Fuzzy C-means clustering

Project report for computational intelligence course

Parsa Eskandarnejad

9531003

Computer Engineering & IT Department, Amirkabir University of Technology

چکیده- این گزارش به بررسی الگوریتم FCM می‌پردازد. الگوریتم FCM یک الگوریتم خوشه‌بندی فازی است که به جای این که تعلق داده‌ها را به خوشه‌ها صفر و یکی در نظر بگیرد، آن‌ها را به صورت فازی و بین مقادیر صفر و یک حساب می‌کند. در برنامه پیاده‌سازی شده به زبان پایتون نیز این الگوریتم پیاده شده و عملکرد آن بر روی مجموعه داده‌های مختلف آزمایش و گزارش شده است.

کلمات کلیدی: خوشه‌بندی، خوشه‌بندی فازی، یادگیری بدون نظارت، فازی‌سازی

مقدمه

در حالت کلی خوشه‌بندی[1] یک روش یادگیری ماشین بدون نظارت است. در این روش داده‌هایی که از قبل برچسبی ندارند دسته‌بندی می‌شوند. اعضای هر دسته با یک دیگر تشابه و با اعضای بقیه دسته‌ها تفاوت بیشتری خواهند داشت. از کاربردهای خوشه‌بندی می‌توان به موارد زیر اشاره کرد [1].

۱. زیست‌شناسی و بیوانفورماتیک: تشخیص و توصیف تجمع ارگانیسم‌ها در محیط‌های زیستی، آنالیز ژن‌ها
۲. بازاریابی: دسته‌بندی مشتریان و ارائه خدمات مرتبط به همان دسته، دسته‌بندی محصولات
۳. اینترنت: تشخیص جوامع مختلف در شبکه‌های اجتماعی، دسته‌بندی نتایج جستجو
۴. علوم کامپیوتر: قسمت‌بندی تصاویر دیجیتال به بخش‌های مجزا، الگوریتم‌های ژنتیک، سیستم‌های توصیه‌کننده
۵. علوم اجتماعی: تشخیص جرم، دسته‌بندی جوامع آموزشی

الگوریتم‌های خوشه‌بندی را می‌توان به دو دسته‌ی الگوریتم‌های خوشه‌بندی سخت و الگوریتم‌های خوشه‌بندی نرم تقسیم کرد. در الگوریتم‌های خوشه‌بندی سخت، هر شی یا عضو یک دسته است یا عضو آن نیست. ولی در الگوریتم‌های خوشه‌بندی نرم، عضویت یک شی در دسته‌ها به صورت فازی تعریف می‌شود. این به معنی است که یک شی می‌تواند در دسته‌های مختلف با درجات مختلف حضور داشته باشد. در حقیقت در الگوریتم‌های خوشه‌بندی سخت عضویت یک شی در یک گروه موجب آن می‌شود که تشابه آن شی با اعضای دیگر دسته‌ها بررسی نشود و این یک نقطه ضعف برای این نوع از الگوریتم‌ها است. یک روش برای در نظر گرفتن تشابه عضو یک دسته به اعضای بقیه دسته‌ها را زاده در سال ۱۹۶۵ ارائه کرد. نکته کلیدی این روش این است که میزان تشابه یک عضو با بقیه دسته‌ها را با یک تابع به نام تابع تعلق نمایش دهیم که مقادیر این تابع که به آن میزان تعلق می‌گوییم اعدادی بین ۰ و ۱ باشند. در واقع در این روش هر شی دارای یک میزان تعلق در هر دسته است. میزان تعلق هر چه به ۱ نزدیک‌تر باشد نشان‌دهنده درجه

[1] Clustering

شباهت عضو به آن دسته است و هر چه این مقدار به ۰ نزدیک باشد نشان‌دهنده عدم شباهت است [2].

