دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

دانشکده مهندسی کامپیوتر

گزارش پروژه‌ی پایانی درس سیستم‌ها و روش‌های فازی

عنوان‌مقاله:

خوشه‌بندی فازی با استفاده از ماتریس کوواریانس فازی

Gustafson-Kessel Clustering Algorithm

دانشجو:

سید احمد نقوی نوزاد

ش-د: ۹۴۱۳۱۰۶۰

استاد درس:

دکتر عبادزاده

بهار ۱۳۹۶

بسمه تعالی

فهرست‌عناوین

۱) مقدمه

خوشه‌بندی داده‌ها از جمله روش‌های به اصطلاح بدون ناظر[۱] می‌باشد که در آن به دانستن برچسب داده‌ها احتیاجی نبوده و داده‌ها بنا به میزان شباهتی که با یکدیگر دارند، درون یک خوشه قرار می‌گیرند. یکی از معمول‌ترین روش‌های خوشه‌بندی، الگوریتم **K-Means** می‌باشد که نام دیگر آن **HCM**[۲] بوده و در آن هر داده پس از اتمام خوشه‌بندی، تنها می‌تواند به یک کلاس داده تعلق داشته و به سایر کلاس‌ها هیچ تعلقی نخواهد داشت. همین مسئله سبب می‌گردد تا مجموعه‌ی خوشه‌بندی‌های ممکن برای یک مجموعه‌داده با استفاده از **HCM** شدیدا بزرگ بوده و حالت بهینه در آن گاهی حتی قابل یافت‌شدن نمی‌باشد، خصوصا برای مواردی که برخی خوشه‌ها با یکدیگر همپوشانی بسیار دارند. در این حالت مسئله‌ی خوشه‌بندی به لحاظ الگوریتمی به نوعی رام‌ناشدنی[۳] می‌باشد. لذا منطقی‌تر خواهد بود اگر به هر داده، یک سری مقادیر تعلق به هر یک از کلاس‌ها نسبت داده شود تا آن‌که فقط به صورت به اصطلاح **Crisp** و یا همان صفر و یکی باشد. با این کار دیگر مرز بین کلاس‌ها در خوشه‌بندی نهائی، به اصطلاح سخت و دقیق نبوده و بلکه فازی‌تر خواهد بود، و البته داده‌هائی که به مراکز خوشه‌ها نزدیک می‌باشند، به مراتب مقدار تعلق بیشتری به خوشه‌ی مربوطه خواهند داشت. فایده‌ی دیگر خوشه‌بندی فازی در مورد داده‌های پرت می‌باشد که در روش خوشه‌بندی **HCM** شدیداً مشکل‌ساز بوده و حتی شاکله‌ی نهائی خوشه‌ها را به هم می‌ریزد. در حالت فازی با توجه به این‌که مقادیر تعلق پیوسته می‌باشند، داده‌های پرت نیز به هر خوشه تا حدی تعلق داشته و در نتیجه شاکله‌ی نهائی خوشه‌ها، ساختار معقول‌تری خواهد داشت.

حال در این‌جا قصد داریم تا الگوریتم جدیدی را مبتنی بر تئوری فازی جهت خوشه‌بندی معرفی نمائیم، که در واقع بهبودی بر خوشه‌بندی معمول فازی یا همان **FCM**[۴] می‌باشد. در این روش یک مفهوم جدید تحت عنوان ماتریس کوواریانس فازی معرفی می‌گردد و با استفاده از آن خوشه‌بندی فازی با دقت بیشتری نسبت به **FCM** در بسیاری موارد حاصل می‌گردد.

