به نام خدا

**تمرین سوم درس مبانی داده کاوی**

**امیررضا حسینی ۹۸۲۰۳۶۳**

## سوالات تئوری

سوال ۱)

باتوجه به فرمول اصلی F-measure، Precision و Recall داریم:

محاسبه recall برای کلاس i برای خوشه j:

$$R(i,j) = \frac{n_{ij}}{n_i}$$

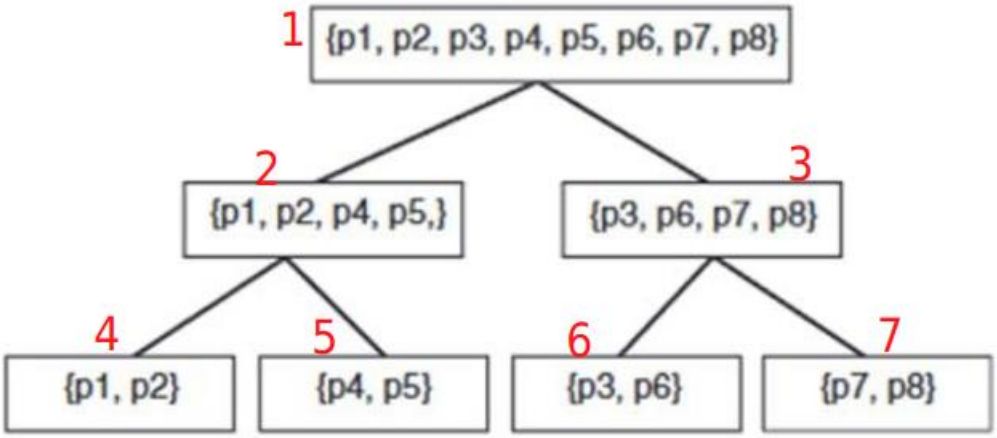
رابطه precision کلاس i برای خوشه j:

$$P(i,j) = \frac{n_{ij}}{n_j}$$

فرمول F-measure برای کلاس i و خوشه j:

$$F(i,j) = \frac{2R(i,j) \times P(i,j)}{R(i,j) + P(i,j)}$$

نحوه ترتیب دهی این کلاسترها برای بدست آوردن مقادیر:




حال از کلاستر بالا به پایین به محاسبه این مقادیر برای هرکدام از کلاس‌ها به صورت جداگانه میپردازیم:

Cluster1={P1, P2, P3, P4, P5, P6, P7, P8}

$$R(A,1) = \frac{3}{3} = 1$$

$$P(A,1) = \frac{3}{8} = 0.375$$

$$F(A,1) = \frac{2 \times 1 \times 0.375}{1 + 0.375} = 0.55$$

$$R(B,1) = \frac{5}{5} = 1$$

$$P(B,1) = \frac{5}{8} = 0.625$$

$$F(B,1) = \frac{2 \times 1 \times 0.625}{1 + 0.625} = 0.77$$

Cluster2={P1, P2, P4, P5}

$$R(A,2) = \frac{2}{3} = 0.667$$

$$P(A,2) = \frac{2}{4} = 0.5$$

$$F(A,2) = 0.57$$

$$R(B,2) = \frac{2}{5} = 1$$

$$P(B,2) = \frac{2}{4} = 0.5$$

$$F(B,2) = 0.44$$

Cluster3={P3, P6, P7, P8}

$$R(A,3) = \frac{1}{3} = 0.33$$

$$P(A,3) = \frac{1}{4} = 0.25$$

$$F(A,3) = 0.29$$

$$R(B,3) = \frac{3}{5} = 0.6$$

$$P(B,3) = \frac{3}{4} = 0.75$$

$$F(B,3) = 0.67$$

Cluster4={P1, P2}

$$R(A,4) = \frac{2}{3} = 0.667$$

$$P(A,4) = \frac{2}{2} = 1$$

$$F(A,4) = 0.8$$

$$R(B,4) = \frac{0}{5} = 0$$

$$P(B,4) = \frac{0}{2} = 0$$

$$F(B,4) = 0$$

Cluster5={P4, P5}

$$R(A,5) = 0$$

$$P(A,5) = 0$$

$$F(A,5) = 0$$

$$R(B,5) = \frac{2}{5} = 0.4$$

$$P(B,5) = \frac{2}{2} = 1$$

$$F(B,5) = 0.57$$

Cluster6={P3, P6}

$$R(A,6) = 0.33$$

$$P(A,6) = 0.5$$

$$F(A,6) = 0.4$$

$$R(B,6) = \frac{1}{5} = 0.2$$

$$P(B,6) = \frac{1}{2} = 0.5$$

$$F(B,6) = 0.29$$

Cluster7={ P7, P8}

$$R(A,7) = 0$$

$$P(A,7) = 1$$

$$F(A,7) = 0$$

$$R(B,7) = \frac{2}{5} = 0.4$$

$$P(B,7) = \frac{2}{2} = 1$$

$$F(B,7) = 0.57$$

حال با داشتن تمامی مقادیر برای تمامی کلاسترها، به پیدا کردن مقادیر نهایی برای هر دو کلاس A و B:

$$F(A) = max\{F(A,j)\} = max\{0.55,0.57,0.29,0.8,0,0.4,0\} = 0.8$$

$$F(B) = max\{F(B,j)\} = max\{0.77,0.44,0.67,0,0.57,0.29,0.57\} = 0.77$$

همچنین برای F-measure کلی را محاسبه میکنیم:

$$F = \sum_{1}^{2} \frac{n_i}{n} \max_i F(i,j) = \frac{3}{8}F(A) + \frac{5}{8}F(B) = 0.78$$

سوال ۲)

به دلیل نوع پراکندگی خاصی که این ۱۰۰ رکورد دارند، الگوریتم‌هایی که مبنای آنها شباهت بر حسب فاصله (k-means و یا نسخه‌های مشابه آن مثل kmeans++) است، دچار اشکال در یافتن تمایز بین داده‌ها شده و بهتر است از روش‌های دیگری مانند DBSCAN که به توده داده‌ها(نواحی پر چگال) نگاه میکند را استفاده کرد.

برای مثال، اگر داده ها به صورت یک خط بلند و باریک در فضای دو بعدی قرار داشته باشند، الگوریتم K-means با هر تعداد kتنها یک خوشه غیر خالی را بر می گرداند. و یا اگر تمامی داده‌ها در فضا نسبت به هم با یک فاصله پراکنده شده باشند. و یا نویز بسیار زیاد باشد به طوری که فاصله نقاط از یکدیگر به شدت زیاد باشد.

