دانشگاه صنعتی امیرکبیر

(پلی تکنیک تهران)

دانشکده مدیریت علم و فناوری

گزارش کار هفته هشتم

خوشه بندی

نگارش

رضا اکبری مقدم

استاد

دکتر مهدی قطعی

دی ماه ۹۹

فهرست مطالب

فهرست اشکال

مقدمه

خوشه‌بندی، فرآیندی است که به کمک آن می‌توان مجموعه‌ای از اشیاء را به گروه‌های مجزا افراز کرد. هر افراز یک خوشه نامیده می‌شود. اعضاء هر خوشه با توجه به ویژگی‌هایی که دارند به یکدیگر بسیار شبیه هستند و در عوض میزان شباهت بین خوشه‌ها کمترین مقدار است. در چنین حالتی هدف از خوشه‌بندی، نسبت دادن برچسب‌هایی به اشیاء است که نشان دهنده عضویت هر شیء به خوشه است.

به این ترتیب تفاوت اصلی که بین تحلیل خوشه‌بندی و تحلیل طبقه‌بندی (Classification Analysis) وجود دارد، نداشتن برچسب‌های اولیه برای مشاهدات است. در نتیجه براساس ویژگی‌های مشترک و روش‌های اندازه‌گیری فاصله یا شباهت بین اشیاء، باید برچسب‌هایی بطور خودکار نسبت داده شوند. در حالیکه در طبقه‌بندی برچسب‌های اولیه موجود است و باید با استفاده از الگوی‌های پیش‌بینی قادر به برچسب گذاری برای مشاهدات جدید باشیم.

معرفی مجموعه داده

مجموعه داده Mturk User-Perceived Clusters over Images که تحلیلی از مجموعه تصاویر ۳۲۵ کاربر میباشد. دارای ۱۸۰ نمونه و ۵۰۰ ویژگی میباشد.

	0	1	2	3	4	5	6	7	8	9	...	490	491	492	493	494	495	496	497	498	499
0	2	0	0	0	0	1	1	5	0	1	...	0	0	0	0	1	0	0	1	1	0
1	1	1	1	0	1	0	1	0	0	4	...	1	1	0	0	0	1	0	0	0	0
2	0	0	1	4	0	1	0	4	3	1	...	1	1	0	1	0	0	0	0	0	0
3	0	0	0	0	0	0	1	0	0	1	...	0	0	2	0	0	4	0	0	0	1
4	1	1	0	0	1	0	1	0	0	2	...	2	1	0	0	3	0	1	1	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
175	0	1	0	1	1	0	0	0	0	1	...	2	0	0	3	5	2	1	1	0	0
176	2	0	0	1	1	4	2	2	1	2	...	2	1	1	0	1	0	0	1	0	0
177	0	1	2	0	0	0	0	1	0	2	...	2	1	0	0	0	1	1	1	1	1
178	1	0	2	1	1	2	1	4	0	0	...	0	0	0	0	4	2	0	0	1	2
179	1	2	1	1	0	0	0	0	1	1	...	0	3	3	0	0	0	0	1	1	0

180 rows × 500 columns

شکل ۱

گزارش مراحل کار

پیش پردازش

در مرحله پیش پردازش با استفاده از فاصله های اقلیدسی و کسینوسی فاصله بین هر نمونه را حساب و ماتریسی از فاصله بین نمونه ها را بدست آوردیم.