۲) شرح روش و پارامترها

در این قسمت ابتدا به معرفی مختصر الگوریتم **FCM** می‌پردازیم و سپس مزیت الگوریتم جدید را با اندکی تغییر در تابع هزینه‌ی **FCM** نشان خواهیم داد. معادله‌ی بهینه‌سازی الگوریتم **FCM** به صورت زیر می‌باشد:

$$J(\omega,\lambda)=\sum_{i=1}^{N}\sum_{j=1}^{k}\omega_{ij}^{\alpha}d_{ij}+\sum_{i=1}^{N}\lambda_i\left(\sum_{j=1}^{k}\omega_{ij}-1\right) \qquad (1)$$

که در آن به دنبال مقادیر بهینه برای مقادیر تعلق یا همان پارامتر ω و در نتیجه‌ی آن موقعیت مراکز فازی می‌باشیم. پس از حل معادله‌ی بهینه‌سازی مقادیر بهینه برای مقادیر تعلق و به دنبال آن مراکز فازی به صورت زیر خواهد بود:

---

1 Unsupervised
2 Hard C-Means
3 Intractable
4 Fuzzy C-Means

$$\omega_{ij}^{\ *} = \frac{1}{\sum_{\ell=1}^{k}\left(d_{ij}/d_{i\ell}\right)^{1/(\alpha-1)}} \qquad (2)$$

$$v_j^{\ *} = m_{fj}^{\ *} = \frac{\sum_{i=1}^{N}\omega_{ij}^{*\alpha}x_i}{\sum_{i=1}^{N}\omega_{ij}^{*\alpha}} \qquad (3)$$

که در آن علامت * به معنای مقادیر بهینه می‌باشد. ضمناً پارامتر α مشخص‌کننده‌ی میزان فازی‌بودن الگوریتم بوده و هر چه بیشتر باشد، مقادیر تعلق، فازی‌تر خواهند بود و زمانی که برابر با یک باشد، الگوریتم ما همان **C-Means** می‌باشد.

اما از آن‌جا که معیار فاصله در الگوریتم **FCM** که در فرمول بالا با d_{ij} نشان داده شد، از معیار فاصله‌ی اقلیدسی تبعیت می‌کند، لذا این الگوریتم طبعاً قادر به یافتن خوشه‌های کروی خواهد بود نه خوشه‌های بیضی‌وار. در این‌جا در الگوریتم پیشنهادی از یک معیار فاصله‌ی جامع‌تر تحت عنوان معیار فاصله‌ی ماهالانوبیس استفاده می‌کنیم، که اجازه می‌دهد تا شکل خوشه‌ها در ابعاد گوناگون به اندازه‌های متفاوت رشد نموده و به عبارتی خوشه‌های بیضی‌وار را نیز کشف خواهد نمود. جدای از این مسائل از آن‌جا که خوشه‌بندی خود یک مسئله‌ی بدون ناظر می‌باشد، لذا می‌بایست محدودیتی را به مسئله‌ی بهینه‌سازی مربوطه اضافه نمائیم تا از رشد بی‌رویه‌ی خوشه‌ها جلوگیری نماید. این شرط اضافه، ایجاد یک محدودیت بر روی دترمینان ماتریس کوواریانس هر خوشه می‌باشد که می‌بایست از یک مقدار مشخص اولیه بیشتر نگردد. معیار فاصله‌ی ماهالانوبیس و مسئله‌ی بهینه‌سازی نهائی به صورت زیر می‌باشند:

Mahalanobis Distance:

$$d_{ij}(\theta_j) = \left(x_i - v_j\right)^T M_j^{-1}\left(x_i - v_j\right), \quad 1 \le j \le k \qquad (4)$$

Constraint on Covariance Matrix Determinant:

$$\left|M_j\right| = \rho_j \quad , \quad \rho_j > 0 \qquad (5)$$

The Augmented Cost:

$$J\left(\omega,\lambda,\beta\right) = \sum_{i=1}^{N}\sum_{j=1}^{k}\omega_{ij}^{\ \alpha} d_{ij}\left(\theta_j\right) + \sum_{i=1}^{N}\lambda_i\left(\sum_{j=1}^{k}\omega_{ij} - 1\right) + \sum_{j=1}^{k}\beta_j\left(\left|M_j\right| - \rho_j\right) \qquad (6)$$

در نهایت پس از حل مسئله‌ی بهینه‌سازی مربوطه، مقادیر مراکز فازی همان مراکزی خواهند بود که در مورد **FCM** حاصل گردید. ولی مقدار بهینه‌ی معکوس ماتریس کوواریانس یا همان M_j^{*-1} به صورت زیر خواهد بود:

$$M_j^{*-1} = \frac{1}{\beta_j\left|M_j^*\right|}\sum_{i=1}^{N}\omega_{ij}^{\ \alpha}\left(x_i - v_j^*\right)\left(x_i - v_j^*\right)^T \qquad (7)$$

