التحليل العنقودي الهرمي

Hierarchical Cluster Analysis

سمنار مقدم

من قبل

المدرس

بهاء عبد الرزاق قاسم

Bahaa Abdulrasaq Qasim

المدرس المساعد

أحمد هشام محمد

Ahmed Husham Albasri

قسم الإحصاء – كلية الادارة والاقتصاد

جامعة البصرة

التحليل العنقودي الهرمي

1- ما المقصود بالتحليل العنقودي .

2- فؤاد استخدام التحليل العنقودي ومجالات تطبيقه.

3- انواع التحليل العنقودي

- التحليل العنقودي الهرمي
- التحليل العنقودي باستخدام طريقة المتوسطات

4- مراحل تنفيذ التحليل العنقودي الهرمي

- تكوين مصفوفة القرابة Proximities Matrix
- طرق العنقدة الهرمية

5- مثال تطبيقي

1 – ما المقصود بالتحليل العنقودي

يقصد به إجراءات تهدف الى تصنيف مجموعة حالات او متغيرات بطرق معينة و ترتيبها داخل عناقيد بحيث تكون الحالات المصنفة داخل العنقود الواحد متجانسة [أي متشابه نسبياً فيما بينها] فيما يتعلق بخصائص محددة إلا أنها مختلفة عن حالات أو متغيرات أخرى موجودة في العناقيد الأخرى، أي تكون ذات تشابه اقل نسبياً بين حالات أو متغيرات العناقيد الأخرى. وعليه يمكن الاستنتاج بأن اساس التحليل العنقودي يتمثل في ترتيب الحالات او المتغيرات بشكل عناقيد بحيث يعمل على تصغير التباين داخل العنقود الواحد و تعظيم التباين بين العناقيد المختلفة [سليمان،2012،ص143].

2- فوائد استخدام التحليل العنقودي ومجالات تطبيقه

تتحدد فائدة التحليل العنقودي في :

1. ايجاد المجموعات الحقيقة على وفق محددات معينة .
2. واختصار البيانات .

اما مجالات تطبيقه :

يطبق التحليل العنقودي في مجالات متعددة نذكر منها

1. علم الاحياء .
2. استرجاع المعلومات .
3. المناخ .
4. علم النفس والطب .
5. إدارة الاعمال .

3- أنواع التحليل العنقودي

يوجد نوعان من التحليل العنقودي المتبعة في تصنيف البيانات هما :

1. **التحليل العنقودي الهرمي** . Hierarchal Cluster Analysis
2. **التحليل العنقودي غير الهرمي [طريقة المتوسطات]** K- Means

التحليل العنقودي الهرمي

يعد هذا الأسلوب من الأساليب المفضلة في التحليل العنقودي وفيه يتم عنقدة n من المفردات و بشكل متسلسل في m من العناقيد اذ يكون c_1 اضعف العناقيد و c_m أكثرها قوة .(مصطفى، 2005، ص28)

أساليب العنقده الهرمية :

(1) أسلوب التقسيم [الخلاف] . **The a divisive technique**

يتلخص عمل هذا الأسلوب على افتراض وجود عنقود واحد للمفردات ثم يتم تجزئة هذا العنقود الى عناقيد جزئية ثم بعد ذلك تجزء العناقيد إلى عناقيد اصغر وهكذا نستمر لحين يتكون لكل مفردة عنقود خاص بها (صوار وآخرون،2013 ،ص13).

(2) أسلوب التكتل [التجميع] . **Agglomerative technique**

مبدأ عمل هذا الأسلوب مغاير تماماً للأسلوب أعلاه اذ يفترض ان كل مفردة تصف عنقود جزئي خاص بها ثم يتم تجميع العناقيد الجزئية المتشابهة في عناقيد جزئية أكثر شمولاً، وتكرر العملية لحين الحصول على عنقود واحد يضم جميع المفردات (هيثم، 2011، ص632).

