خوشه‌بندهای K-means و SOM

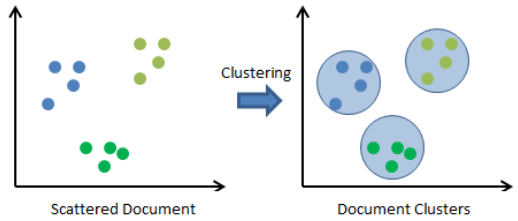
مجید نصیری منجیلی
دانشکده مهندسی کامپیوتر، دانشگاه تربیت دبیر شهید رجایی
majid.nasiri@srttu.edu

چکیده

در این گزارش با استفاده از الگوریتم‌های خوشه-بندی[1] K-means و SOM[2] به خوشه‌بندی دیتاست‌های دو کلاسه با توزیع گاسین، iris و satimage پرداخته می‌شود. در این گزارش نتایج بدست آمده از آنها را مورد بررسی قرار می‌دهیم. این گزارش برمبنای تمرین درس یادگیری ماشین می‌باشد.

۱- مقدمه

اولین بار ایده‌ی خوشه‌بندی در دهه‌ی ۱۹۳۵ ارائه شد و امروزه با پیشرفت‌ها و جهش‌های عظیمی که در آن پدید آمده، خوشه‌بندی در کاربردها و جنبه-های مختلفی حضور یافته است. خوشه‌بندی یکی از شاخه های یادگیری بدون نظارت[3] می‌باشد و فرآیند خودکاریست که در طی آن، نمونه ها به دسته‌هایی که اعضای آن مشابه همدیگر می‌باشند تقسیم می‌شوند. که به این دسته‌ها خوشه گفته می‌شود. بعبارتی دیگر خوشه‌بندی، فرآیند دسته-بندی مجموعه‌ای از اشیاء به خوشه‌هایی است که اعضای درونی هر خوشه بیشترین شباهت را به یکدیگر و کمترین شباهت را نسبت به اعضای سایر خوشه‌ها داشته باشد. نمونه‌ای از عمل خوشه‌بندی در شکل ۱ آورده شده است.



شکل ۱: خوشه‌بندی دیتا

در ادامه پیاده‌سازی الگوریتم های K-means و SOM و سپس نتایج آنها برای خوشه‌بندی دیتاست های مختلف ارائه شده است. در بخش انتهایی هم این دو خوشه‌بند با هم مقایسه شده‌اند.

۲- پیاده‌سازی

۲-۱- K-means

الگوریتم K-means یک روش پایه برای بسیاری از روش‌های خوشه‌بندی دیگر محسوب می‌شود. این الگوریتم داری شامل مراحل زیر می‌باشد.

- بدست آوردن نقاطی به عنوان مراکز خوشه‌ها که این نقاط در واقع همان میانگین نقاط متعلق به هر خوشه هستند.
- نسبت دادن هر نمونه داده به یک خوشه که آن داده کمترین فاصله اقلیدوسی را تا مرکز آن خوشه را دارا باشد.
- قرار دادن میانگین داده‌های اختصاص داده شده به مراکز خوشه های فعلی بعنوان مراکز خوشه‌های جدید.

[1] Clustering
[2] Self-organized mapping
[3] Unsupervised

- تکرار فرآیند بروز رسانی مراکز خوشه‌ها تا جایی که دیگر تغییر چندانی در آنها ایجاد نشود.

SOM -۲-۲

در SOM از روش یادگیری رقابتی برای آموزش استفاده می‌شود و مبتنی بر مشخصه‌های خاصی از مغز انسان، توسعه یافته است. در این روش به تعداد ویژگی‌های نمونه‌ها، ورودی داریم و به تعداد خوشه‌های دیتا در لایه خروجی، نورون قرار می-دهیم. در این الگوریتم نورون‌ها توسط یک تابع همسایگی به یکدیگر متصل شده‌اند. و هر بردار ورودی بر اساس بیشترین شباهت، نورونی را در لایه خروجی را که سلول برنده خوانده می‌شود، فعال می‌کند. که در شباهت توسط فاصله اقلیدوسی (رابطی ۱) سنجیده می‌شود. همچنین از رابطی ۲ برای بروزرسانی وزن‌ها استفاده می-شود.

