"homogeneity"

معیار یکنواختی یا Homogeneity یکی از معیارهایی است که برای ارزیابی کیفیت خوشه‌بندی مورد استفاده قرار می‌گیرد. این معیار به ما می‌گوید که چقدر خوشه‌ها یکنواخت هستند و تا چه حد هر خوشه شامل عناصر یا کلاس (یا دسته‌بندی) مشابه است.

مفهوم homogeneity: معیار یکنواختی نشان می‌دهد که هر خوشه چقدر شامل داده‌هایی است که فقط از یک دسته‌بندی یا کلاس می‌آیند. به عبارت دیگر، این معیار می‌گوید که آیا همه عناصر در یک خوشه متعلق به یک کلاس هستند یا خیر.

نحوه محاسبه یکنواختی: معیار یکنواختی با استفاده از اطلاعات متقابل (Mutual Information) بین خوشه‌ها و کلاس‌های واقعی محاسبه می‌شود. سپس با اطلاعات کلاس‌های واقعی نرمال‌سازی می‌شود. نتیجه یک عدد بین ۰ و ۱ است:

* یک مقدار نزدیک به ۱ نشان می‌دهد که هر خوشه شامل داده‌هایی از یک کلاس است (خوشه‌ها یکنواخت هستند)
* یک مقدار نزدیک به ۰ نشان می‌دهد که خوشه‌ها حاوی داده‌هایی از کلاس‌های مختلف هستند (خوشه‌ها مخلوط هستند)

کاربرد homogeneity برای ارزیابی خوشه‌بندی: اگر داده‌های شما دارای کلاس‌های واقعی هستند (مثلاً اگر این کلاس‌ها برای مثال به عنوان برچسب‌های آموزشی نبوده‌اند) معیار یکنواختی می‌تواند به شما کمک کند تا کیفیت خوشه‌بندی را ارزیابی کنید.

** برای استفاده از این معیار، به داده‌هایی با برچسب نیاز داریم تا بتوانیم یکنواختی را اندازه‌گیری کنیم. این برچسب‌ها کلاس‌های واقعی داده‌ها را نشان می‌دهند.

محدودیت‌ها: معیار یکنواختی تنها زمانی مفید است که داده‌های شما دارای برچسب‌های واقعی باشند. این معیار برای داده‌های بدون برچسب کاربرد ندارد.

"DBSCAN clustering"

برای پیدا کردن ε در این الگوریتم ۲ روش وجود دارد که اولین روش آن مانند k-means که از نمره سیلوئت استفاده کرد و در این حالت بیشترین نمره سیلوئت در یک ε را می‌بینیم و آن بیشترین ε را انتخاب می‌کنیم.

Subject:
Date

روش دوم: رسم نمودار فاصله بین نقاط نزدیک‌ترین همسایه. با رسم نمودار فاصله بین نقاط نزدیک‌ترین همسایه‌ی توانیم بازای تغییرات در فواصل بین نقاط را بسنجیم. به طور مثال، یک بله یا تغییر ناگهانی در فاصله‌ها نشان دهنده بهترین ε است.

۶- تشنه خوشه‌ی K-means با K برابر با ۸ به صورت است که تعدادی خوشه در یک دسته مناسب قرار گرفته اند و بعضی جدین cluster دیگر در کنار هستند و به صورت نمودار پراکندگی آمده که آن توانسته این ها را cluster بندی کند البته به نظر می‌رسد که انتخاب K در این قسمت تأثیر به سزایی داشته است.

در نمودار DBSCAN او توانسته فقط چند cluster کم را تشخیص داده و برحسب visualize او بهتر عمل کرده است.
ولی این هم نیاز به ε دارد.

«کاهش بعد»

تحلیل مولفه‌های اصلی (PCA) یک تکنیک پرکاربرد در یادگیری ماشین و آمار است که برای کاهش بعد داده‌ها، تجزیه و تحلیل ویژگی‌ها و مصورسازی استفاده می‌شود. PCA با شناسایی مولفه‌های اصلی که بیشترین واریانس را در داده‌ها توضیح می‌دهند کار می‌کند. این تکنیک به ویژه زمانی مفید است که داده‌ها دارای تعداد زیادی ویژگی (ابعاد) باشند و ما بخواهیم آن‌ها را به تعداد کمتری از ویژگی‌های اصلی کاهش دهیم.