4- مراحل تنفيذ التحليل العنقودي الهرمي

1. تكوين مصفوفة القرابة proximities matrix

تكوين مصفوفة التقارب (proximities matrix) وهي مصفوفة مربعة من درجة n متماثلة عناصرها مقياس يعبر عن المسافة بين كل زوجين من البيانات (الحالات) ويرمز لها بالرمز D (هيثم واخرون،2011، ص630) و تعرف بالشكل أدناه:

$$D = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ \vdots & \vdots & \cdots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{bmatrix}$$

وتوجد هناك عدد من الطرق لقياس المسافة بين عناصر مصفوفة القرابة منها :

1- المسافة الاقليدية **Euclidean Distance.**

2- صيغة منكوسكي **Minkowsski metric .**

3- مسافة القطاع **City Block distance**

- **المسافة الاقليدية Euclidean Distance**

نعد صيغة المسافة الاقليدية من اشهر مقاييس المسافة وعرفت بتلك التسمية نسبة إلى العالم الرياضي الإسكندنافي اقليدس وتعرف رياضيا كالاتي:

$$D(x,y) = \left[\sqrt{\sum_{j=1}^{p}(x_j - y_j)^2}\right] \quad ; \quad p = \text{عدد الحالات}$$

- **صيغة منكوسكي Minkowsski metric**

$$D(x,y) = \left[\sum_{j=1}^{p}|x_j - y_j|^r\right]^{\frac{1}{r}} \quad ; \quad r > 0$$

- **مسافة القطاع City Block Distance**

$$D(x,y) = \left[\sum_{j=1}^{p}|x_j - y_j|\right]$$

(2) تحديد طريقة ربط العناقيد مع بعضها البعض بالاعتماد على المسافة المحسوبة بين العناقيد ومن هذه الطرق الاتي :

- **طريقة الجوار الاقرب (الربط المنفرد)**

حيث يتم على اساسها ايجاد اصغر مسافة لكل زوج من العناقيد و ربطهما معاً بحسب الصيغة الآتية:

$$D(A,B) = min_{x_i \in A\ ,y_i \in B}(d(x_i\ ,y_i)$$

- **طريقة الجوار الابعد (الربط الشامل)**

يتم في هذه الطريقة ايجاد اكبر مسافة بين العناقيد و ربطهما معاً بحسب الصيغة الاتية:

$$D(A,B) = MAX_{x_i \in A\ ,y_i \in B}(d(x_i\ ,y_j)$$

- **طريقة الربط باستخدام المتوسط (المعدل) Average**

تستخدم هذه الطريقة لربط عنقودين وذلك بالاعتماد على متوسط المسافة بين نقطة من العنقود الاول (A) n_A و نقطة من العنقود لثاني (B) n_B ووفق الصيغة الاتية:

$$D(A,B) = \frac{\sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d(x_i, y_j)}{n_A n_B}$$

مثال تطبيقي

في دراسة إحصائية لتصنيف عدد من الحقول الزراعية (O_i) وفقاً لنسبة ملوحة التربة (x_i) ونسبة السماد الكيماوي(y_i) ، ولإجراء عملية التصنيف سيتم استخدام أسلوب التحليل العنقودي ، مستخدمين صيغة المسافة الاقليدية ، تقنية ربط الأوساط لإيجاد العناقيد، مع رسم مخطط العناقيد

	x_i	y_i
	0.40	0.53
	0.22	0.38
	0.35	0.32
	0.26	0.19
	0.08	0.41
	0.45	0.30

الحل :
تكوين مصفوفة القرابة

$$D = \begin{pmatrix} \mathbf{0} & P_1P_2 & P_1P_3 & P_1P_4 & P_1P_5 & P_1P_6 \\ & \mathbf{0} & P_2P_3 & P_2P_4 & P_2P_5 & P_2P_6 \\ & & \mathbf{0} & P_3P_4 & P_3P_5 & P_3P_6 \\ & & & \mathbf{0} & P_4P_5 & P_4P_6 \\ & & & & \mathbf{0} & P_5P_6 \\ & & & & & \mathbf{0} \end{pmatrix}$$