	0	1	2	3	4	5	6	7	8	9	...	170	171	172	173	174	175	176	177	178	179
0	1.000000	0.237885	0.208438	0.201264	0.281089	0.365448	0.335341	0.313754	0.212186	0.301170	...	0.217798	0.176942	0.197043	0.602527	0.332556	0.506395	0.354644	0.228271	0.413220	0.097628
1	0.237885	1.000000	0.273183	0.193587	0.480202	0.252448	0.326174	0.253749	0.238817	0.453497	...	0.350192	0.158480	0.427339	0.107060	0.128029	0.176891	0.398888	0.324681	0.350918	0.081289
2	0.208438	0.273183	1.000000	0.287826	0.387185	0.312553	0.333587	0.422768	0.279988	0.345265	...	0.209153	0.196041	0.450284	0.122518	0.119667	0.160568	0.451560	0.351559	0.295955	0.133620
3	0.201264	0.193587	0.287826	1.000000	0.306142	0.222414	0.275459	0.411003	0.259746	0.361447	...	0.361786	0.096971	0.230980	0.153993	0.122770	0.132927	0.265205	0.289868	0.261093	0.059755
4	0.281089	0.480202	0.387185	0.306142	1.000000	0.236305	0.367224	0.246005	0.238998	0.444225	...	0.243926	0.190607	0.372054	0.146137	0.211030	0.145785	0.344603	0.404834	0.311383	0.142364
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
175	0.506395	0.176891	0.160568	0.132927	0.145785	0.283780	0.206668	0.277951	0.170576	0.144268	...	0.177463	0.100239	0.121308	0.789496	0.279518	1.000000	0.209301	0.135467	0.221200	0.043514
176	0.354644	0.398888	0.451560	0.265205	0.344603	0.409622	0.332565	0.510672	0.356750	0.397409	...	0.210373	0.178586	0.326726	0.200607	0.282191	0.209301	1.000000	0.381335	0.398204	0.142485
177	0.228271	0.324681	0.351559	0.289868	0.404834	0.266663	0.377237	0.342385	0.291134	0.444425	...	0.231615	0.337783	0.620678	0.150313	0.144187	0.135467	0.381335	1.000000	0.320698	0.306974
178	0.413220	0.350918	0.295955	0.261093	0.311383	0.371216	0.302448	0.297509	0.329075	0.300460	...	0.428285	0.134736	0.253101	0.211006	0.233218	0.221200	0.398204	0.320698	1.000000	0.140167
179	0.097628	0.081289	0.133620	0.059755	0.142364	0.076116	0.213586	0.086361	0.303871	0.308576	...	0.107716	0.720666	0.308489	0.044691	0.080166	0.043514	0.142485	0.306974	0.140167	1.000000

180 rows × 180 columns

شکل ۲

در شکل ۲ فاصله کسینوسی نمونه ها محاسبه شده است.

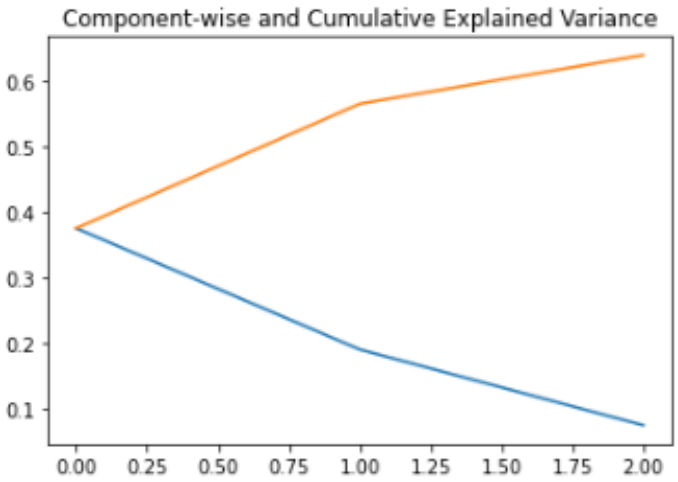
	0	1	2	3	4	5	6	7	8	9	...	170	171	172	173	174	175	176	177	178	179
0	0.000000	24.738634	30.066593	29.816103	24.899799	31.527766	25.553865	27.055499	24.799194	24.919872	...	26.000000	39.076847	32.664966	53.833075	24.535688	45.088801	27.892651	26.888659	28.896367	51.797683
1	24.738634	0.000000	24.576411	25.199206	16.186414	30.951575	20.904545	23.537205	17.691806	17.117243	...	18.220867	36.262929	25.159491	64.776539	21.771541	51.662365	23.706539	20.074860	27.549955	49.467161
2	30.066593	24.576411	0.000000	28.407745	23.452079	32.924155	25.865034	25.059928	24.228083	24.474477	...	26.457513	38.845849	27.294688	66.196677	28.319605	53.860932	25.922963	25.000000	31.764760	51.137071
3	29.816103	25.199206	28.407745	0.000000	24.351591	34.626579	26.532998	24.959968	24.000000	23.748684	...	23.515952	40.669399	31.874755	65.322278	27.802878	54.313902	29.647934	25.690465	32.218007	52.554733
4	24.899799	16.186414	23.452079	24.351591	0.000000	31.749016	21.142375	24.535688	19.104973	18.193405	...	20.832667	36.318040	26.476405	64.404969	21.908902	52.449976	25.139610	19.773720	28.653098	48.959167
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
175	45.088801	51.662365	53.860932	54.313902	52.449976	52.659282	51.961524	50.764161	51.672043	52.630789	...	51.951901	60.133186	56.017854	39.711459	50.229473	0.000000	53.037722	53.122500	53.814496	69.641941
176	27.892651	23.706539	25.922963	29.647934	25.139610	31.112698	26.739484	23.790755	24.186773	24.474477	...	27.422618	39.812058	30.805844	64.560050	26.645825	53.037722	0.000000	25.317978	29.983329	51.371198
177	26.888659	20.074860	25.000000	25.690465	19.773720	31.984371	22.135944	24.062419	20.199010	19.493589	...	22.561028	34.234486	21.771541	64.645185	24.392622	53.122500	25.317978	0.000000	29.223278	46.303348
178	28.896367	27.549955	31.764760	32.218007	28.653098	33.808283	30.000000	30.967725	27.748874	29.086079	...	26.627054	42.755117	34.641016	65.061509	30.380915	53.814496	29.983329	29.223278	0.000000	52.858301
179	51.797683	49.467161	51.137071	52.554733	48.959167	55.578773	48.373546	51.604263	46.151923	46.151923	...	49.588305	33.585711	47.791213	79.050617	50.309045	69.641941	51.371198	46.303348	52.858301	0.000000