در اینجاست که می‌توانیم مفهوم جدید ماتریس کوواریانس فازی را از فرمول آخر برداشت نموده و به صورت زیر نمایش دهیم:

$$P_{fj}=\frac{\sum_{i=1}^{N}\omega_{ij}^{\alpha}\left(x_i-v_j^*\right)\left(x_i-v_j^*\right)^T}{\sum_{i=1}^{N}\omega_{ij}^{\alpha}}\quad ;\quad \alpha>1 \qquad (8)$$

در نهایت پس از یک سری محاسبات که در اصل مقاله نیز قید نگردیده است، به فرمول نهائی زیر جهت محاسبه‌ی معکوس ماتریس کوواریانس خواهیم رسید:

$$M_j^{*-1}=\left(\frac{1}{\rho_j\left|P_{fj}\right|}\right)^{1/n}P_{fj} \qquad (9)$$

به طوری که **n** برابر تعداد ابعاد یا همان ویژگی‌های مجموعه‌داده‌ی مورد استفاده می‌باشد. در اینجا نیز همان‌طور که پیش از این نیز قید گردید، به ازای مقدار $\alpha=1$ ماتریس کوواریانس فازی ما تبدیل به یک ماتریس کوواریانس سخت یا همان به اصطلاح **Crisp** خواهد شد که به آن ماتریس کوواریانس نمونه[5] نیز می‌گویند.

در این‌جا با توجه به مطالب قیدشده لازم است تا الگوریتم تکرارشونده‌ی پیشنهادی را معرفی نمائیم، تا در نهایت به مقادیر نهائی بهینه به ازای پارامترهای میانگین فازی و مقادیر تعلق دست یابیم. حال اگر پارامترهای مدنظر مربوط به هر کلاس در الگوریتم را در قالب $\theta_j=\left\{m_{fj},P_{fj}\right\}$ نمایش دهیم، با داشتن مجموعه‌داده‌ی ورودی $\left\{x_i\right\}$ و مقادیر اولیه برای پارامترهای هر کلاس به صورت $\theta_j^{\,0}=\left\{m_{fj}^{\,0},P_{fj}^{\,0}\right\}$ به صورت زیر عمل می‌کنیم:

(i) مقادیر $\left\{d_i\left(\theta_j^{(k)}\right)\right\}$ یا همان فاصله‌ی هر داده را از تمامی کلاس‌ها با استفاده از (۴) محاسبه می‌کنیم؛

(ii) مقادیر $\left\{\omega_{ij}^{(k)}\right\}$ یا همان مقادیر تعلق هر داده به هر یک از کلاس‌ها را با استفاده از (۲) محاسبه می‌کنیم. در موارد خاصی که فاصله‌ی یک داده از یکی از مراکز فازی برابر با صفر باشد، مقدار تعلق آن داده را به کلاس مربوطه برابر یک و به ازای سایر کلاس‌ها برابر صفر در نظر می‌گیریم.