المرحلة الأولى حساب عناصر مصفوفة القرابة حسب صيغة المسافة الاقليدية

$$D(P_i, P_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

- حساب المسافة بين الحقل 1 والحقل 2 $D(P_1, P_2)$

$P_1 = (x_1, y_1)$; $P_2 = (x_2, y_2)$

$P_1 = (0.40, 0.53)$; $P_2 = (0.22, 0.38)$

$$D(P_1, P_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

$$D(P_1, P_2) = \sqrt{(0.40 - 0.22)^2 + (0.53 - 0.38)^2}$$

$$D(P_1, P_2) = \sqrt{0.0549}$$

$$D(P_1, P_2) = 0.234$$

- حساب المسافة بين الحقل 1 والحقل 3 $D(P_1, P_3)$

$$P_1 = (x_1, y_1) \quad ; \quad P_3 = (x_3, y_3)$$
$$P_1 = (0.40, 0.53) \quad ; \quad P_3 = (0.35, 0.32)$$
$$D(P_1, P_3) = \sqrt{(x_1 - x_3)^2 + (y_1 - y_3)^2}$$
$$D(P_1, P_2) = \sqrt{(0.40 - 0.35)^2 + (0.53 - 0.32)^2}$$
$$D(P_1, P_3) = \sqrt{0.0466}$$
$$D(P_1, P_3) = 0.216$$

- حساب المسافة بين الحقل 2 والحقل 4 $D(P_2, P_4)$

$$P_2 = (x_2, y_2) \quad ; \quad P_4 = (x_4, y_4)$$
$$P_2 = (0.22, 0.38) \quad ; \quad P_4 = (0.26, 0.19)$$
$$D(P_2, P_4) = \sqrt{(x_2 - x_4)^2 + (y_2 - y_4)^2}$$
$$D(P_2, P_4) = \sqrt{(0.22 - 0.26)^2 + (0.38 - 0.19)^2}$$
$$D(P_2, P_4) = \sqrt{0.0377}$$
$$D(P_2, P_4) = 0.194$$

وبنفس الطريقة يمكن ايجاد المسافات بين ازواج المفردات الحقول لغرض تكوين مصفوفة التقارب، وعليه سنحصل على المصفوفة التالية :

مصفوفة القرابة

Proximate Matrix

	P1	P2	P3	P4	P5	P6
P1	0	0.234	0.216	0.368	0.342	0.235
P2		0	0.143	0.194	0.143	0.244
P3			0	0.158	0.285	0.102
P4				0	0.284	0.220
P5					0	0.386
P6						0

Min Value (Element)

$$D(P3, P6) = 0.102$$

المرحلة الثانية : تحديث مصفوفة القرابة لايجاد العناقيد حسب طريقة ربط الأوساط Average Linkage

(5 دمج الحقل 3 مع الحقل

	P1	P2	P3,P6	P4	P5
P1	0	0.234		0.368	0.342
P2		0		0.194	0.143
P3,P6			0		
P4				0	0.284
P5					0

ويتم التحديث على وفق الصيغة التالية

$$D(P_k,(P_i,P_j)) = \frac{1}{2}\left((P_k,P_i)+(P_k,P_j)\right)$$

وعليه سيتم تحديث النقاط التالية

$D(P_1,(P_3,P_6))$ ، $D(P_2,(P_3,P_6))$ ،$D((P_3,P_6),P_4)$، $D((P_3,P_6),P_5)$

- حساب المسافة $D(P_1,(P_3,P_6))$

$$D(P_1,(P_3,P_6)) = \frac{1}{2}\left((P_1,P_3)+(P_1,P_6)\right)$$
$$= \frac{1}{2}(0.216+0.235)$$
$$=0.226$$

- حساب المسافة $D(P_2,(P_3,P_6))$

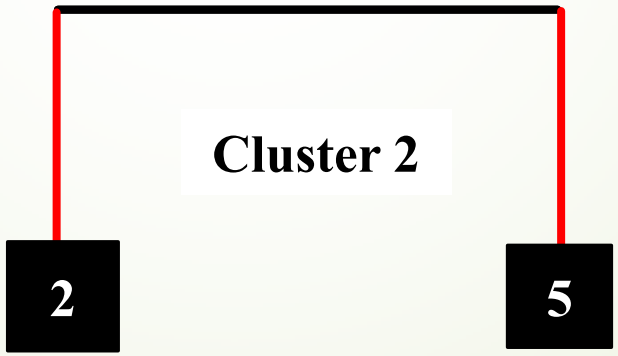
$$D(P_2,(P_3,P_6)) = \frac{1}{2}\left((P_2,P_3)+(P_2,P_6)\right)$$
$$= \frac{1}{2}(0.143+0.244)$$
$$=0.194$$