180 rows × 180 columns

شکل ۳

در شکل ۳ فاصله اقلیدسی نمونه ها محاسبه شده است.

همچنین برای مقایسه بیشتر مجموعه داده از روش pca برای کاهش ابعاد استفاده شده است که با استفاده از معیارهای Component-wise و Cumulative Explained Variance که در شکل ۴ نمایش داده شده است به این نتیجه میرسیم که ۶۰ درصد داده ها در Cumulative Explained Variance در ۳ بعد قرار دارند.



شکل ۴

در شکل ۵ جدول کاهش ابعاد نمایش داده شده است.

	x	y	w
0	0.459239	1.206536	-0.181593
1	-0.199418	0.159961	0.342806
2	-0.372305	0.142478	0.050542
3	0.982712	0.577481	0.159428
4	-0.542169	-0.113620	-0.061346
...	...	...	...
175	1.906089	1.594927	0.360806
176	-0.022504	0.382784	-0.306103
177	-0.447423	-0.934597	0.254611
178	0.583521	0.662826	-0.060425
179	2.944003	-1.458807	-0.266318

180 rows × 3 columns

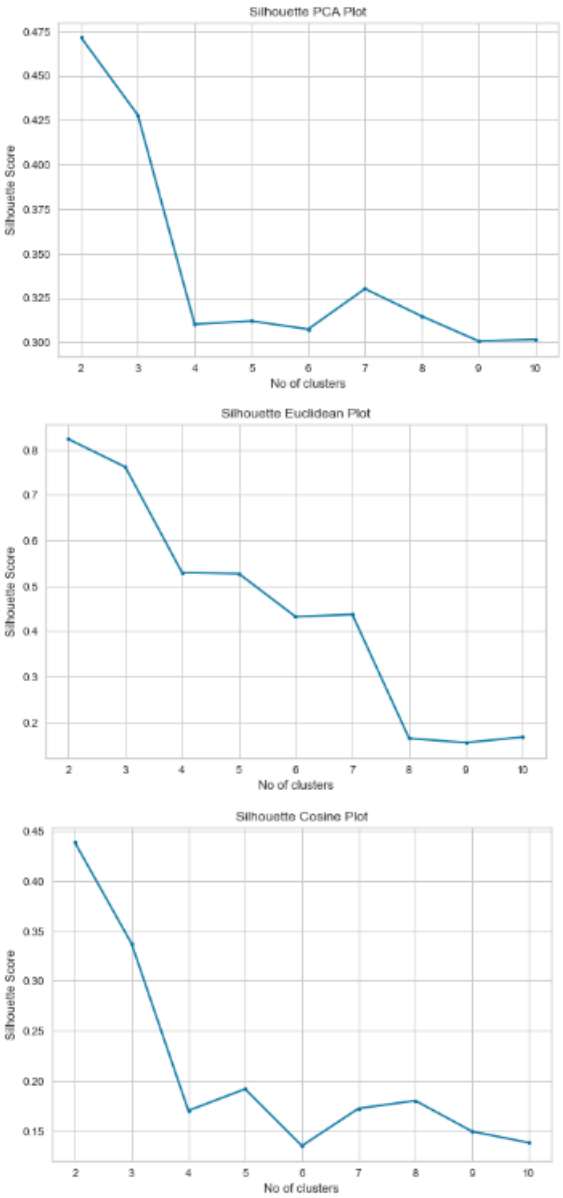
شکل ۵

خوشه بندی

Partitioning

در این بخش از الگوریتم kmeans برای بدست آوردن خوشه های مجموعه داده استفاده شده است. برای این الگوریتم از هر سه روش پیش پردازش انجام شده بر روی مجموعه داده استفاده شده است تا نتایج هر کدام بررسی شود.