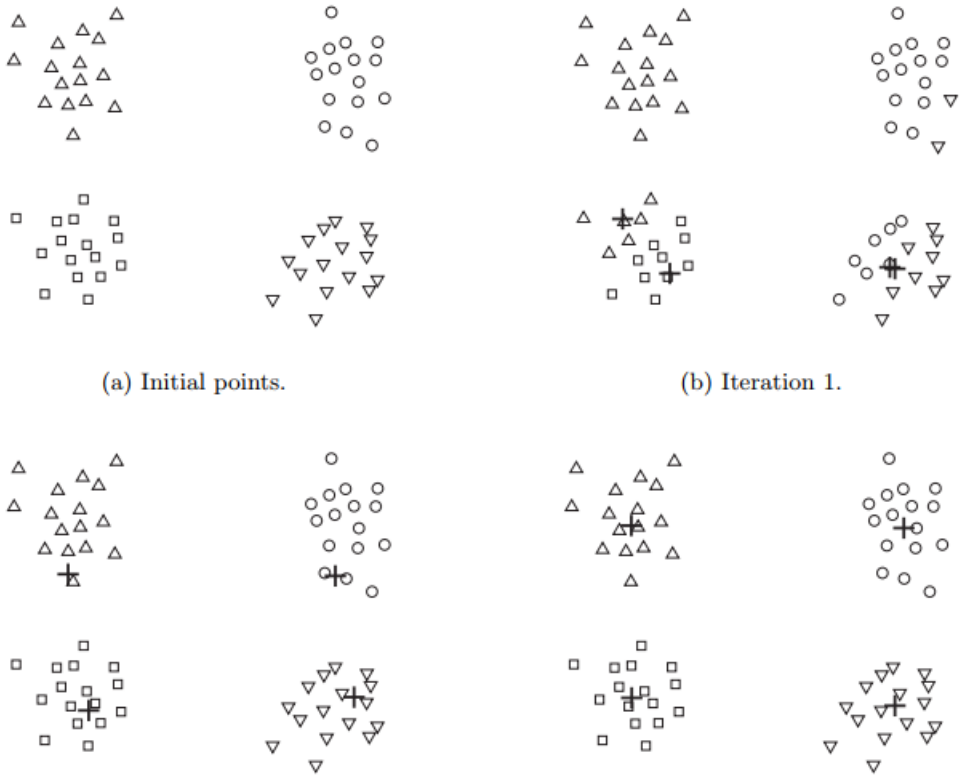
در حالت دیگر میتوان از روش‌های سلسله مراتبی مانند single link که به هرکدام از داده‌ها به عنوان یک خوشه نگاه میکند، استفاده کرد.

سوال ۳)

همانطور که میدانیم انتخاب نقاط مرکزی مهمترین بخش همگرایی به جواب گلوبال در الگورتیم k-means پایه است.

طبق مثال صفحات ۳۲۲ تا ۳۲۴ کتاب آقای تان میتوان دید که اجرای متفاوت میتواند مقدار SSEهای مختلفی را بدهد. و حالا ممکن است در کلاستر بندی، مانند هر الگوریتم هیوریستیک دیگری، در نقاط بهینه محلی گیر بیوفتیم(local minima of SSE) و نتوانیم به بهینه سراسری( global minima of SSE) حتی با اجراهای متوالی برسیم. اما در اینجا چون در هر اجرا، نقاط ابتدایی را رندوم فرض میکنیم، ممکن است نتیجه به بهترین حالت همگرا نشود. البته این حالت به نحوه پراکندگی دیتا اولیه و تعداد کلاستربندی‌ها میباشد.

در مثال ۵.۲ کتاب آقای تان به این نکته یادآور میشود که مثلا برای حالت زیر که داده ها شامل دو جفت خوشه است، جایی که خوشه ها در هر جفت (بالا–پایین) نسبت به خوشه های جفت دیگر، به یکدیگر نزدیک ترند. در شکل (b-d) نشان می دهد که اگر با دو مرکز اولیه برای هر جفت از خوشه ها شروع کنیم، حتی زمانی که هر دو مرکز در یک خوشه واحد قرار دارند، مراکز نقطه های خوشه ها از هم جدا می شوند.

