التحليل العنقودي Cluster Analysis

إن التحليل العنقودي هو مجموعة من الطرائق الرياضية لاستكشاف الخواص الهيكلية للبيانات الاحصائية لمفردات العينة المسحوبة من المجتمع، وذلك من خلال تصنيفها إلى مجموعات (ضمن عناقيد)، بحيث تكون المفردات داخل كل مجموعة متشابهة مع بعضها (وذلك بالنسبة للمتغيرات أو الصفات المعتمدة لذلك)، وبحيث تكون المجموعات مختلفة عن بعضها البعض، وبعبارة أخرى أن هدف التحليل العنقودي هو تجميع مفردات العينة وتصنيفها ضمن مجموعات متجانسة داخلياً ومتباينة خارجياً بين بعضها البعض .

وخلافاً لما ذكرناه حول مسألة التصنيف في التحليل التمييزي، حيث يكون انتماء كل مشاهدة إلى إحدى المجموعات معروفاً، فإن التحليل العنقودي يهدف إلى التنبؤ بالمجموعة التي سينتمي إليها أي عنصر جديد، ويحاول اكتشاف عدد أو تركيبات هذه المجموعات .

كما يستخدم التحليل العنقودي لتجميع أي مجموعة من المتغيرات X ضمن مجموعات متجانسة ومنفصلة. ويطبق هذا الاتجاه في مراجعة وتنقيح الاستبيانات بناءً على إجابات المستجوبين على مسودات الاستبيان، حيث أن تجميع الأسئلة حسب المتوسطات بواسطة التحليل العنقودي يساعدنا في التعرف على الأسئلة الضعيفة وإعادة النظر فيها . وهذا يزيد حظوظ معدل الإجابات الجيدة بالنسبة لإجمالي أسئلة الاستبيان .

وأخيراً يمكننا أن نعرف التحليل العنقودي بما يلي: هو أحد الأساليب الاحصائية الرياضية لتقسيم عناصر المجتمع المدروس إلى عدة مجموعات متعالقة ومتجانسة داخلياً (متشابهة) ومتباينة خارجياً عن بعضها البعض . أي أنه يهدف إلى جعل تباين العناصر داخل كل مجموعة أصغر ما يمكن، وجعل التباين بين المجموعات (بين مراكزها) أكبر ما يمكن . وبصورة عامة يتفرع التحليل العنقودي إلى نوعين أساسيين هما:

- التحليل العنقودي الهرمي (Hierarchical) .
- التحليل العنقودي غير الهرمي (Non- Hierarchical)

ويعتبر أسلوب التحليل العنقودي الهرمي من الأساليب المفضلة في التحليل العنقودي، لأنه يعتمد على أسس بسيطة، ويعمل على عنقدة مفردات العينة (n مفردة). وبشكل متتالي، ضمن m عنقوداً، بواسطة دمج المفردات المتقاربة ضمن مجموعات متعالقة تسمى عناقيد، وبحيث يكون العنقود الأول C_1 أبسط العناقيد، ويكون العنقود الأخير أعقدها وأشملها (لأنه يضم جميع مفردات العينة)، وبحيث يتألف كل

عنقود من عدة مجموعات متقاربة ومرتبطة مع بعضها بواسطة علاقات تحقق شروط التقارب المفضلة (حسب المتحول أو الصفة المدروسة) .

ويستخدم التحليل العنقودي الهرمي لعنقدة مفردات العينة أسلوبين عمليين هما:

1- أسلوب التجميع The Agglomerative Technique :

ويفترض هذا الأسلوب من البداية أن كل مفردة من مفردات العينة تشكل عنقوداً خاصاً بها، ثم يتم دمج أي مفردتين متقاربتين في عنقود خاص (أول)، ثم نضيف إليهما أي مفردة ثالثة متقاربة مع ذلك العنقود فيتشكل لدينا عنقود ثانٍ، وهكذا نتابع إضافة المفردات واحدة بعد الأخرى إلى بعضها أو إلى العناقيد السابقة، مع تحديد العلاقات بينها ضمن العناقيد، حتى نحصل على العنقود الأخير، الذي يضم جميع مفردات العينة (n مفردة) مع العلاقات التي ترتبط بينها. ويعتمد هذا الأسلوب على مصفوفة التقارب بين مفردات العينة حسب المسافات المحسوبة .

2- أسلوب التجزئة أو التقسيم The Divisive Technique :

ويفترض هذا الأسلوب من البداية أن جميع مفردات العينة (n مفردة) تشكل عنقوداً واحداً شاملاً . ثم تتم تجزئته إلى عناقيد جزئية متباينة تتضمن عدداً أقل من المفردات . وبعد فرز هذه العناقيد وتحديد العلاقات بينها تتم تجزئتها إلى عناقيد أصغر فأصغر، وتتابع هذه العملية حتى يتكون عنقود خاص لكل مفردة من مفردات العينة أو نتوقف عن التقسيم عند حد معين .