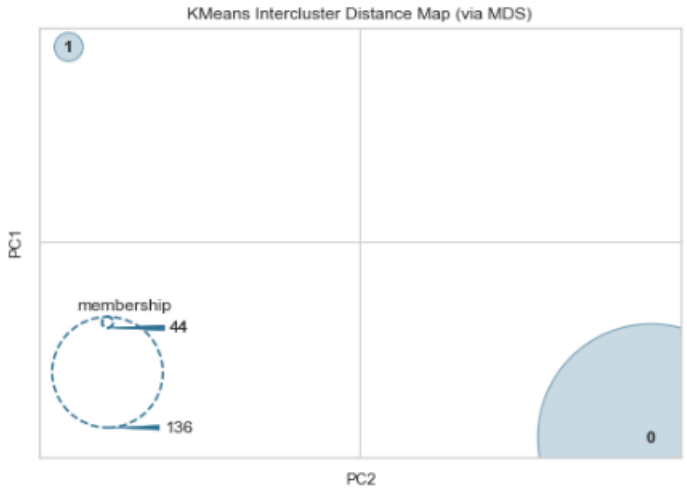
در شکل ۶ مشاهده میشود که برای داده های PCA بهترین خوشه بندی ۲ و برای داده های اقلیدسی بهترین خوشه بندی ۴ و داده های کسینوسی بهترین خوشه بندی ۵ میباشد.



شکل ۶

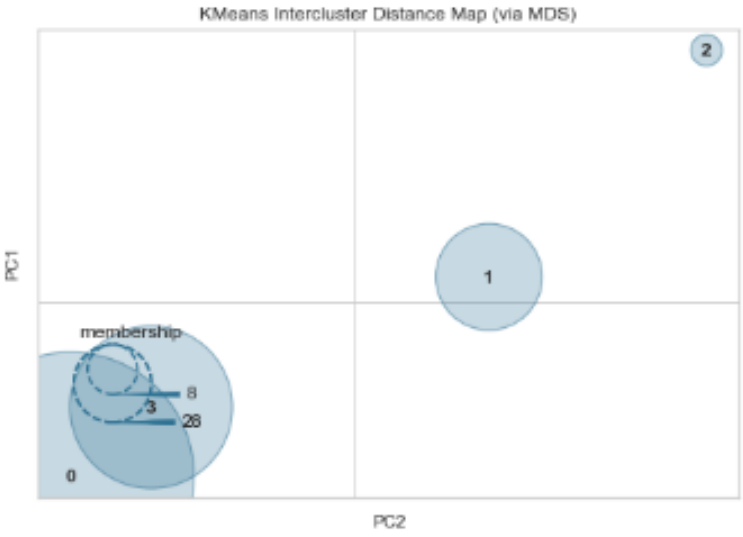
که پس از خوشه بندی نتایج بدست آمده براساس خوشه بندی و همچنین فاصله داخلی و خارجی خوشه ها به شکل زیر میباشد.

شکل ۷ نمایش داده های مجموعه داده pca میباشد.



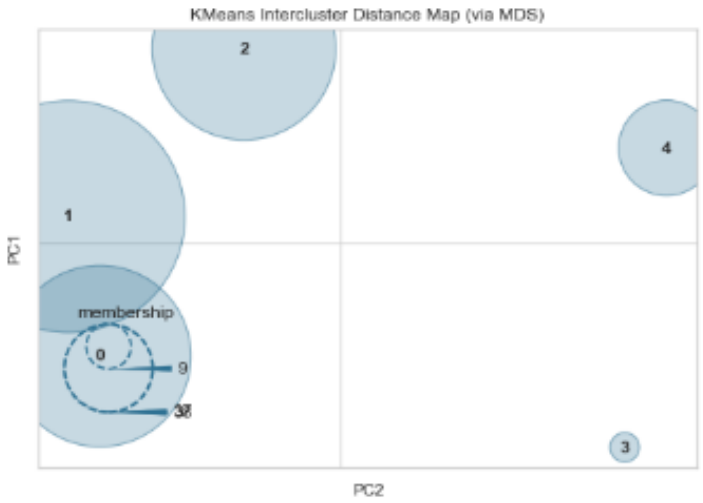
شکل ۷

شکل ۸ نمایش مجموعه داده اقلیدسی میباشد.