الگوریتم

الگوریتم k-means یک الگوریتم خوشه‌بندی سخت است که سعی می‌کند تابع هدف زیر را بهینه کند:

$$J = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik} D_{ik}$$

$$u_{ik} \begin{cases} 1 & x_k \in c_i \\ 0 & x_k \notin c_i \end{cases}$$

$$D_{ik} = \|x_k - v_i\|^2$$

که در این عبارت منظور از x_k داده kام و منظور از v_i مرکز خوشه iام است. همچنین u_{ik} میزان تعلق صفر و یکی داده kام به خوشه iام را مشخص می‌کند. ||*|| نیز اپراتور نورم است که به جای آن هر نورمی مانند نورم اقلیدس یا نورم منهتن می‌تواند بنشیند. ما تا آخر این گزارش برای سادگی کار از نورم اقلیدس استفاده می‌کنیم.

برای محاسبه یک خوشه‌بندی مناسب باید تابع هدف آن را بهینه کنیم. برای بهینه کردن این تابع هدف نمی‌توانیم از مشتق گرفتن و مساوی صفر قرار دادن استفاده کنیم چرا که تابع u_{ik} یک تابع گسسته است و مقادیر گسسته ۰ و ۱ دارد لذا مشتق پذیر نیست. اگر تابع u_{ik} را یک تابع پیوسته در نظر بگیریم می‌توانیم از تابع هدف یک بار بر حسب u_{ik} و یک بار دیگر بر حسب v_i مشتق بگیریم تا تابع هدف را بهینه کنیم. با توجه به این که توان u_{ik} یک است، در مشتق بعدی از بین می‌رود، لذا یک توان m نیز برای رفع این مشکل برای آن قرار می‌دهیم. با توجه به نکات گفته شده، تابع هدف به شکل زیر تغییر می‌یابد:

$$J = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^{m} D_{ik}$$

$$u_{ik} \in [0,1]$$

$$D_{ik} = \|x_k - v_i\|^2$$

این تابع هدف در واقع الگوریتم FCM یا Fuzzy C-means را بیان می‌کند که یک الگوریتم خوشه‌بندی نرم است. این الگوریتم اولین بار در سال ۱۹۷۳ توسط دون[2] ارائه شد و در سال ۱۹۸۱ توسط بزدک[3] بهبود یافت [3].

در مشتق گرفتن از این تابع برای این که محدودیت یک شدن جمع میزان تعلقات یک عضو به همه دسته‌ها را اعمال کنیم از ضرایب لاگرانژ استفاده می‌کنیم. در نتیجه باید از تابع زیر مشتق بگیریم.

$$E = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^{m} D_{ik} - \sum_{k=1}^{n} \lambda_k \left(\sum_{i=1}^{c} u_{ik} - 1\right)$$

ابتدا یک بار نسبت به v_i مشتق می‌گیریم تا مرکز بهینه را پیدا کنیم.

$$\frac{\partial E}{\partial v_i} = \sum_{k=1}^{n} -u_{ik}^{m} \times 2(x_k - v_i) = 0$$

$$v_i \sum_{k=1}^{n} u_{ik}^{m} = \sum_{k=1}^{n} u_{ik}^{m} \times x_k$$

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}^{m} \times x_k}{\sum_{k=1}^{n} u_{ik}^{m}}$$

برای پیدا کردن میزان بهینه تعلقات یک بار نسبت به u_{ik} مشتق می‌گیریم:

$$\frac{\partial E}{\partial \mathrm{u_{ik}}} = m u_{ik}^{m-1} D_{ik} - \lambda_k = 0$$

$$u_{ik} = \left(\frac{\frac{\lambda_k}{m}}{D_{ik}}\right)^{\frac{1}{m-1}} \qquad (A)$$

یک بار بر حسب λ_k مشتق می‌گیریم.

$$\frac{\partial E}{\partial \lambda_k} = -\left(\sum_{i=1}^{c} u_{ik} - 1\right) = 0$$

$$\sum_{i=1}^{c} u_{ik} = 1 \quad (B)$$

[2] Dunn

[3] Bezdek

حالا با گذاشتن (A) در (B) خواهیم داشت.