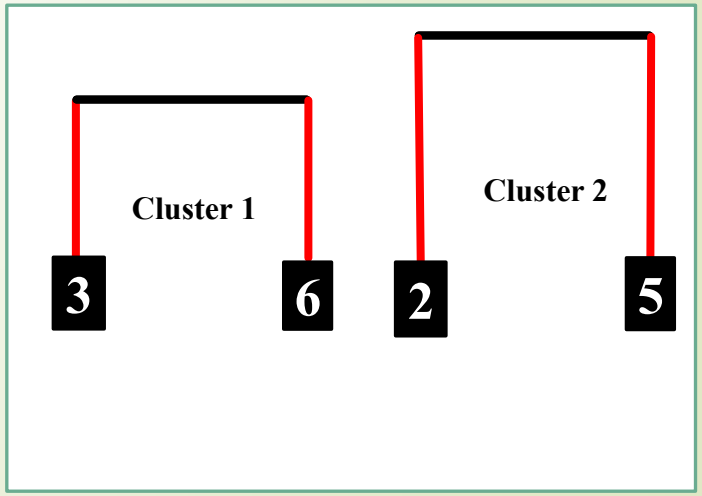
المصفوفة المحدثة

	P1	P2	P3,P6	P4	P5
P1	0	0.234	0.226	0.368	0.342
P2		0	0.193	0.194	0.143
P3,P6			0	0.189	0.335
P4				0	0.284
P5					0

← Min Distance

$$D(P2, P5) = 0.143$$





المرحلة الثالثة : تحديث مصفوفة القرابة من خلال دمج الحقل 2 مع الحقل 5

	P1	P2,P5	P3,P6	P4
P1	0		0.226	0.368
P2,P5		0		
P3,P6			0	0.189
P4				0

وعليه سيتم تحديث النقاط التالية

$$D(P_1,(P_2,P_5))\ ،\ D((P_2,P_5),(P_3,P_6))،\ D((P_2,P_5),P_4)$$

- حساب المسافة $D(P_1,(P_2,P_5))$

$$\boldsymbol{D(P_1,(P_2,P_5))} = \frac{1}{2}\left((\boldsymbol{P_1},\boldsymbol{P_2}) + (\boldsymbol{P_1},\boldsymbol{P_5})\right)$$

$$= \frac{1}{2}\,(0.234+0.342)$$

$$=0.288$$

- حساب المسافة $D((P_2,P_5),(P_3,P_6))$

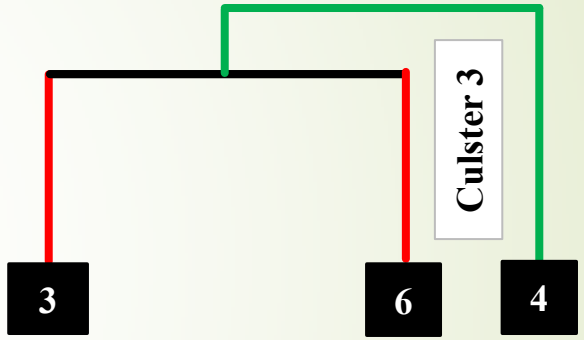
$$D((P_2,P_5),(P_3,P_6))=\frac{1}{2}((P_2(P_3,P_6))+(P_5(P_3,P_6)))$$

