به نام خدا

گزارش فاز اول

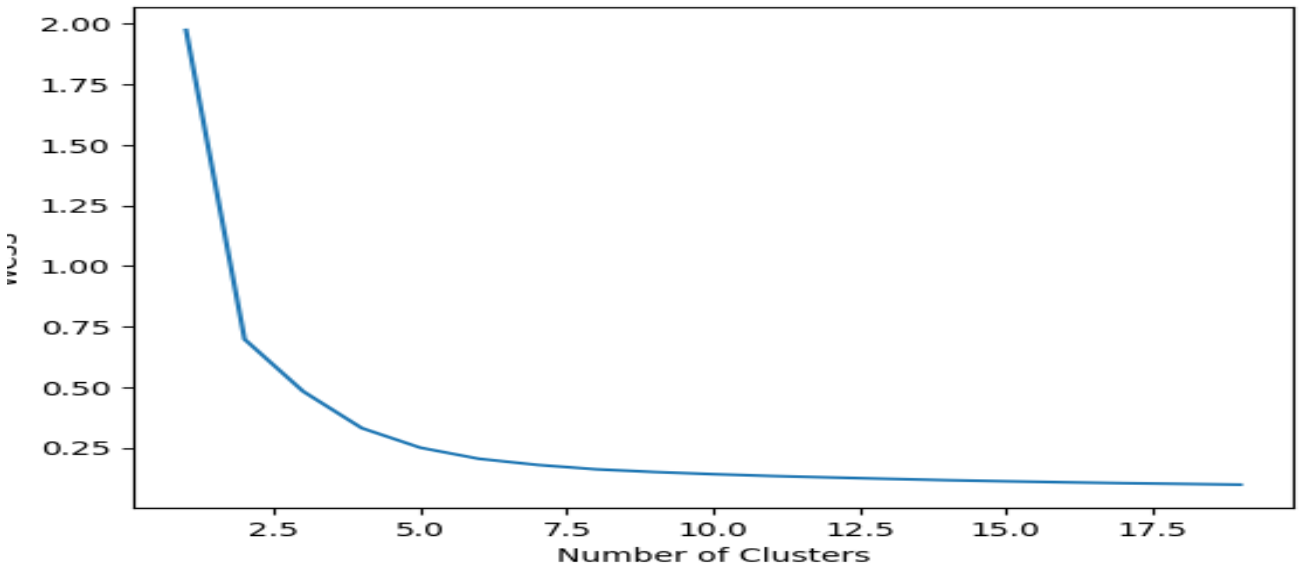
پروژه هوش محاسباتی

تینا توکلی-محمدهادی امینی

در این مسئله ما با داده هایی از تصاویر که لیبل شئ داخل تصویر مشخص است، روبرو هستیم .

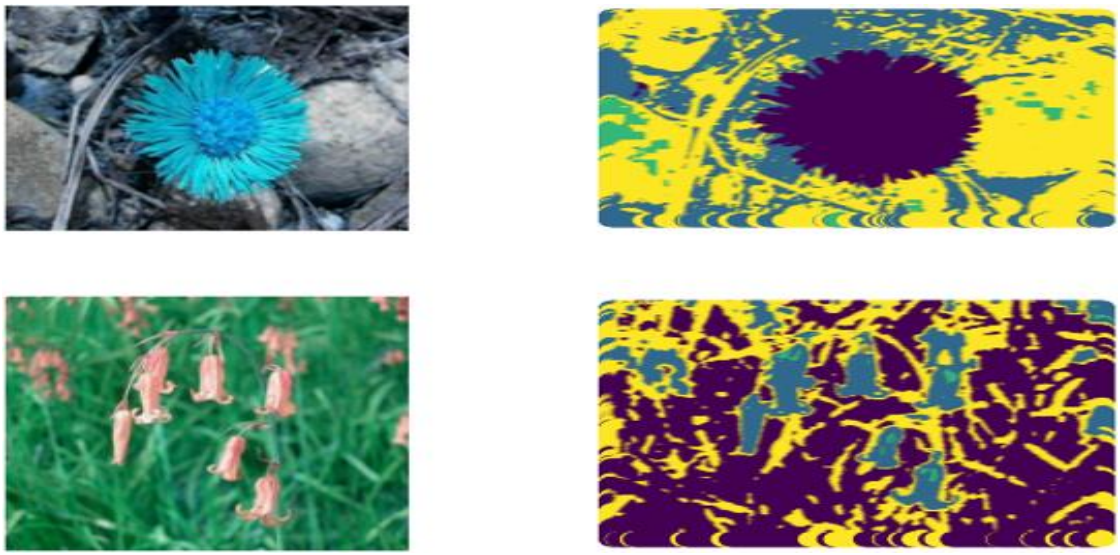
هدف خوشه بندی داده ها بر حسب دامین تصاویر است. منظور از دامین نوع تصویرموجود از آن شئ است.