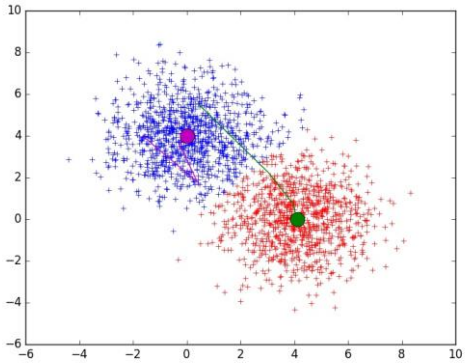
$$d = \sum_{k=1}^{n} (i_{l,k} - w_{j,k}(t))^2 \qquad (1)$$

$$w_j(t+1) = w_j(t) + \eta(t)\left(i_l - w_j(t)\right) \quad (2)$$

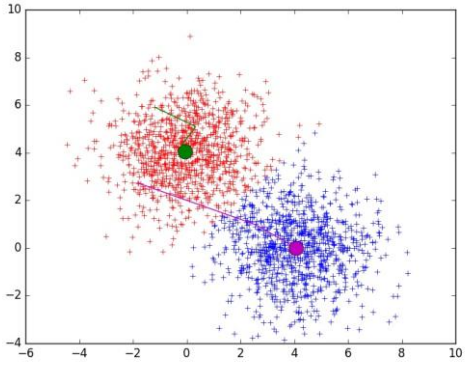
۳- نتایج

K-means -۱-۳

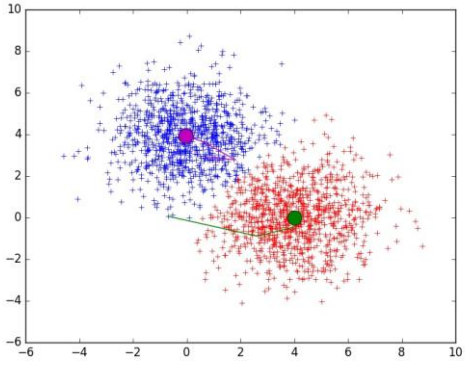
شکل‌های ۲ تا ۴ مربوط به تست‌های مختلف انجام شده بر روی دیتای دوکلاسه با توزیع گوسی می-باشد. با توجه به آنها می‌توان تغییرات مرکز خوشه-ها، و همگرایی آن‌ها را به مرکز خوشه مشاهده کرد. در این تصاویر نمونه ها با رنگهای قرمز و آبی و همچنین جهت حرکت مراکز خوشه ها و مراکز نهایی خوشه‌ها با خطوط و دایره‌هایی با رنگ‌های سبز و فیروزه‌ای نشان داده شده است.



شکل ۲: جهت حرکت مراکز و مراکز نهایی خوشه‌های بدست آمده برای دیتای دو کلاسه با توزیع گاسین (تست ۱)



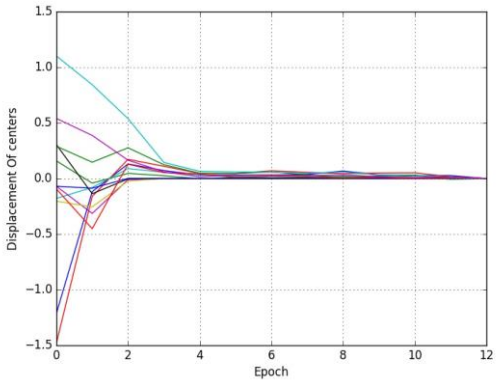
شکل ۳: جهت حرکت مراکز و مراکز نهایی خوشه‌های بدست آمده برای دیتای دو کلاسه با توزیع گاسین (تست ۲)



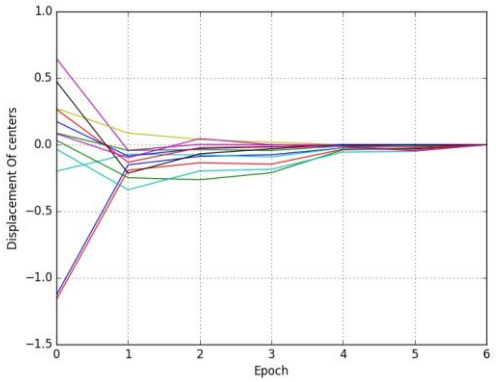
شکل ۴: جهت حرکت مراکز و مراکز نهایی خوشه‌های بدست آمده برای دیتای دو کلاسه با توزیع گاسین (تست ۳)