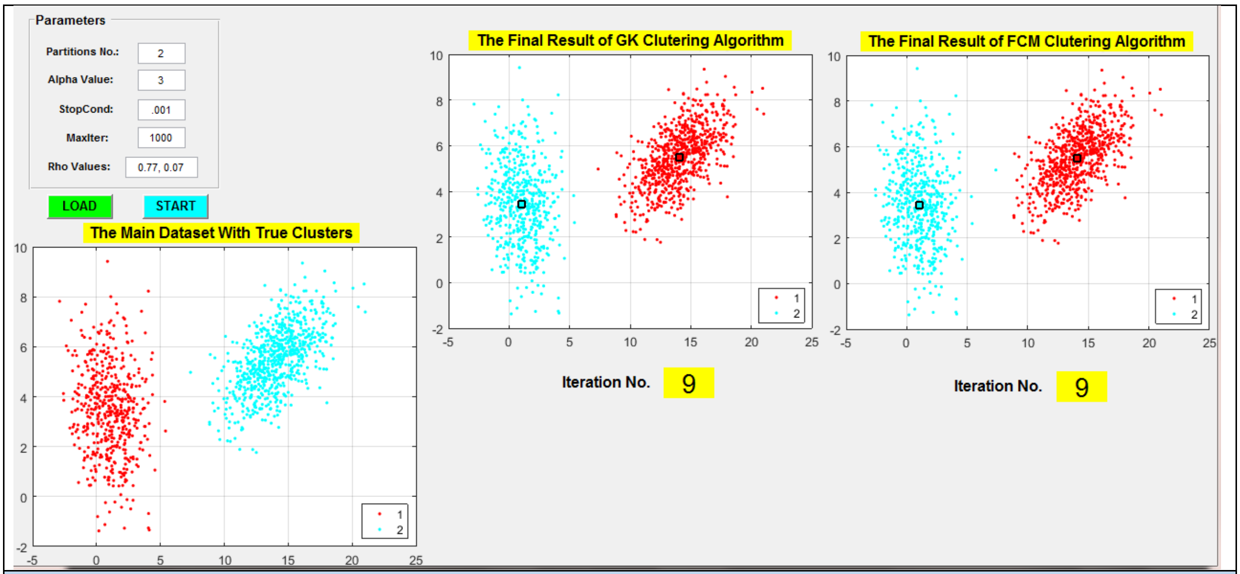
(iii) مقادیر تخمینی جدید به ازای $\theta_j^{(k+1)}$ را با استفاده از (۲)، (۸) و (۹) محاسبه می‌نمائیم. به مرحله‌ی **(i)** رفته و همین رویه را تا زمان رسیدن به یک ضابطه‌ی همگرائی خاص ادامه می‌دهیم.

۳) آزمایشات انجام‌شده

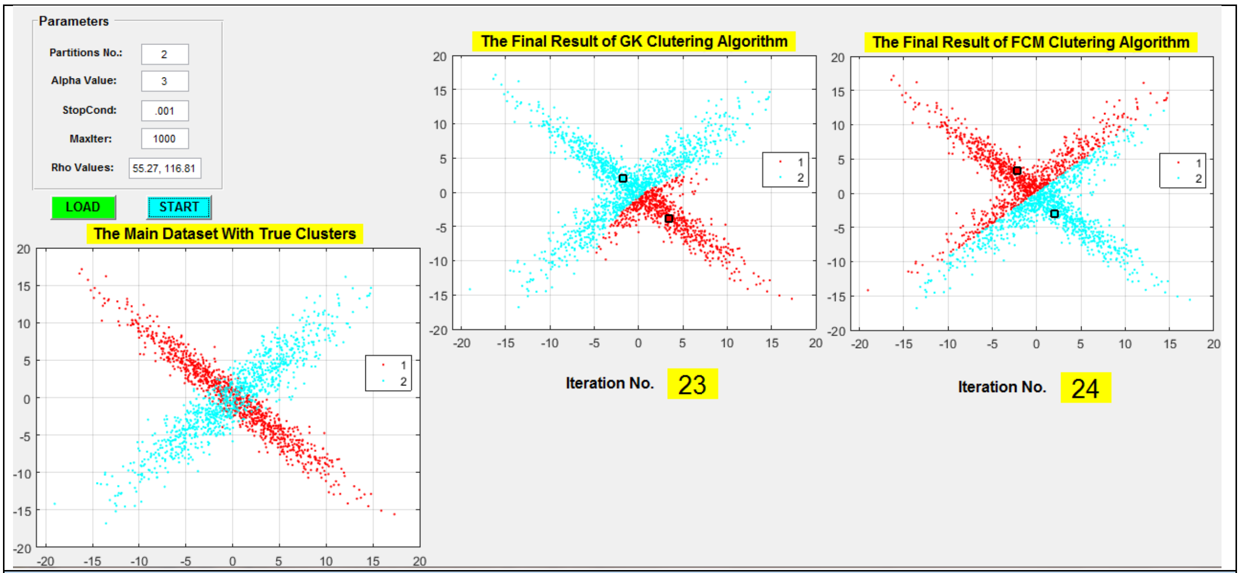
در این قسمت به انجام چند آزمایش بر روی مجموعه‌داده‌های مصنوعی با اشکال مختلف می‌پردازیم و در هر مورد، عملکرد الگوریتم پیشنهادی را با الگوریتم **FCM** مقایسه خواهیم نمود. در هر آزمایش تحلیل مربوطه در ذیل آن قید گردیده است.