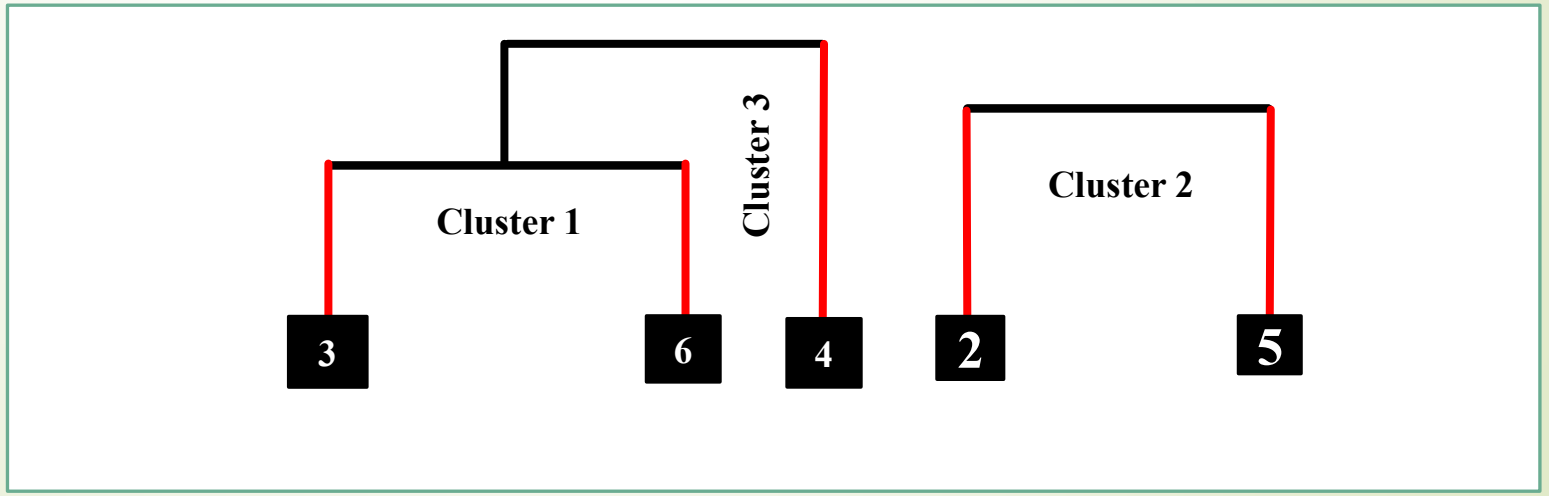
$$=\frac{1}{2}(0.193+0.335)$$

$$=0.264$$

	P1	P2,P5	P3,P6	P4
P1	0	0.288	0.226	0.368
P2,P5		0	0.264	0.239
P3,P6			0	0.189
P4				0

← Min Distance



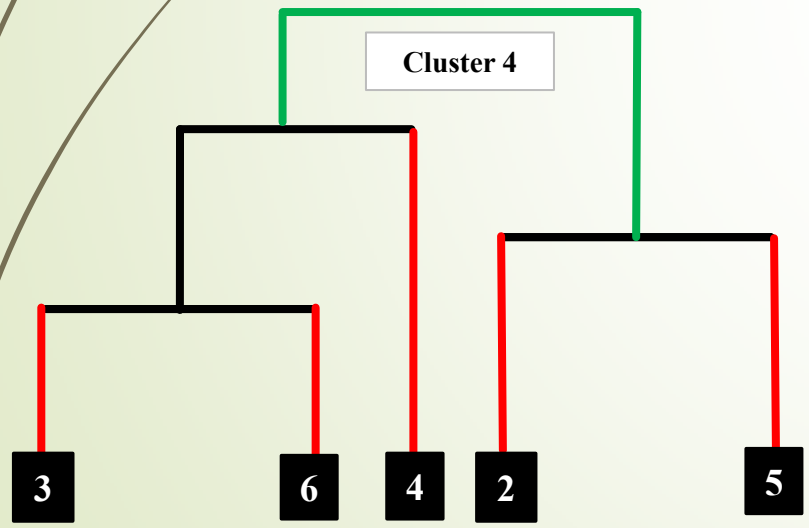


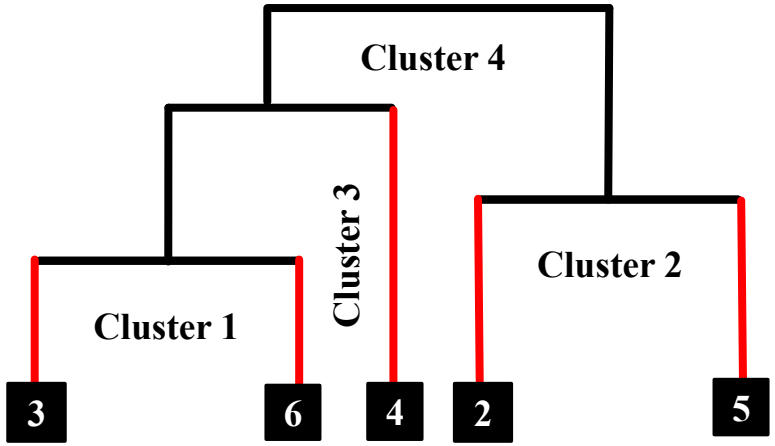
المرحلة الرابعة : تحديث مصفوفة القرابة

	P1	P2,P5	(P3,P6),P4
P1	0	0.288	
P2,P5		0	