نحوه عملکرد PCA: PCA به این صورت عمل می‌کند که مولفه‌های اصلی را شناسایی می‌کند به طور خطی ترکیبی از ویژگی‌های اصلی هستند و بیشترین واریانس را دارند. در اینجا چند گام کلیدی که PCA انجام می‌دهد آورده شده است:

① مرکزیت داده‌ها: داده‌ها معمولاً مرکزیت سازی می‌کنند یعنی میانگین هر ویژگی از داده‌ها کسر می‌شود تا داده‌ها حول مبدا صفر توزیع شوند.

② محاسبه ماتریس کوواریانس: ماتریس کوواریانس نشان می‌دهد که چگونه ویژگی‌ها نسبت به یکدیگر تغییر می‌کنند PCA با محاسبه این ماتریس شروع می‌کند.

③ محاسبه بردارها و مقادیر ویژه: PCA بردارهای ویژه و مقادیر ویژه ماتریس کوواریانس را محاسبه می‌کند این بردارها جهت‌های مولفه‌های اصلی هستند و مقادیر ویژه میزان واریانس توضیح داده شده توسط هر مولفه اصلی را نشان می‌دهد.

④ انتخاب مولفه‌های اصلی: مولفه‌های اصلی به ترتیب کاهش واریانس مرتب می‌شوند. با انتخاب چند مولفه اول، می‌توانیم بیشترین واریانس داده‌ها را توضیح دهیم.

⑤ تبدیل داده‌ها به مولفه‌های اصلی: با استفاده از مولفه‌های اصلی انتخاب شده، داده‌ها به فضای جدیدی با ابعاد کمتر تبدیل می‌شوند

Subject:
Date

چرا از PCA استفاده می‌کنیم؟ PCA برای کاهش ابعاد داده‌ها بدون از دست دادن اطلاعات زیاد است. این کار مزایای مختلفی دارد: ① کاهش پیچیدگی: با کاهش تعداد ویژگی‌ها، محاسبات سریع‌تر و آسان‌تر می‌شوند.

② مصورسازی: PCA می‌تواند مصورسازی داده‌ها در فضای دوبعدی یا سه‌بعدی کمک کند.

③ کاهش نویز: حذف مولفه‌هایی که واریانس کمتری دارند ممکن است به کاهش نویز در داده‌ها کمک کند.

محدودیت‌های PCA:

فقدان تفسیرپذیری: مولفه‌های اصلی ممکن است تفسیرپذیری مشخصی نداشته باشند زیرا ترکیب خطی از ویژگی‌های اصلی هستند.

فرض بر خطی بودن: PCA بر اساس فرض خطی بودن روابط بین ویژگی‌ها عمل می‌کند. اگر روابط غیرخطی باشند این روش ممکن است موثر نباشد.

حساسیت به مقیاس: اگر ویژگی‌ها مقیاس‌های مختلفی داشته باشند نتایج PCA ممکن است تحت تاثیر قرار گیرد.

در داده بعد از انجام PCA می‌بینیم که بسیار بهتر شده‌اند cluster های انجام شده و این به ما کمک کرده است.

ارزیابی و تحلیل:

۸- در قسمت‌های قبل راجب این دو معیار صحبت شده است.

۹- کل کار را [illegible] نمره سیلوئت انجام داده‌ام.

۱۰- برای بهبود عملکرد مدل‌ها و بهینه‌سازی نتایج می‌توان از روش‌ها و استراتژی‌های مختلفی استفاده کرد. در اینجا به چند راهکار مهم برای افزایش عملکرد مدل‌های clustering می‌پردازیم:

① پیش‌پردازش داده‌ها: پاک‌سازی داده‌ها، پیدا کردن ویژگی‌های مرتبط

② مهندسی ویژگی: ایجاد ویژگی‌های جدید، انتخاب ویژگی، کاهش ابعاد

③ انتخاب و تنظیم مدل: تنظیم پارامترها، انتخاب مدل مناسب، استفاده از ترکیب مدل‌ها

④ ارزیابی و اعتبارسنجی

⑤ افزایش داده‌ها

⑥ استفاده از مدل‌های پیچیده‌تر