شکل ۸

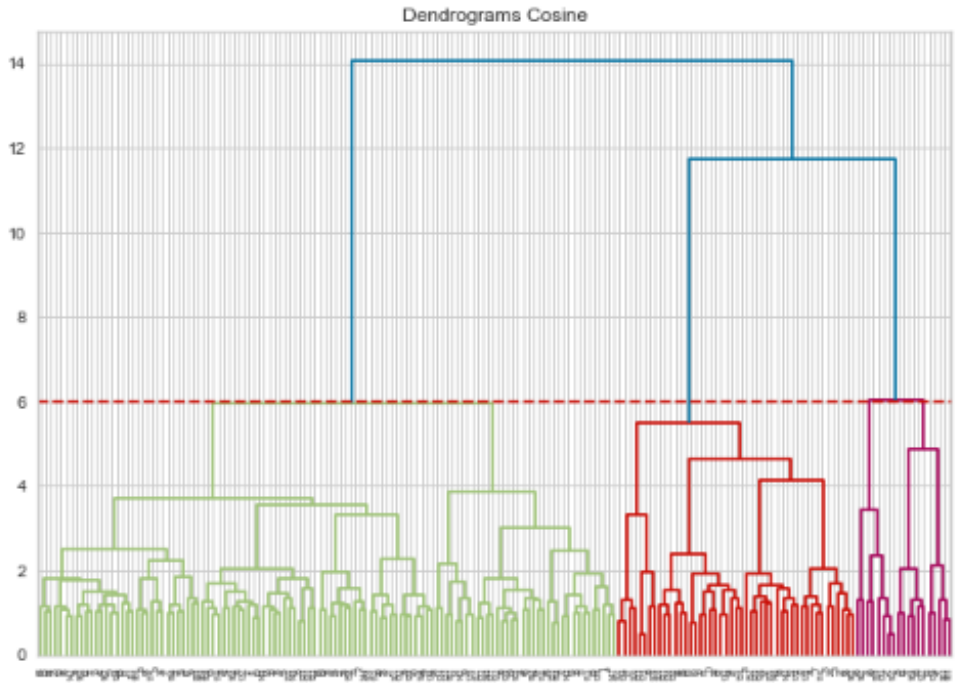
شکل ۹ نمایش مجموعه داده کسینوسی میباشد.



شکل ۹

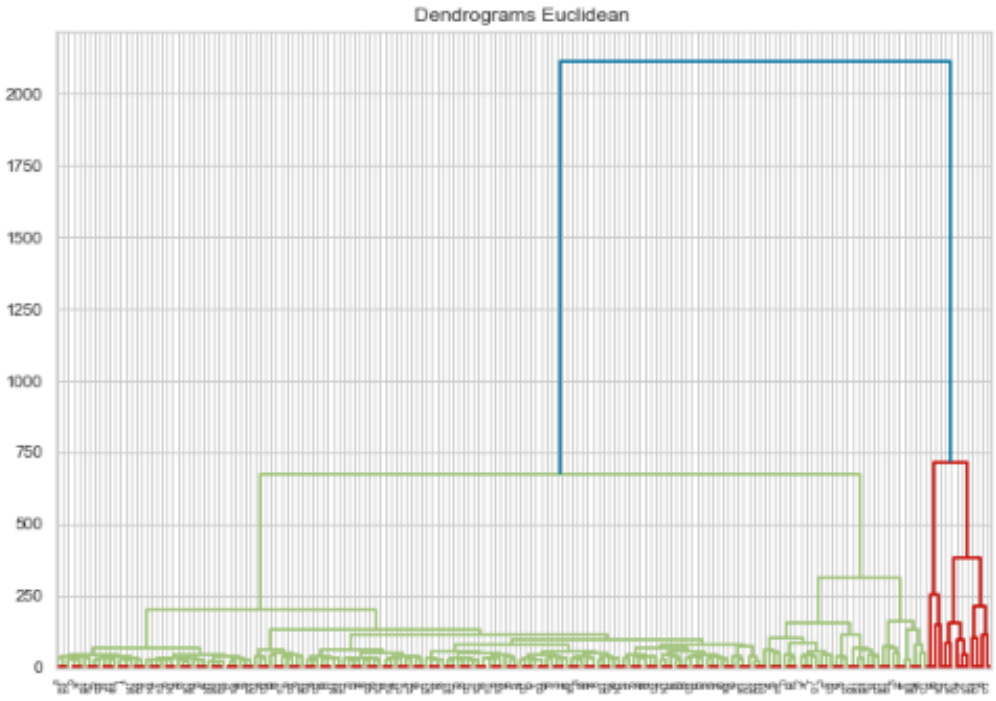
از نتایج خوشه های این مجموعه داده ها میتوان نتیجه گرفت که مجموعه داده کسینوسی نتیجه بهتری در خوشه بندی با استفاده از kmeans را دارد.