(P3,P6),P4			0

	P1	P2,P5	(P3,P6),P4
P1	0	0.288	0.297
P2,P5		0	0.232
(P3,P6),P4			0

← Min Distance

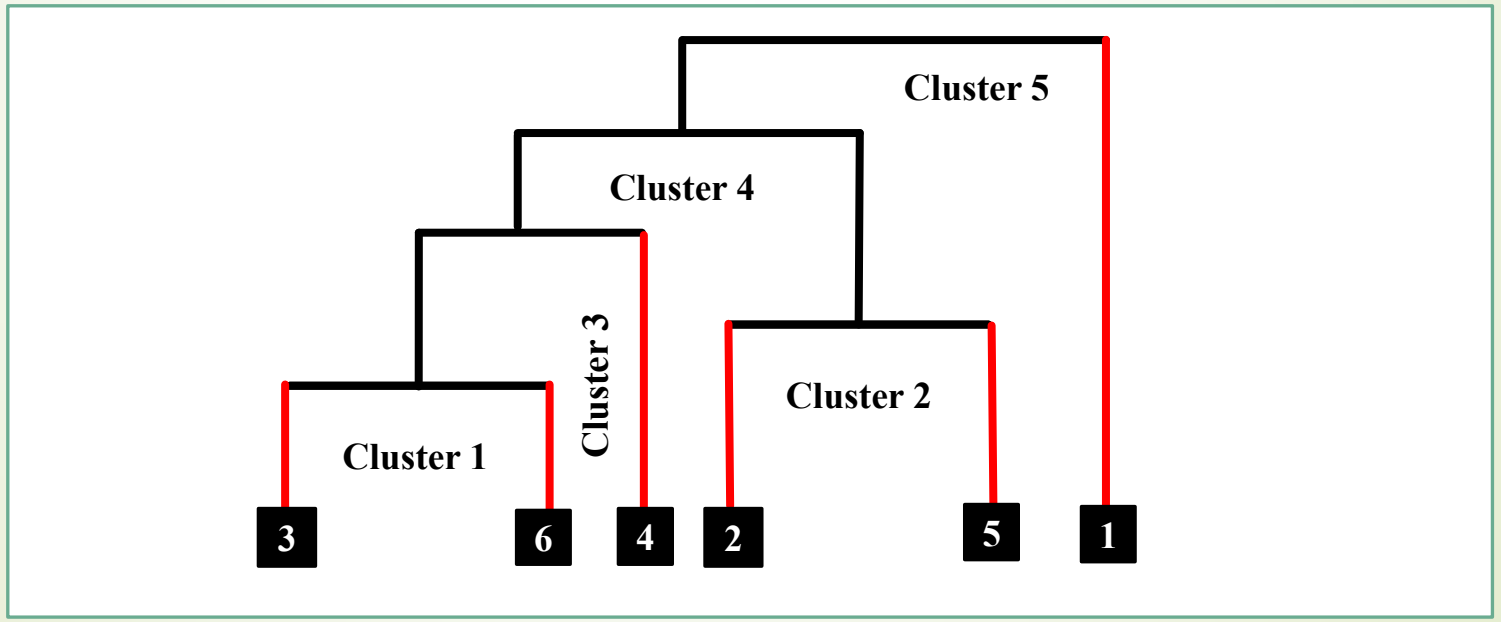




المرحلة الخامسة : تحديث مصفوفة القرابة – ربط الحقل 1 مع جميع الحقول المعنقدة لتشكل عنقود الاخير (البناء الهرمي للعناقيد)

	P1	(P2,P5)(P3,P6),P4)
P1	0	0.292
(P2,P5)(P3,P6),P4)		0

الشجرة الهرمية



شكراً جزيلاً لحسن اصغائكم