خوشه بندی به صورتunsupervised است و از لیبل دامین داده ها اطلاعی نداریم.
چالش هایی که با آنها روبرو هستیم، خوشه بندی به شکل مناسب است و پس از آن پیدا کردن تعداد دامین ها و مپ کردن خوشه ها با لیبل های دامین تصاویر است.
به منظور خوشه بندی و ارزیابی داده ها، راهکارها و الگوریتم های مختلفی را امتحان کرده ایم .در برخی از الگوریتم ها همانند kmeans نیازمند دانستن تعداد مناسب خوشه ها هستیم .به این منظور راهکارهای مختلفی را امتحان میکنیم.
در یکی از شیوه های مورد استفاده قرار گرفته به این شکل پیش رفتیم که داده ها را بر حسب لیبل نوع شئ تصویرجدا کردیم و روی هر بخش جداگانه الگوریتم kmeans را با k های مختلف اجرا کردیم .و با WCSS تشخیص میدهیم که کدام k برای خوشه بندی مناسب تر است، چرا که به این شکل تعداد دامین ها را در داده های محدود بهتر میتوان مشخص کرد(با for برای تعدادی عکس و در رنج 20تا پلات میکند تا بهترین k را به دست بیاوریم با توجه به بیشترین شکستگی)



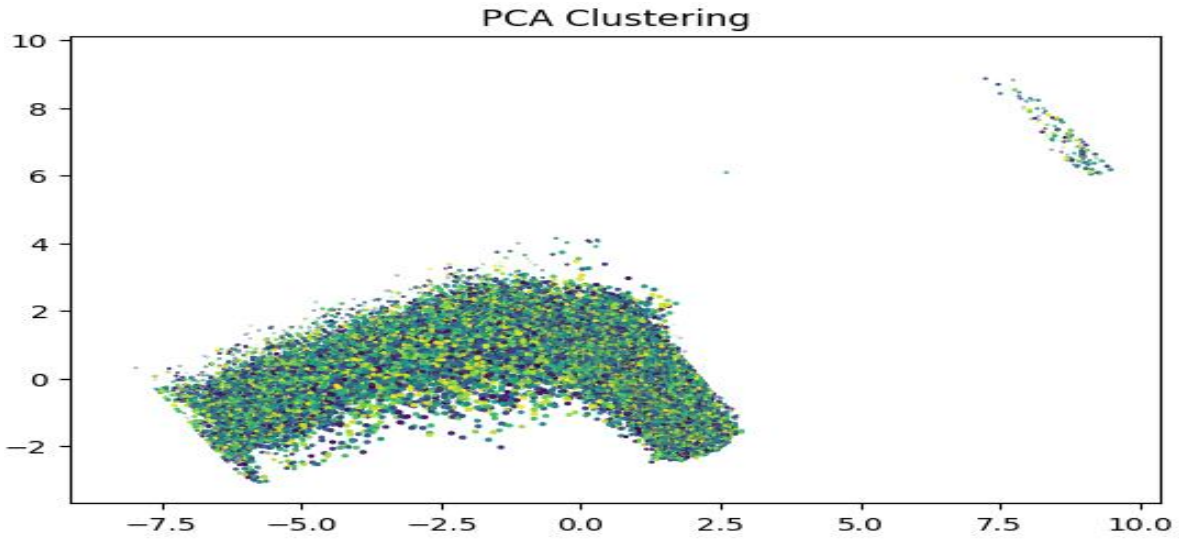
(پیدا کردن *k* مناسب روی کل داده)

نتایج نشان میدهد که در نیمی k برابر با 4 و در نیمی دیگر k برابر با3 ، مطلوب بوده است .این مورد نشان میدهد که میتوان تخمینی از4 دامین داشت.
نتیجه تست ها نشان میدهد برای داده های بیشتر هم همان k برابر با 4 مطلوب است.