سلسله مراتبی



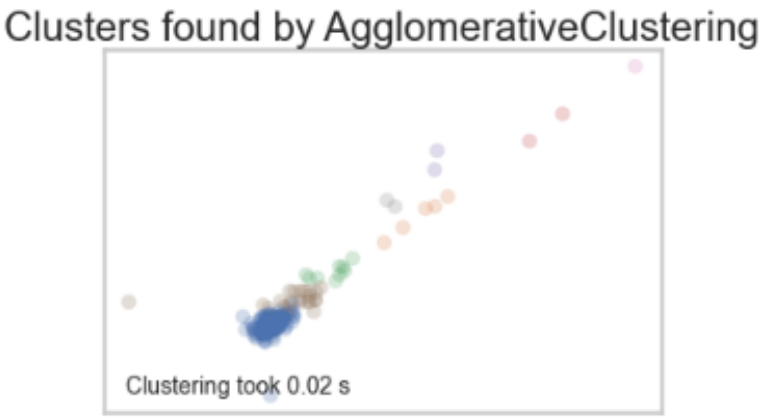
شکل ۱۰

در این روش ابتدا نمودار دندوگرام مجموعه داده کسینوسی را محاسبه کرده ایم که همانطور که مشاهده میشود تعداد ۴ خوشه بهترین خوشه بندی میباشد.



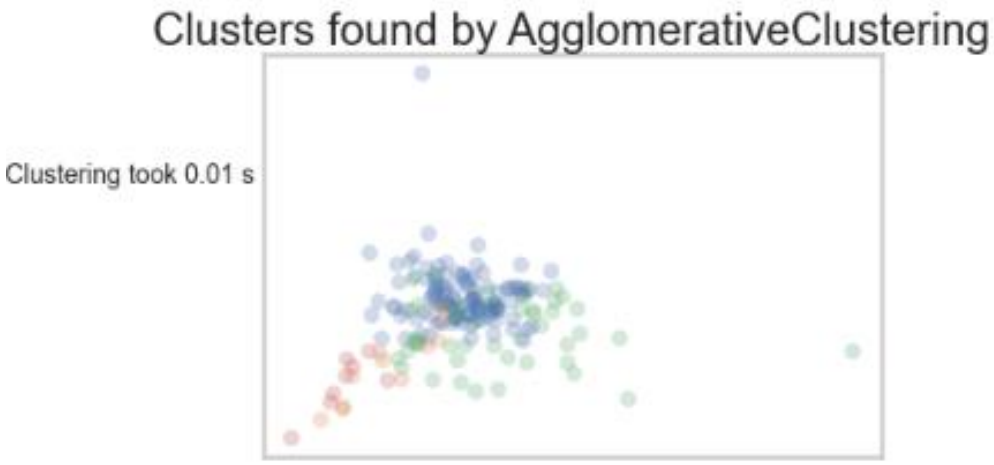
شکل ۱۱

در شکل ۱۱ دندوگرام مجموعه داده اقلیدسی نمایش داده شده است که ۸ خوشه مناسب میباشد.



شکل ۱۲

در شکل ۱۲ نمایش خوشه بندی سلسله مراتبی مجموعه داده اقلیدسی بر اساس ۸ خوشه میباشد.