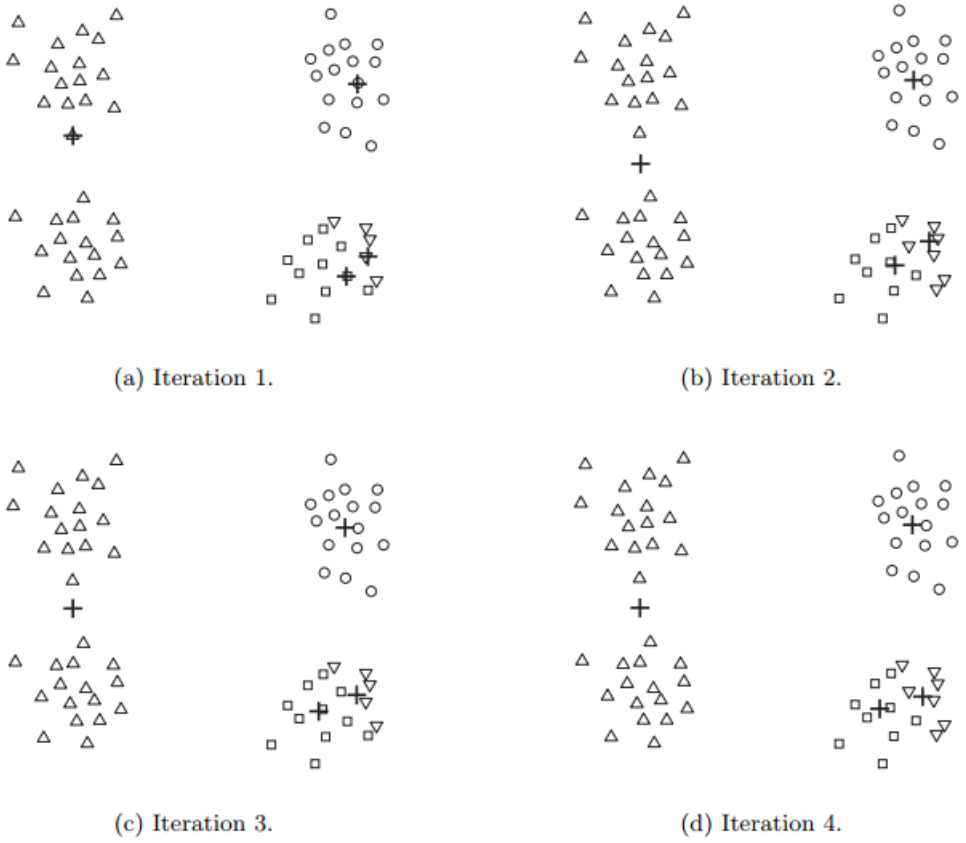


(a) Initial points. (b) Iteration 1.

(c) Iteration 2. (d) Iteration 3.

**Figure 5.6.** Two pairs of clusters with a pair of initial centroids within each pair of clusters.

اما همانطوری که در شکل زیر مشخص است، اگر یک زوج از کلاسترها یک مرکز و بقیه ۳تا داشته باشند، خوشه بندی نهایی به درستی صورت نمیگیرد.



(a) Iteration 1. (b) Iteration 2.

(c) Iteration 3. (d) Iteration 4.

**Figure 5.7.** Two pairs of clusters with more or fewer than two initial centroids within a pair of clusters.

این مثال به وضوح نشان میدهد که با انتخاب رندوم مراکز خوشه‌ها ممکن است بقیه مراحل نیز تحت تاثیر قرار بگیرند و به جواب بهینه سراسری همگرا نشویم. در این صورت حتی ممکن است اجراهای متوالی نیز به راحتی به جواب نرسند و در بهینه محلی گیر کنند.

یکی از راه‌هایی که الگوریتم را از دام بهینه‌های محلی خارج میکند و به سمت بهینه سراسری همگرا میکند، انتخاب هوشمندانه‌تر نقاط ابتدایی است که یکی از آنها k-means++ میباشد.

سوال۴)

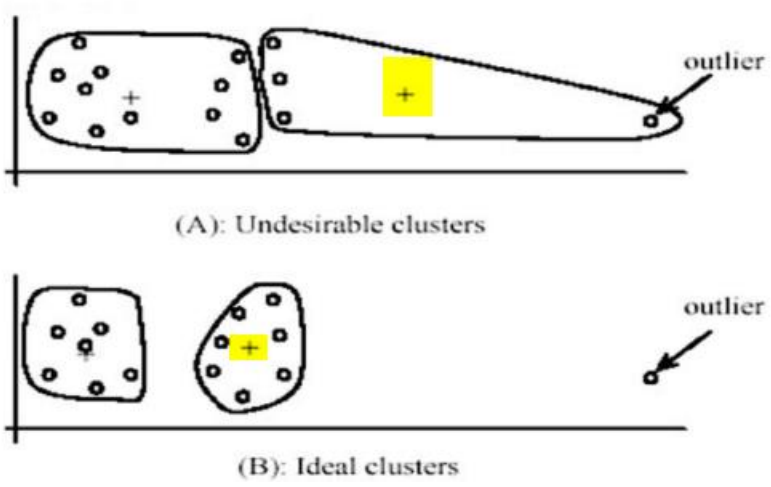
الف) باتوجه به مطالب داخل اسلایدها، K-means و K-medoids شباهت‌های زیادی در خوشه‌بندی دارند اما تفاوت بارز آنها در انتخاب نقاط مرکزی برای هر خوشه میباشد.
یکی از آنها میانگین را انتخاب میکند و دیگری میانه.