(نمونه ای از کلاسترینگ بخش اول)

PCA: تجزیه و تحلیل مولفه های اساسی برای کاهش ابعاد مجموعه های بزرگ داده که داده ها دارای همبستگی هستند به کار میرود . الگوهای موجود در یک مجموعه داده را مشخص کنید تا بتوان بدون از بین رفتن هرگونه اطلاعات مهم به مجموعه ای از داده ها با ابعاد کمتری دست پیدا کرد.
بنابراین اگر بخواهیم طبق تعاریف از pca یا tsne استفاده میکنیم، چه بسا که نیازی هم به استاندارد سازی نیست و حتی کلاسترینگ بدتری نیز میشود (با توجه به تراکم است).



(pca با n=3)

ویژگی های بخش اول:

ویژگی هایی که در اولین بخش مشخص میکنیم رنگها(h,s,v) وفاصله از مرکز و زاویه است (در اولین k_means)

علت تبدیل مختصات به فاصله از مرکز و زاویه (مختصات قطبی) این بوده که : مختصات x,y براساس فاصله از گوشه عکس هستند وما نیاز داشتیم که نقطه ها را براساس فاصله از مرکز محاسبه کنیم و علت تبدیل r,g,b به h,s,v هم این است که فضای رنگی اولی گسسته است ولی ما نیاز به فضای رنگی پیوسته داریم(طیف رنگی در این سری داده ها مهم است)

بخش دوم:

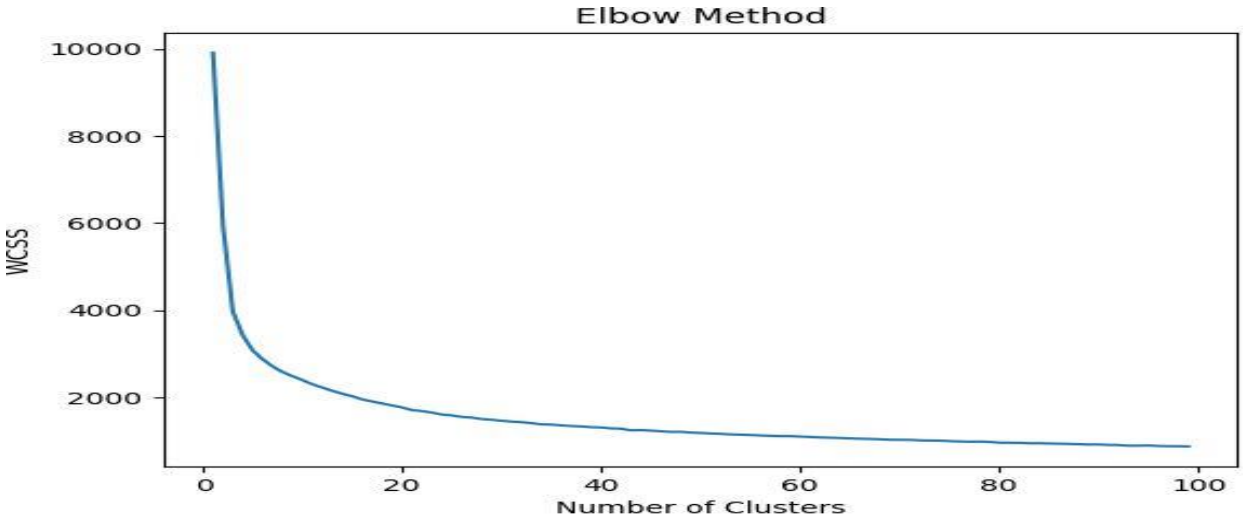
علت انتخاب الگوریتم k-means برای بخش دوم:

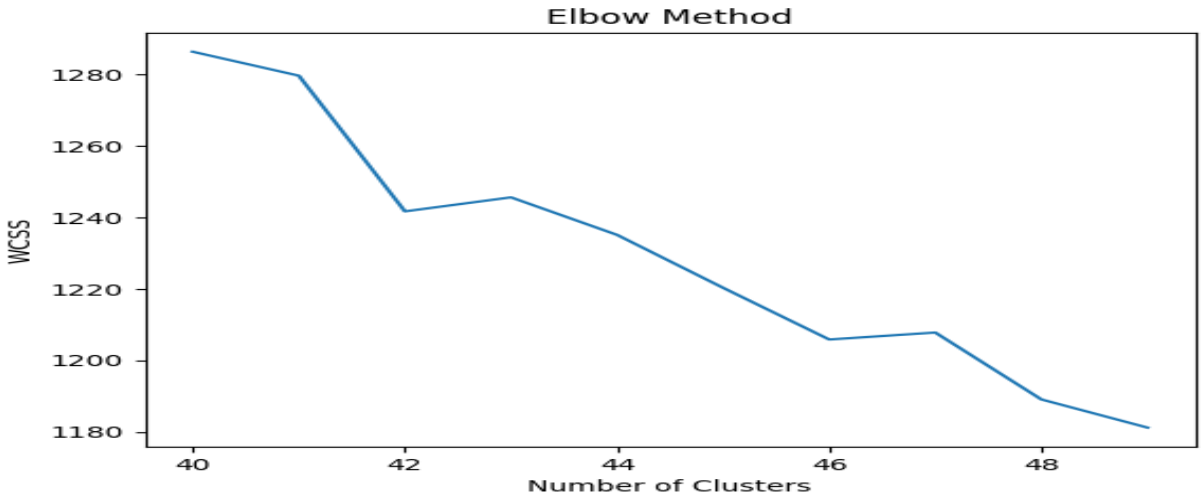
Dbscan: با توجه به ساختار داده ها این الگوریتم توانایی جداسازی درست داده ها را ندارد .

mean-shift: در این الگوریتم تمرکز روی مراکز چگالی است درصورتیکه برای دسته بندی گلها با توجه به عکسها، پراکندگی انها نسبت به هم زیاد است(ما اینکار را با واریانس هندل میکنیم)

k-median: به علت پراکندگی گلها در عکس اینکه میانه را درنظر بگیریم باعث میشود که داده های مشابه در یک دسته قرار نگیرند.

بنابراین از **k-means** استفاده میکنیم که براساس نزدیکی داده ها دسته بندی میکند(k برابر 40)





(پیدا کردن k مناسب برای بخش دوم)

تحلیل ویژگیهای بررسی شده :

1)واریانس h,s,v: نیاز به محاسبه پراکندگی رنگ برای تشخیص نواحی مشابه

ولی ویژگی مناسبی نیست زیرا ما براساس رنگ گل هم دسته بندی میکنیم و همچنین طیف و شدت ان پس توجه به پراکندگی تنها باعث بدتر دسته بندی شدن گلها میشود.

2)میانگین h و مینیمم و ماکسیمم طیف h: اصلی ترین شاخصه جداسازی گلها رنگ انهاست و معیار h رنگ را نشان میدهد.

تاثیر این 3 بر دسته بندی درست گلها : بیشترین تاثیر را hو سپس s و در اخرv است .

علت استفاده نکردن از مینیمم و ماکسیمم این است که گاهی نویزها کار را خراب میکنند.

3) میانگین فاصله از مرکز و زاویه قرارگیری در عکس: نیاز به موقعیت مکانی داریم برای تشخص گلها از هم ،همچنین به دلیل اینکه عموما گلهای داخل یک دسته از لحاظ سایز مشابه هستند بهترین حالت درصورتی به وجود میاد که ضریب یکسان با میانگین hداشته باشد(زیرا با یک نسبت باعث جداسازی و لیبل گذاری داده ها میشوند).

4)شباهت به دایره : هدف اصلی ما دسته بندی گلهاست بنابراین اهمیت شکل گل ها نیز برای ما مهم تر است که این معیار ،معیاردرست تری نسبت به مقایسه محیط و مساحت است ویا نسبت به میانگین زاویه قرارگیری گلها.

5) aspect-ratio : کاری که انجام میدهیم نسبت طول به عرض بود که با توجه به پراکندگی قرارگیری گلها در عکس ،تقریبا مختصات خود عکس را برمیگرداند بنابراین ویژگی درستی نیست (در اخر باعث بسیار بدتر کردن دقت و دسته بندی میشود زیرا تقربا برای بیشتر داده ها یکی برمیگرداند).