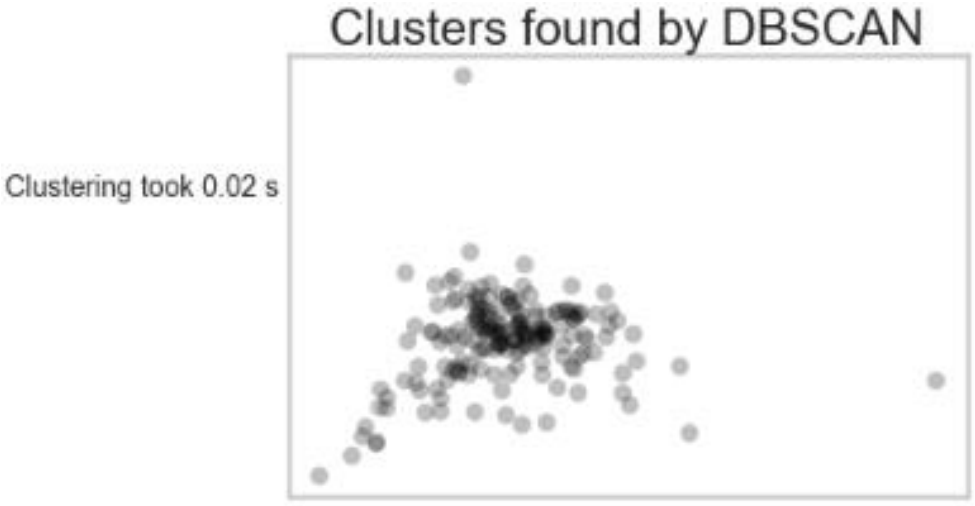


شکل ۱۳

در شکل ۱۳ نمایش خوشه بندی سلسله مراتبی مجموعه داده کسینوسی بر اساس ۴خوشه میباشد.

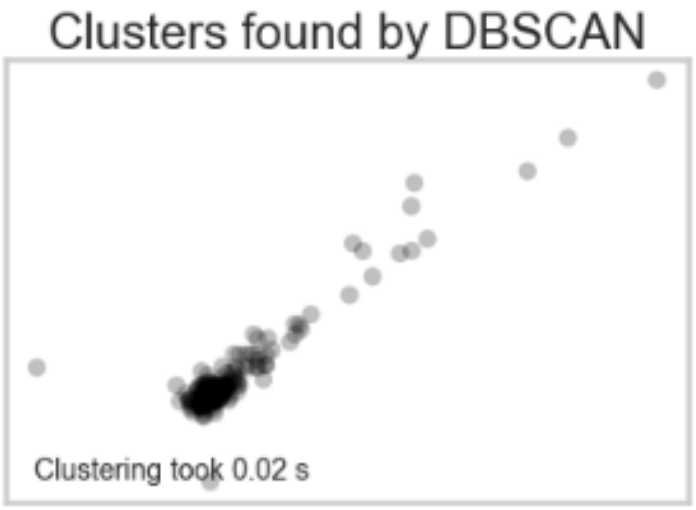
DBScan

در این خوشه بندی با eps ۰,۰۲۵ خوشه بندی را انجام داده ایم.



شکل ۱۴

شکل ۱۴ نمایش اسکترپلات خوشه بندی مجموعه داده کسینوسی میباشد.