$$\sum_{i=1}^{c}\left(\frac{\frac{\lambda_k}{m}}{D_{ik}}\right)^{\frac{1}{m-1}} = 1$$

$$1 = \left(\frac{\lambda_k}{m}\right)^{\frac{1}{m-1}} \sum_{i=1}^{c}\left(\frac{1}{D_{ik}}\right)^{\frac{1}{m-1}} \quad (C)$$

با تقسیم (A) بر (C) می‌توانیم u_{ik} بهینه را محاسبه کنیم.

$$\frac{(A)}{(C)} = u_{ik} = \frac{\left(\frac{1}{D_{ik}}\right)^{\frac{1}{m-1}}}{\sum_{j=1}^{c}\left(\frac{1}{D_{jk}}\right)^{\frac{1}{m-1}}}$$

$$= \frac{1}{\sum_{j=1}^{c}\left(\frac{\|x_i - v_k\|^2}{\|x_i - v_j\|^2}\right)^{\frac{1}{m-1}}}$$

$$= \frac{1}{\sum_{j=1}^{c}\left(\frac{\|x_i - v_k\|}{\|x_i - v_j\|}\right)^{\frac{2}{m-1}}}$$

توجه کنید که مقدار m میزان فازی‌سازی تعلقات را نمایش می‌دهد و معمولا بین ۲ و ۳ انتخاب می‌شود. (هر چه بزرگ‌تر باشد نافازی‌تر می‌شود.) اگر مقادیر را بزرگ‌تر انتخاب کنیم نتیجه علاوه بر این که نسبت به حالت خوشه‌بندی سخت بهبود نمی‌یابد بلکه بدتر نیز می‌شود.

با محاسبه مقدار مراکز خوشه و ماتریس تعلقات که انجام شد، می‌توانیم این محاسبات را در یک حلقه به طور مداوم انجام دهیم. شرط پایان حلقه می‌تواند عدم تغییر کمتر از یک اپسیلون مشخص شده برای تعلقات باشد و یا این که تعداد تکرار حلقه از پیش مشخص باشد.

الگوریتم FCM با توجه به محاسبات بالا به شرح زیر است:

۱. یک ماتریس تعلق اولیه با ابعاد $n \times c$ درست کن.
۲. مرکز تعلق را با توجه به فرمول زیر محاسبه کن:

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}^m \times x_k}{\sum_{k=1}^{n} u_{ik}^m}$$

۳. ماتریس تعلق را با توجه به فرمول زیر به‌روزرسانی کن:

$$u_{ik} = \frac{1}{\sum_{j=1}^{c}\left(\frac{\|x_i - v_k\|}{\|x_i - v_j\|}\right)^{\frac{2}{m-1}}}$$

۴. اگر شرط خاتمه (مثلا تعداد تکرار به تعداد ماکسیمم تکرار رسیده است.) برآورده شده است الگوریتم را تمام کن. اگر نه برو به ۲.

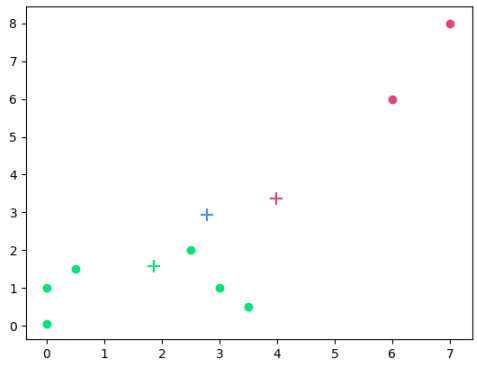
بررسی نتایج

در این پروژه این الگوریتم توسط زبان پایتون پیاده‌سازی شده است.