6)مساحت : ویژگی درستی نیست زیرا ممکن است در چند عکس یک گل باشد ولی با تعداد و یا سایز متفاوت که ممکن است در عکس های مختلف،متفاوت باشد بنابراین معیار درستی نیست .

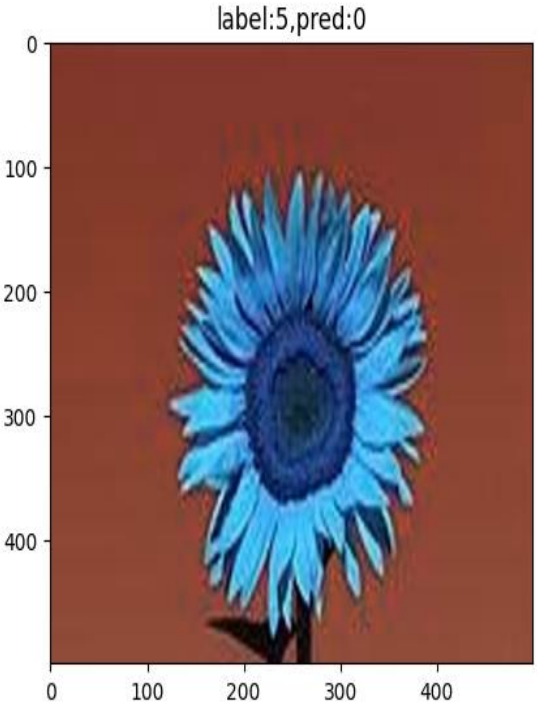
7)مرکز کلاستر: ویژگی درستی نیست زیرا احتمال دارد دو کلاستر با پراکندگی نقاط زیاد و کم مرکز یکسانی داشته باشند در حالی که یک دسته نیستند.(به علت پراکندگی زیاد به احتمال بیشتری گلهای یکسان را در دسته های متفاوتی قرار میدهد)

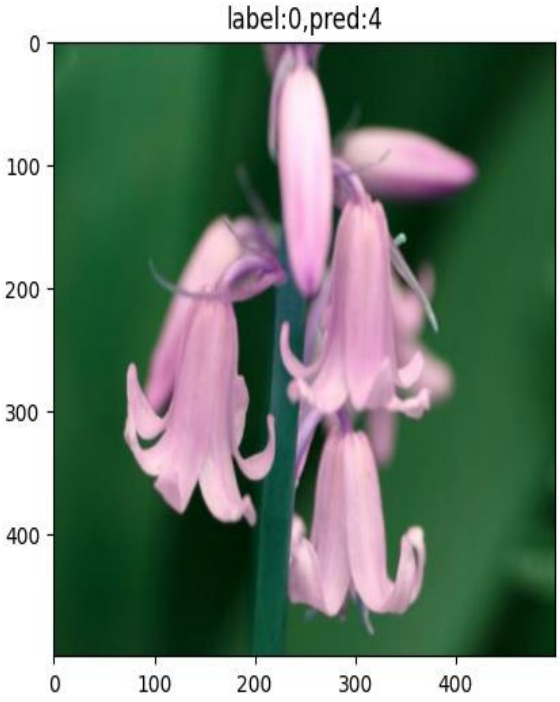
***بهترین دقت با ضریب :** h=1 , s=0.8, v=0.4, distance=1, angle=0.2, circlelarity=0.8

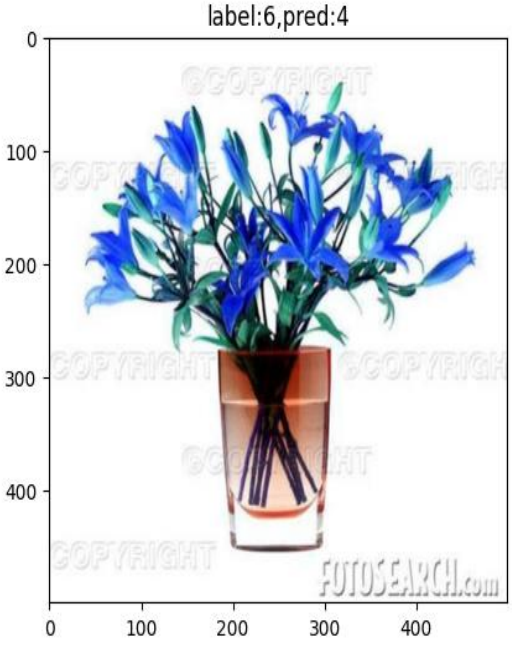
***بدترین دقتها با ضریب :** h=1 , s=1, v=0.4, distance=0.6, angle=1, circlelarity=5