یکی از مهمترین مزایای K-means این است که سرعت بسیار بالایی دارد و به راحتی قابل پیاده‌سازی است. همچنین، این الگوریتم به دلیل سادگی طراحی، به طور کلی برای داده‌هایی مناسب است که خوشه‌های به صورت کروی هستند. با این حال، یکی از محدودیت‌های اصلی K-means این است که آن حساس به انتخاب نقاط اولیه است و نتایج آن به شدت به انتخاب اولیه متمرکز است.

از سوی دیگر، K-medoids، یک نسخه اصلاح شده از K-means است (مانند الگوریتم PAM صفحه ۴۵۷ کتاب آقای هان) که در آن، به جای استفاده از میانگین نقاط به عنوان نقطه مرکزی خوشه، از یک نقطه واقعی در داده‌ها استفاده می‌شود. این باعث می‌شود که K-medoids نسبت به پارامترهای خود مقاوم تر باشد و نتایج آن کمتر به انتخاب اولیه بستگی داشته باشد.

با این حال، K-medoidsبه طور کلی سرعت پایین‌تری نسبت به K-means دارد و به دلیل پیچیدگی محاسباتی آن، برای داده های بزرگ غیر عملی است. همچنین، به دلیل اینکه درK-medoids ، از یک نقطه واقعی به عنوان مرکز خوشه استفاده می‌شود، این الگوریتم در مواجهه با داده‌هایی که خوشه‌های غیر منظمی دارند، مشکل دارد.

به طور کلی، هر یک از این الگوریتم‌ها برای شرایط خاصی مناسب هستند K-means .برای داده هایی مناسب است که همگن هستند و خوشه های آن‌ها به صورت کروی هستند و همچنین داده‌های پرت (outlier) و نویز در آن به ندرت وجود دارد.، در حالی که K-medoids برای داده هایی با خوشه های منظم و یا غیر منظم پیشنهاد می شود.




همانطور که مشاهده میشود وجود داده پرت به شدت میتواند نتیجه clustering را تحت تاثیر قرار بدهد.

ب) باتوجه به صفحه ۴۶۰ کتاب هان دو مورد الگوریتم AGNES و به طور کلی روش‌های سلسله مراتبی(بالا به پایین و پایین به بالا) معمولاً سرعت بالاتر، صرفه‌جویی در منابع سیستم و قابلیت اجرای بهینه‌تر(مثلا موازی سازی) را دارند. با این حال، تفاوت‌های اصلی بین طرح‌های خوشه‌بندی سلسله مراتبی و پارتیشن در نحوه الگوریتم استفاده شده برای جداسازی داده‌هاست.

مثلا در روش‌های خوشه بندی سلسله مراتبی نیازی به دانستن تعداد خوشه‌ها (k) از ابتدا نیست ولیکن امکان دارد کیفیت خوشه بندی و تفکیک بین داده‌ها به خوبی الگوریتم‌های الگوریتم‌های کلاسترینگ (مثل k-means) نباشد. همچنین برای الگوریتم‌های سلسله مراتبی خوشه هایی با هر مدل شکل (shape) هندسی دلخواه ایجاد کنیم. ولی در پارتیشنینگ معمولا کلاسترها کروی شکل میشوند.