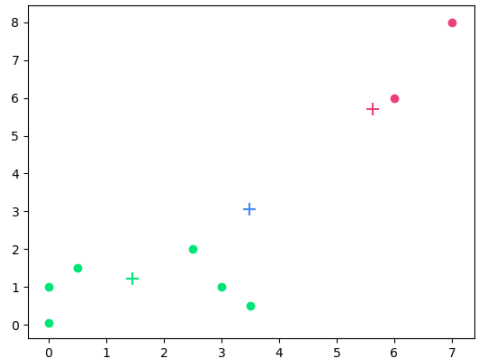
برای بررسی الگوریتم ۳ دسته مجموعه داده انتخاب شده است که هر کدام به این شرح است:

۱. یک دسته داده کوچک با ۳ مرکز فرضی که در کد هاردکد شده است.
۲. یک دسته داده نسبتا بزرگ به اندازه ۱۰۰ داده که حول سه مرکز فرضی به صورت نرمال پخش شده است. این مجموعه داده توسط تابع generate_2d_dataset_with_3_centers() تولید می‌شود.
۳. یک دسته داده بزرگ با ۲۰۰ داده که به صورت کاملا تصادفی پخش شده است. این مجموعه داده توسط تابع generate_dataset() تولید می‌شود.

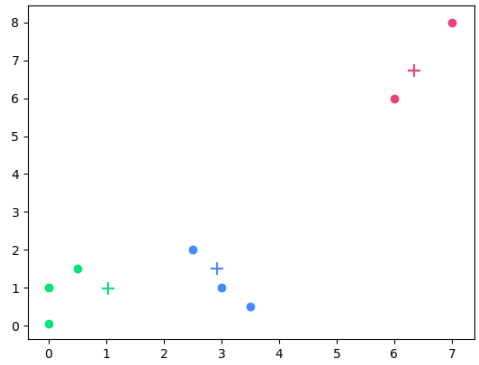
نتیجه اجرای الگوریتم بر روی مجموعه داده اول که کوچک است و قابلیت دیباگ را دارد به این شرح است:



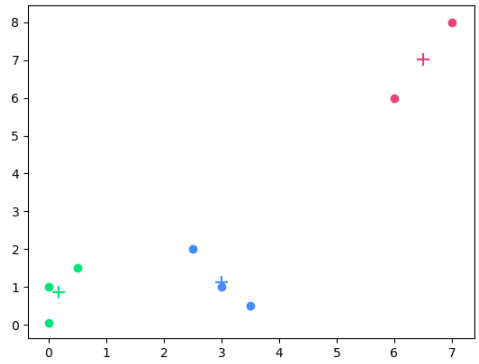
مرحله اول اجرا بر روی مجموعه داده اول



مرحله دوم اجرا بر روی مجموعه داده اول



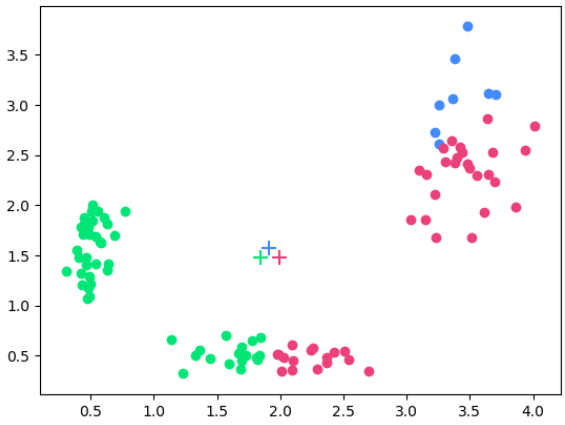
مرحله سوم اجرا بر روی مجموعه داده اول



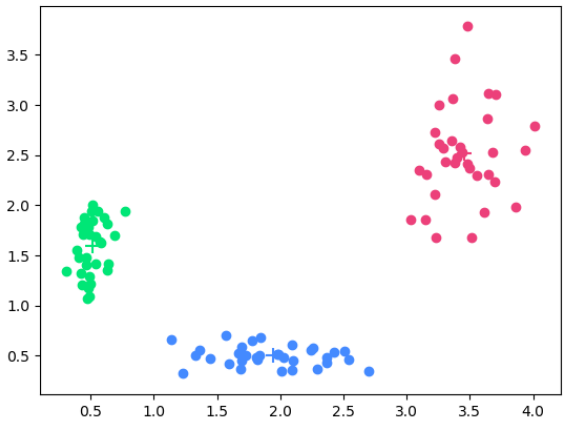
مرحله چهارم اجرا بر روی مجموعه داده اول

مراحل اجرا بر روی این مجموعه داده به صورت تصویر متحرک gif در این آدرس موجود است:
https://ceit.aut.ac.ir/~9531003/uploads/ci/2.gif