***بدترین دقتها با ضریب :** h=1 , s=1, v=5, distance=4, angle=1, circlelarity=5

چند نمونه از دسته بندی اشتباه:

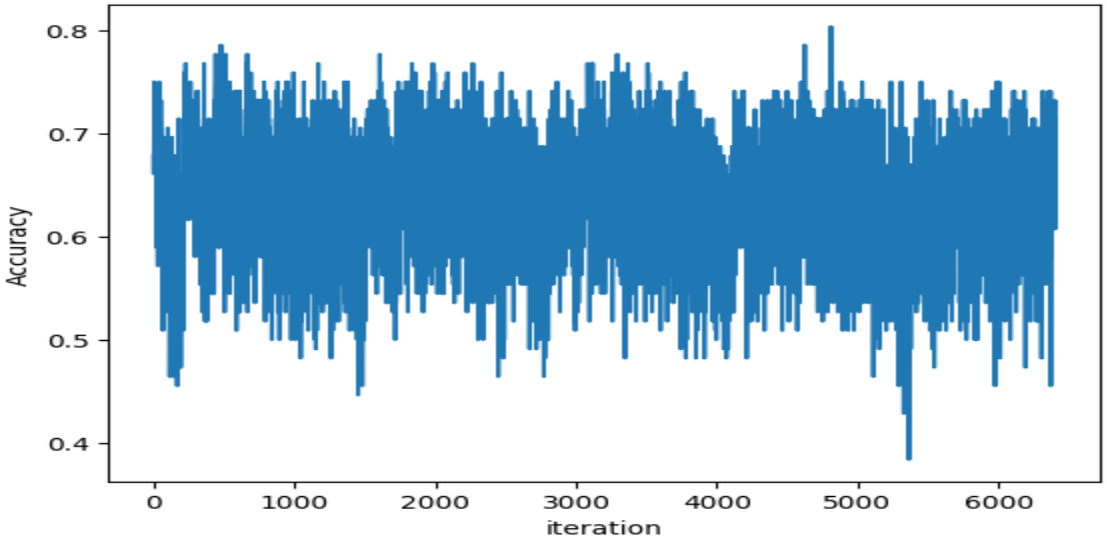






همانطور که مشاهده میشود برای نمونه ای که گلدان دارد باعث شده درست کلاستر نشود و یا به علت رنگ متفاوت گل افتابرگردون باعث شده که درست لیبل گذاری نشود.

پلات برای دقت محاسبه شده برای ضریبهای متفاوت:



:Correlation

نشان دهنده ارتباط ویژگیها نسبت به هم است،اگر 1 باشد یعنی ارتباط مستقیم دارد و اگر -1 باشد یعنی ارتباط معکوس دارد(وچون مشابه هم رفتار میکنند بنابراین یکی از انها را در نظر میگیریم)

(بنابراین با توجه به اعداد ، نتیجه این است که ویژگیها به درستی انتخاب شدند (بهترین حالت)و نیازی به حذف نداریم).

ماتریکس گمراهی:

Accuracy: نسبت تعداد مشاهداتی است که به درستی تشخیص داده شده‌اند به کل تعداد مشاهدات است.

Precision: نسبت تعداد مشاهداتی است که به درستی تشخیص داده شده‌اند به کل تعداد مشاهداتی که به عنوان مثبت پیش‌بینی شده‌اند است .

Recall: نسبت تعداد مشاهداتی است که به درستی تشخیص داده شده‌اند به کل تعداد مشاهداتی که در واقع باید به عنوان مثبت شناخته می‌شدند .

F1-SCORE: یک معیار ترکیبی است که همواره بین بازخوانی و صحت متوازن می‌کند.(نشان دهنده عملکرد بی‌نقص است).