---

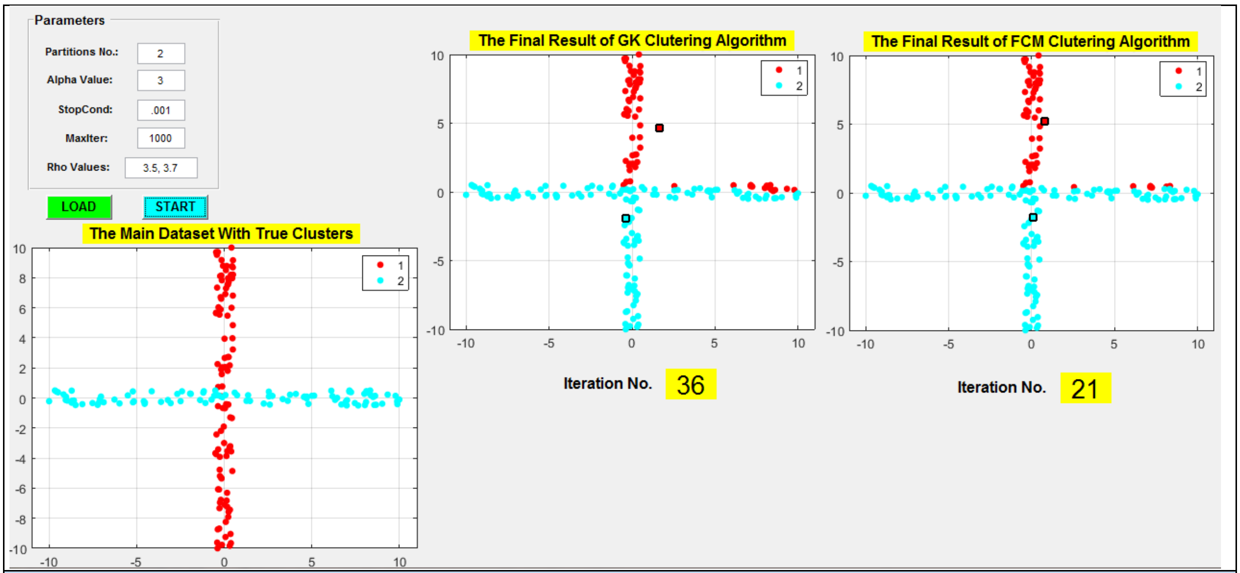
[5] Sample Class Covariance Matrix



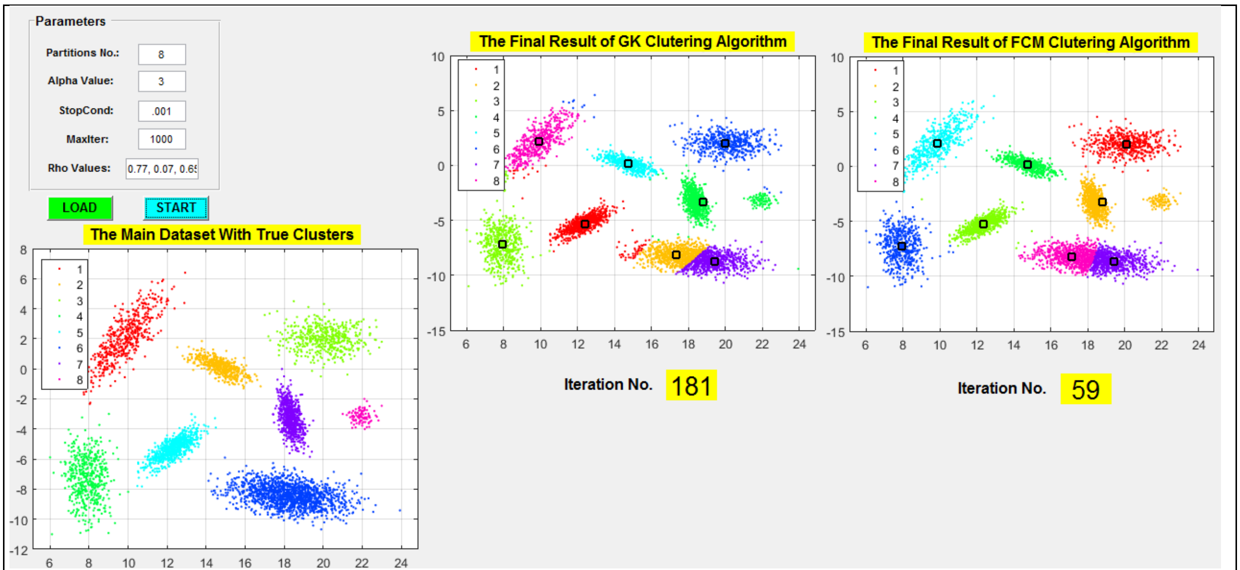
شکل ۱- همانطور که پیداست، هر دو الگوریتم پیشنهادی و FCM در یافتن خوشه‌های صحیح موفق عمل کرده‌اند. البته انتظار ما این بود که با توجه به عدم کروی بودن شکل خوشه‌ها، الگوریتم FCM نتواند موفق عمل کند که این گونه نشد.