نتایج بدست آمده برای خوشه‌بندی دیتاست iris با الگوریتم K-means در شکل‌های ۵ تا ۷ آمده است. در این تصاویر محور افقی تعداد تکرار الگوریتم برای رسیدن به همگرایی و محور عمودی مقادیر تغییرات هر یک از مختصات می‌باشد. دیده می‌شود

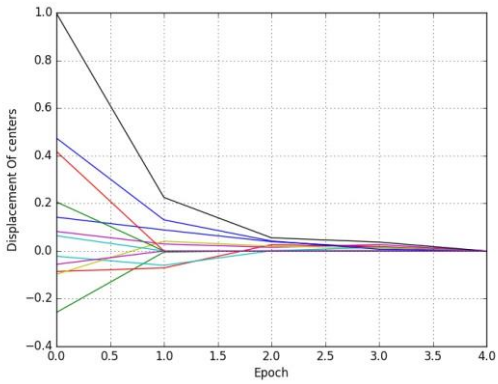
که بعد از چندین تکرار تغییرات مختصات مراکز خوشه‌ها ناچیز خواهد بود.



شکل ۵: تغییرات مراکز خوشه‌ها در تکرار الگوریتم برای دیتای iris (تست ۱)

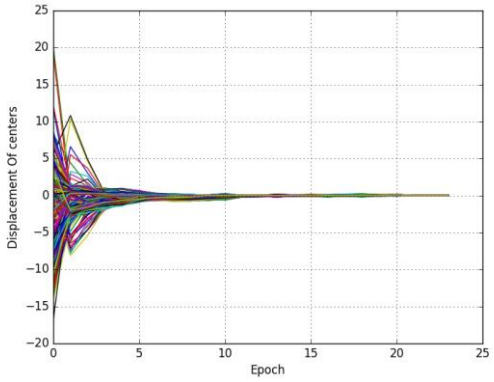


شکل ۶: تغییرات مراکز خوشه‌ها در تکرار الگوریتم برای دیتای iris (تست ۲)

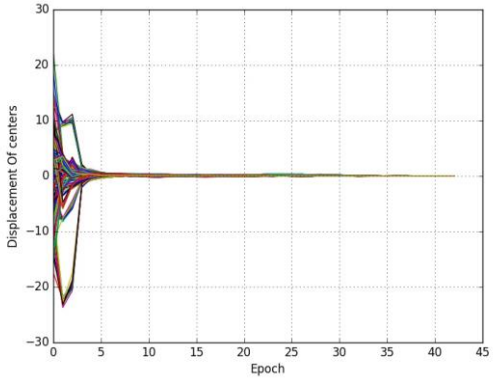


شکل ۷: تغییرات مراکز خوشه‌ها در تکرار الگوریتم برای دیتای iris (تست ۳)

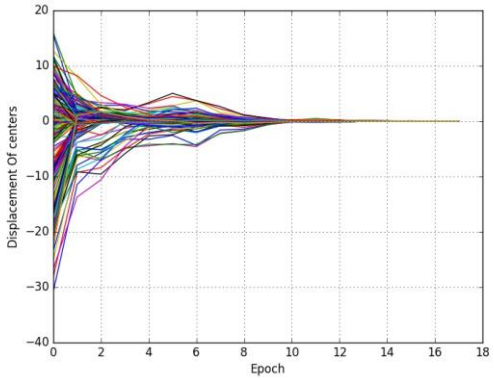
در ادامه با استفاده از الگوریتم K-means به خوشه‌بندی دیتاست satimage پرداختیم که نتایج همگرایی مراکز خوشه‌ها در شکل های ۸ تا ۱۰ آورده شده است.