وأخيراً نشير إلى أن هذين الأسلوبين يعتمدان على بيانات العينة المدروسة . وعلى طبيعة المتغيرات المستقلة $X_1\ X_2\ X_3\ \dots .\ X_P$ المستخدمة في عملية العنقدة .

فإذا كانت المتغيرات $X_1\ X_2\ \dots .\ X_P$ متغيرات كمية . فإننا نقوم بحساب عناصر المصفوفة D التي تسمى مصفوفة التباعد Dissimitlarity وهي عبارة عن المسافات التي تفصل بين مفردات العينة .

أما إذا كانت متغيرات $X_1\ X_2\ \dots .\ X_P$ نوعية أو مختلطة فإننا نقوم بحساب عناصر مصفوفة أخرى S والتي تسمى بمصفوفة التشابه أو التقارب Similarity، وهي عبارة عن أوزان التكرارات التي تقابل الأزواج المتشابهة $(j,\ k)$، ولهذا فإننا سنقوم بتقديم كيفية حساب عناصر هاتين المصفوفتين حسب المتغيرات المؤثرة على مفردات العينة: أي حسب المتغيرات الكمية أو النوعية أو حسب المتغيرات المختلطة من هذين النوعين .

كما نشير إلى أن بيانات العينة لـ n مفردة أو مشاهدة تُنظم حسب المتغيرات النظامية (المعيارية أو الثنائية) $X_1\ X_2\ X_3\ \dots .\ X_P$ المعتمدة في عملية العنقدة وتُوضع في جدول مناسب كما يلي:

جدول (1) نموذج جدول البيانات اللازمة للتحليل العنقودي :

المتغيرات / رقم المفردة	X_1	X_2	X_3	…..	X_i	….	X_P
1	x_{11}	x_{12}	x_{13}	…	x_{1i}	…	x_{1P}
2	x_{21}	x_{22}	x_{23}	…	x_{2i}	…	x_{2P}
3	x_{31}	x_{32}	x_{33}	…	x_{3i}	…	x_{3P}
j	…..	…..	…..	….	x_{ji}	….	…..
n	x_{n1}	x_{n2}	x_{n3}	…	x_{ni}	…	x_{nP}

2- حساب مصفوفة التباعد D (Dissimilarity) للمتحولات الكمية

تتألف مصفوفة التباعد من قيم المسافات بين أزواج مفردات العينة وتحسب من قيم المتغيرات المستخدمة في عملية العنقدة . لذلك نفترض أنه لدينا n مفردة هي: 1 2 3 … j…n ونريد تصنيفها ضمن عناقيد حسب قيم المتحولات المؤثرة عليها، والتي سنرمز لها بـ $X_1\ X_2\ X_3 \ldots . X_P$ ، وسنستخدم قيم هذه المتغيرات . لحساب عناصر مصفوفة التباعد أو مصفوفة المسافات، والتي سنرمز لها بـ D ونكتبها كمايلي:

$$D = \begin{array}{c c} \text{المفردات} & \begin{array}{ccccc} 1 & 2 & 3 & k & n \end{array} \\ \begin{array}{c} 1 \\ 2 \\ \ldots \\ j \\ \vdots \\ n \end{array} & \begin{bmatrix} d_{11} & d_{12} & d_{13} & \ldots & d_{1n} \\ d_{21} & d_{22} & d_{23} & \ldots & d_{2n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & d_{jk} & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ d_{n1} & d_{n2} & d_{n3} & \ldots & d_{nn} \end{bmatrix} \end{array} \qquad (1 \quad)$$

حيث أن: العنصر d_{jk} هو المسافة بين المفردتين (k , j). وهنا نلاحظ أن هذه المصفوفة هي مصفوفة مربعة ومتناظرة (لأن المسافات: $d_{jk} = d_{kj}$) وان عناصر قطرها الرئيسي تساوي أصفاراً (لأن المسافة بين النقطة j ونفسها $d_{jj} = 0$) . ولهذا فإن معظم المراجع العلمية تكتبها اختصاراً على شكل مصفوفة مثلثية عليا كما يلي:

$$D = \begin{bmatrix} 0 & d_{12} & d_{13} & \ldots & d_{1n} \\ & 0 & d_{23} & \ldots & d_{2n} \\ & & 0 & d_{jk} & d_{jn} \\ & & & 0 & \\ & & & & 0 \end{bmatrix} \qquad (2 \quad)$$

ولكن حساب هذه المسافات يتعلق بطبيعة المتغيرات $X_1\ X_2\ X_3 \ldots . X_P$. فهل هي مستمرة أم منقطعة؟ كما أنها تتعلق بطبيعة المسافة المراد حسابها. فهل هي المسافة النظرية أم المسافة الفعلية ؟

لذلك فإننا نعرف هذه المسافات حسب هذه الحالات، فعندما تكون المتغيرات X كمية (عددية)، فإن المشكلة الوحيدة التي تعترض حساب المسافة بين أي مفردتين هي واحدات القياس لتلك المتغيرات .