شکل ۲- در این شکل دو خوشه داریم که با یکدیگر همپوشانی دارند. با توجه به آزمایشاتی که در خود مقاله‌ی مربوطه قید شده است، الگوریتم پیشنهادی در مورد این گونه خوشه‌ها و حتی با مقادیر اولیه‌ی دترمینان واقعی خود خوشه‌ها (که در این‌جا نیز اعمال شده است) می‌تواند بهینه عمل کند. اما همان طور که پیداست، نتیجه‌ی الگوریتم پیشنهادی با الگوریتم FCM چندان تفاوتی ندارد.



شکل ۳- در این شکل نیز مانند شکل ۲، دو خوشه داریم که دارای همپوشانی می‌باشند و البته از الگوی مجموعه‌داده‌ی مصنوعی مورد استفاده در اصل مقاله پیروی می‌نمایند. اما باز هم با وجود این‌که مقادیر اولیه‌ی دترمینان ها، دترمینان های اصلی خود خوشه ها می‌باشند، نتایج حاصله از دو الگوریتم چندان با یکدیگر تفاوتی ندارند.



شکل ۴- در این شکل شاهد تعداد زیادی خوشه های بیضی‌وار و با جهت های چرخش متفاوت می‌باشیم که به عبارت دیگر ماتریس کوواریانس هر یک از آن ها قطری نمی‌باشد. با کمال تعجب شاهدیم که الگوریتم **FCM** نیز قادر بوده است تا تعداد زیادی از این خوشه های به هم پیوسته را به درستی تشخیص دهد و البته الگوریتم پیشنهادی نیز خیلی متفاوت و حتی بهتر عمل نکرده است.

از آزمایشات انجام‌شده می‌توان نتیجه گرفت که با وجود آن‌که در الگوریتم پیشنهادی یک معیار فاصله‌ی جدید تحت عنوان فاصله‌ی ماهالانوبیس جهت کشف خوشه‌های ناموزون معرفی شده است و البته شرطی نیز جهت کنترل اندازه‌ی این خوشه‌ها اعمال شده است، اما به نظر می‌رسد که این الگوریتم نسبت به مقادیر اولیه شدیداً حساس می‌باشد و در نتیجه رسیدن به یک خوشه‌بندی ایده‌ال نیازمند دانش اولیه‌ای بسیار دقیق می‌باشد. اما در مورد الگوریتم خوشه‌بندی **FCM** نیز با تعجب شاهدیم که تقریبا در همگی

آزمایشات انجام‌شده، در مورد خوشه‌های به‌هم‌پیوسته نیز موفق عمل نموده است، و این مسئله خود با توجه به بررسی‌های مکرر پیاده‌سازی انجام‌شده، جای سوال دارد!؟

۴) مراجع

[1] **Gustafson, Donald E., and William C. Kessel. "Fuzzy clustering with a fuzzy covariance matrix."** ***Decision and Control including the 17th Symposium on Adaptive Processes, 1978 IEEE Conference on*****. IEEE, 1979.**