شکل ۸: تغییرات مراکز خوشه‌ها در تکرار الگوریتم برای دیتای satimage (تست ۱)



شکل ۹: تغییرات مراکز خوشه‌ها در تکرار الگوریتم برای دیتای satimage (تست ۲)

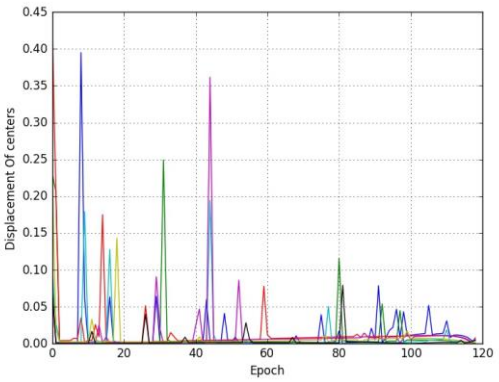


شکل ۱۰: تغییرات مراکز خوشه‌ها در تکرار الگوریتم برای دیتای satimage (تست ۳)

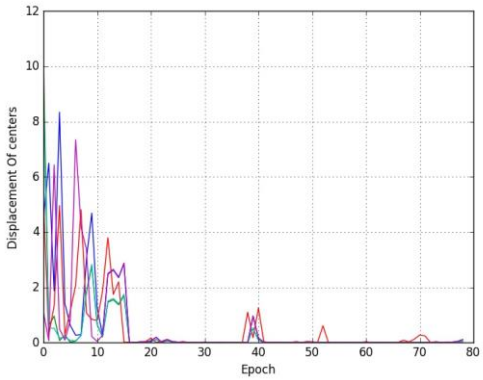
SOM-۲-۳

در ادامه کار دیتاست‌های موجود را با استفاده از الگوریتم SOM خوشه بندی کردیم که نتایج آن‌ها در شکل‌های ۱۱ تا ۱۳ آورده شده است. در این شکل‌ها محور افقی تعداد تکرارهای الگوریتم و

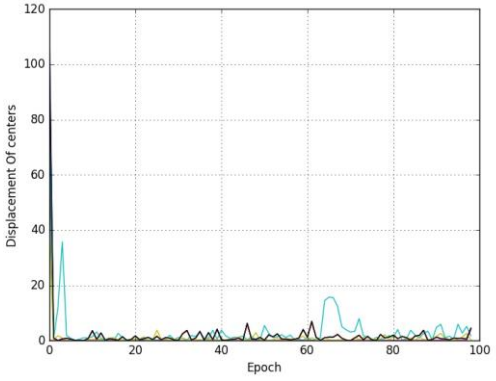
محور افقی تغییرات مختصات وزن‌های بدست آمده می‌باشد. دیده می‌شود که در شکل‌های بدست آمده این جابجایی‌ها در مختصات مراکز خوشه‌ها به صفر رسیده است.



شکل ۱۱: همگرایی وزن‌های الگوریتم SOM برای خوشه‌بندی دیتای دوکلاسه با توزیع گاسی



شکل ۱۲: همگرایی وزن‌های الگوریتم SOM برای خوشه‌بندی دیتای iris



شکل ۱۳: همگرایی وزن‌های الگوریتم SOM برای خوشه‌بندی دیتای satimage

۴- نتیجه‌گیری

با توجه به نتایج بدست آمده، مراکز بدست آمده در الگوریتم خوشه‌بندی K-means با توجه به توزیع دیتا و انتخاب اولیه نمونه‌ها تغییر می‌کند، باید تعداد خوشه‌ها (K) را مشخص کنیم و هزینه پردازشی بالایی دارد.

الگوریتم SOM دیتای با ابعاد بالا را به فضای دو بعدی نگاشت می‌دهد. این الگوریتم نیز توپولوژی دیتا را حفظ می‌کند. البته در الگوریتم SOM باید دیتای زیاد و خوب در اختیار داشته باشیم.

در صورتی که تعداد خوشه‌ها زیاد می‌شوند الگوریتم K-means نسبت به SOM پاسخ بهتری دارد. الگوریتم K-means نسبت به SOM به نویز و یا دیتای پرت حساسیت بیشتری دارد.

با توجه به تست‌های انجام شده و نتایج بدست آمده که در مقاله [1] آمده است، SOM کارایی بهتری از K-means ندارد. در این مقاله این دو روش از لحاظ تعداد خوشه‌ها، تعداد ویژگی‌ها و خطا بر روی ۱۰۸ دیتاست مختلف مورد بررسی قرار گرفته است.

مراجع

[1] Mingoti, Sueli A., and Joab O. Lima. "Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms." *European journal of operational research* 174.3 (2006): 1742-1759.