مراحل ابتدایی و پایانی اجرای الگوریتم بر روی مجموعه داده دوم به شرح زیر است:



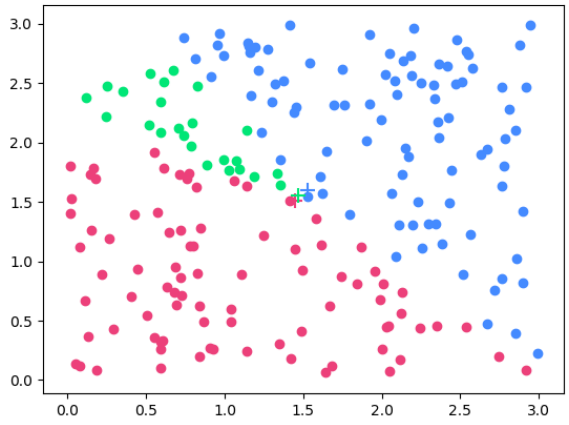
مرحله ابتدایی اجرا بر روی مجموعه داده دوم



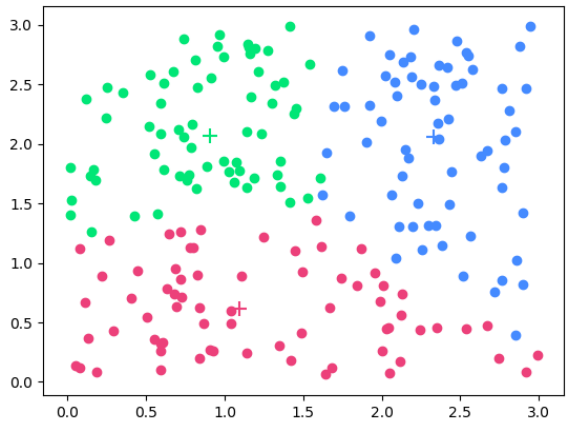
مرحله پایانی اجرا بر روی مجموعه داده دوم

مراحل اجرا بر روی این مجموعه داده به صورت تصویر متحرک gif در این آدرس موجود است:
https://ceit.aut.ac.ir/~9531003/uploads/ci/3.gif

مراحل ابتدایی و پایانی اجرای الگوریتم بر روی مجموعه داده سوم به شرح زیر است:



مرحله ابتدایی اجرا بر روی مجموعه داده سوم



مرحله پایانی اجرا بر روی مجموعه داده سوم

مراحل اجرا بر روی این مجموعه داده به صورت تصویر متحرک gif در این آدرس موجود است:
https://ceit.aut.ac.ir/~9531003/uploads/ci/1.gif

نتیجه‌گیری و کارهای آینده

الگوریتم FCM با توجه به این که در خوشه‌بندی از اطلاعات بیشتری مانند تاثیر یک عضو بر خوشه‌های دیگر استفاده می‌کند معمولا خوشه‌بندی بهتری را از k-means انجام می‌دهد.

برای بهتر کردن این پروژه و در آینده پیشنهاد می‌شود امکان کشیدن نمودار برای داده‌های با بیش از ۲ بعد فراهم شود.

همچنین امکان انجام این الگوریتم به صورت تعاملی برای مسائل آموزشی یکی دیگر از امکاناتی است که به بهبود این پروژه کمک می‌کند.

منابع

[1] “Cluster analysis,” *Wikipedia*, 06-Apr-2019. [Online]. Available: https://en.wikipedia.org/wiki/Cluster_analysis. [Accessed: 19-Apr-2019].

[2] J. C. Bezdek, R. Ehrlich, and W. Full, “FCM: The fuzzy c-means clustering algorithm,” *Computers & Geosciences*, vol. 10, no. 2-3, pp. 191–203, 1984.

[3] “Fuzzy C-Means Clustering,” *Clustering - Fuzzy C-means*. [Online]. Available: https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/cmeans.html. [Accessed: 19-Apr-2019].