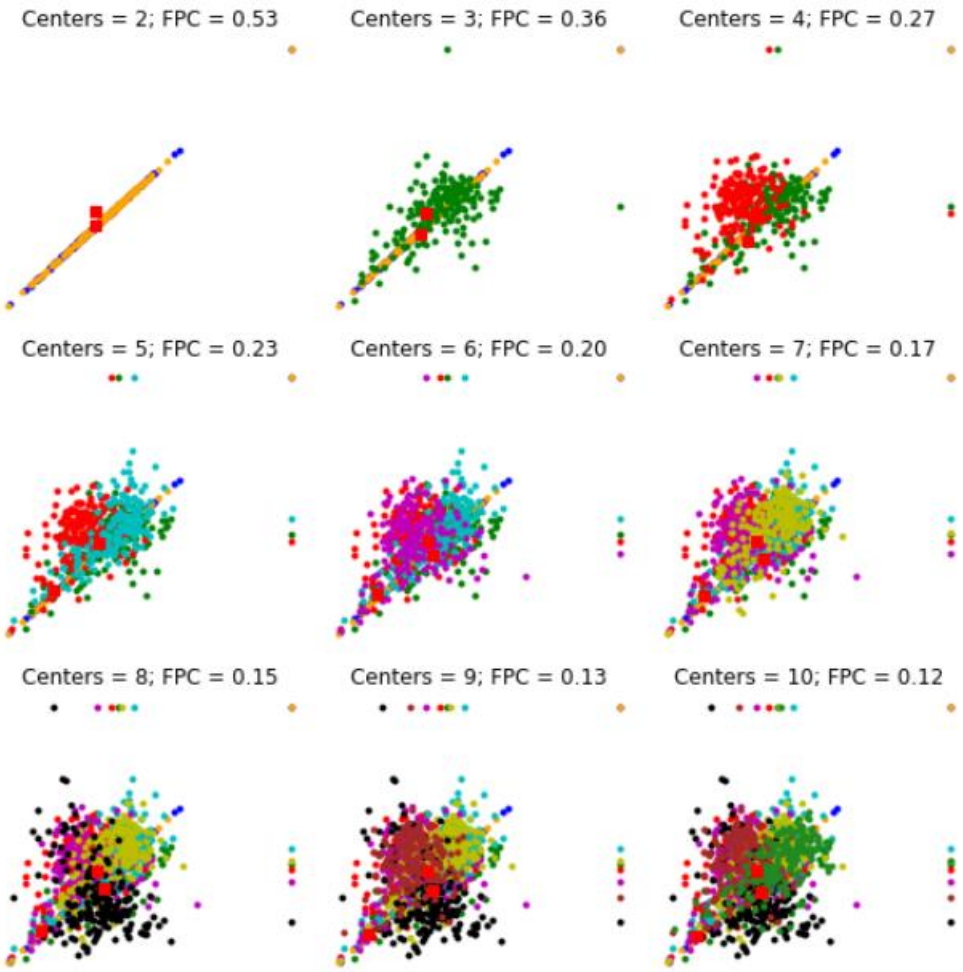
شکل ۱۵

شکل ۱۵ نمایش اسکترپلات خوشه بندی مجموعه داده اقلیدسی میباشد.

در خوشه بندی به روش DBScan نتیجه مطلوبی از جهت خوشه بندی حاصل نشده است و تمامی داده ها در یک خوشه قرار گرفته اند.

خوشه بندی فازی

C-Means



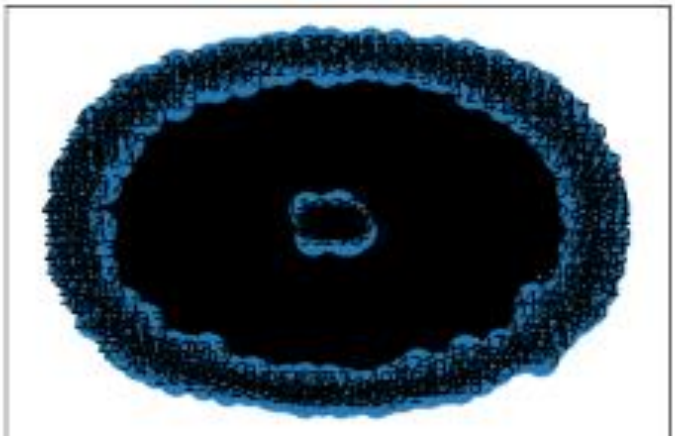
شکل ۱۶

در شکل ۱۶ مجموعه ای از الگوریتم های اجرا شده بر روی مجموعه داده کسینوسی میباشد که نحوه خوشه بندی را مشاهده میفرمایید.

در مدل های احتمالاتی هر نمونه فقط عضو یک خوشه ممکن است باشد اما در خوشه بندی فازی ممکن است یک نمونه عضو هر خوشه ای شود و بر اساس آن درصدی که میگیرد متفاوت است. اما در این مجموعه داده همانطور که مشاهده شد خوشه بندی فازی بهتر از خوشه بندی احتمالاتی که در قسمت اول ذکر شد عمل کرده است و داده ها عملکرد بهتری نشان داده اند. لذا از نظر بنده خوشه بندی فازی مناسبتر میباشد چرا که باعث میشود نمونه ها فرصت بهتری داشته باشند که تحت شرایط گوناگون در خوشه های دیگر قرار بگیرند.

گراف

در زمینه گراف بنده ابتدا هر نمونه را به صورت یک گره در نظر گرفته و ارتباطی که با هر ویژگی دارد را به صورت یال در نظر گرفتم و نمودار گراف آن را رسم کردم. اما نمودار بسیار شلوغی بدست آمد و همچنین با اعمال این گراف بر روی الگوریتم های خوشه بندی نتیجه مطلوبی دریافت نشد. شکل ۱۷ نمایش خروجی گراف مجموعه داده میباشد.



شکل ۱۷

نتیجه گیری

بر اساس مشاهدات انجام شده و همچنین آزمون و خطا روش های خوشه بندی بایستی بر روی فاصله داده ها اعمال شود تا نتیجه درستی داشته باشد همچنین روش های خوشه بندی برای هر مجموعه داده نتیجه متفاوتی دارد و نمیشود انتظار خروجی مطلوبی از هر الگوریتم خوشه بندی داشت. در خصوص روش های فازی به نظر بنده نسبت به روش های احتمالاتی در مجموعه داده هایی که درصد خطای متغیر دارند بهتر از روش های احتمالاتی عمل میکنند.

منابع

۱. https://archive.ics.uci.edu/ml/datasets/NIPS+Conference+Papers+۱۹۸۷-#۲۰۱۵

۲. https://medium.com/pursuitnotes/k-means-clustering-model-in-۶-steps-ad۸cfa۵۳۲b۳۵with-python-

۳. https://towardsdatascience.com/pca-using-python-scikit-learn-۶۰e۸۹۸۹f۶۵۳e