

برای الگوریتم های مورد نظر من از کتابخانه *sklearn* استفاده کردم. Representation ای از داده ها که به عنوان ورودی به الگوریتم ها داده میشد همانطور که در صورت تمرین آمده بود صرفا پیکسل های آنها بود به طوری که هر عکس 70x80 به یک بردار (1x5600) reshape میشد و از آنجایی که کلا 410 عکس در این دیتاست موجود بود، کل دیتا به ابعاد (410x5600) درآمد که به شبکه ها داده شد. من به ازای 3 حالت مختلف داده ها را به الگوریتم ها دادم که نتایج هریک در ادامه گزارش آمده است. این حالا عبارتند از:

1. None: در این حالت هیچ نرمال سازی ای بر روی داده ها انجام نمیشود
2. StandardScaler: در این حالت داده ها به فرمی نرمال سازی میشوند که بعد از تبدیل دارای میانگین 0 و واریانس 1 باشند
3. Range [0,1]: با یک تقسیم ساده پیکسل ها بر عدد 255.0، بازه پیکسل ها بین عدد 0 و 1 قرار میگیرد. لازم به ذکر است که با بررسی نتایج مشخص شد که به ازای این representation ساده از داده ها، این حالت از نرمال سازی تقریباً تاثیری بر نتیجه نهایی نخواهد گذاشت.

Table 1: Comparison of results

ID	Algorithm	Normalization	Parameters	Rand index
1.1	KMeans	-	clusters=41	<b>0.9814</b>
1.2	KMeans	StandardScaler	clusters=41	0.9784
1.3	KMeans	Range [0,1]	clusters=41	<b>0.9814</b>
2.1	DBSCAN	-	$\epsilon=2000$ , MinPts=2	0.9237
2.2	DBSCAN	StandardScaler	$\epsilon=50$ , MinPts=2	0.9294
2.3	DBSCAN	Range [0,1]	$\epsilon=8$ , MinPts=2	0.9286
3.1.1	Agglomerative <sub>complete</sub>	-	clusters=41	0.9732
3.1.2	Agglomerative <sub>complete</sub>	StandardScaler	clusters=41	0.9683
3.1.3	Agglomerative <sub>complete</sub>	Range [0,1]	clusters=41	0.9732
3.2.1	Agglomerative <sub>single</sub>	-	clusters=41	0.6329
3.2.2	Agglomerative <sub>single</sub>	-	clusters=41	0.5367
3.2.3	Agglomerative <sub>single</sub>	Range [0,1]	clusters=41	0.6329
3.3.1	Agglomerative <sub>group</sub>	-	clusters=41	0.9560
3.3.2	Agglomerative <sub>group</sub>	StandardScaler	clusters=41	0.9474
3.3.3	Agglomerative <sub>group</sub>	Range [0,1]	clusters=41	0.9560

من الگوریتم DBSCAN را برای بهبود انتخاب کردم که توضیحات مربوطه در خود نوت بوک آمده است.

Table 2: Comparison of DBSCAN results with the proposed approach

ID	Algorithm	Normalization	Rand index
2.1	DBSCAN	-	0.9237
2.2	DBSCAN	StandardScaler	0.9294
2.3	DBSCAN	Range [0,1]	0.9286
4	DBSCAN <sub>enhanced</sub>	StandardScaler	<b>0.9430</b>

برای پیاده سازی متریک rand index یک دیکشنری در نظر گرفتم که کلید های آن همان TP, FP (برای حالاتی که pair مورد نظر دارای ground truth یکسانی بوده اند) FP, FN (برای حالاتی که pair مورد نظر دارای ground truth متفاوتی از هم بوده اند) و بنابر ground truth و همچنین کلاست رهایی که به هر سمپل اختصاص شده بود، مقادیر این دیکشنری رو بروز و پر کردم و در نهایت مقدار rand index را طبق رابطه زیر محاسبه کردم:

$$RI = \frac{TP + TN}{TP + TN + FP + FN}$$