فإذا كانت واحدات القياس لجميع المتغيرات X موحدة (كدخل الأسرة أو نفقاتها على الغذاء أو الكساء أو السكن أو النقل والاتصالات ...الخ)، فإننا نقوم بحساب المسافة بين أي مفردتين حسب الصيغ اللاحقة .

أما إذا كانت واحدات القياس لـ X مختلفة (كالدخل وعدد أفراد الأسرة ومساحة السكن ...الخ) فإننا نقوم بتحويل هذه المتغيرات إلى متغيرات معيارية Z وفق العلاقة التالية:

$$Z_i = \frac{X_i - \bar{X}_i}{\sigma_i} \qquad (3)$$

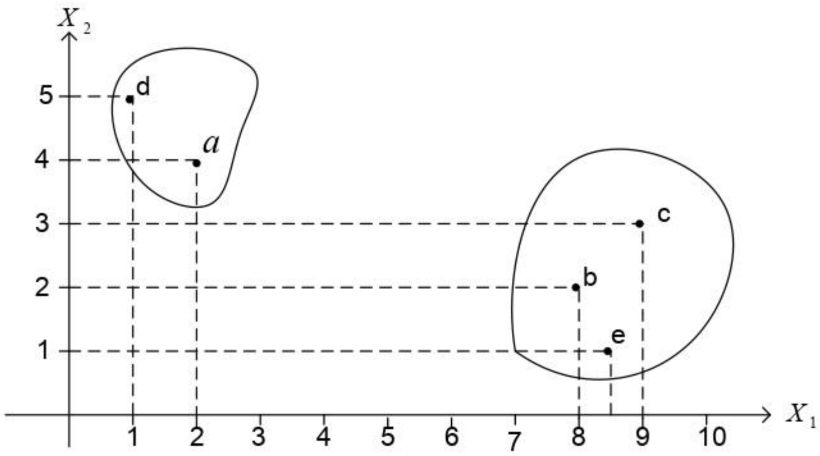
فنحصل على المعيارية: $Z_1\ Z_2\ Z_3\ ....\ Z_P$ ، التي تتميز بأن متوسط كل منها يساوي الصفر $(\bar{Z_i} = 0)$ وأن تباينه يساوي الواحد $\left(\sigma_i^2 = 1\right)$. ثم نتعامل معها كالعادة .

مثال (1): لنفترض أننا نريد تصنيف (5) طلاب ضمن عناقيد حسب متغيرين فقط هما: X_1 نفقات الطالب على الغذاء، وX_2 نفقات الطالب على الاتصالات، وذلك حسب البيانات المبينة في الجدول التالي:

جدول (2): بيانات نفقات (5) طلاب (الف ليرة والأرقام فرضية)

رمز الطالب أو رقمه	X_1 = نفقاته على الغذاء	X_2 = نفقاته على الاتصالات
1 = a	2	4
2 = b	8	2
3 = c	9	3
4 = d	1	5
5 = e	8.5	1

والآن نقوم برسم مواقع هؤلاء الطلاب على المستوى $X_1 0 X_2$ حسب إحداثيات كل منها $(X_1\ X_2)$ فنحصل على الشكل التالي:



الشكل (1): التمثيل البياني لنفقات (5) طلاب

ومن الشكل (1) نلاحظ أن مواقع هؤلاء الطلاب الخمسة تشكل حسب قيم نفقاتهم على الغذاء والاتصالات مجموعتين منفصلتين هما: $G_1(a\ ,d)$ و $G_2(e\ ,b\ ,c)$.

والسؤال الآن هل يمكن تصنيف هؤلاء الطلاب حسب X_1 وX_2، ضمن عناقيد متشابهة داخلياً ومتباينة خارجياً ؟

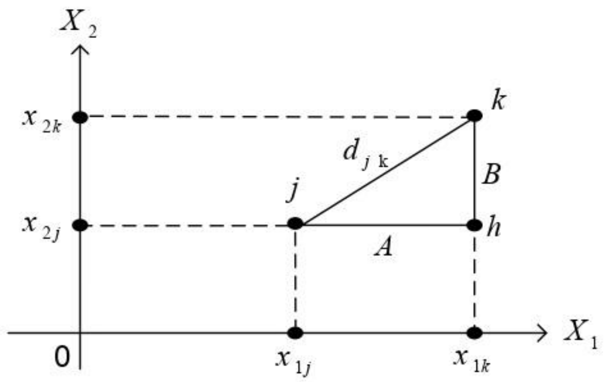
وحتى نجيب على هذا السؤال علينا أن نتبع منهجية العنقدة، والتي تقتضي حساب المسافات المختلفة بين مواقع هؤلاء الطلاب للحصول على مصفوفة التباعد D، ثم تجميع الطلاب حسب الأقرب فالأقرب . ورياضياً تعتبر النقطة j أقرب إلى النقطة k من أية نقطة أخرى ℓ إذا كانت المسافة بينهما d_{jk} تحقق العلاقة:

$$d_{jk} < d_{j\ell} \qquad \ell = 1\ 2 \dots n \qquad \ell \neq k \qquad (4)$$

وإن أهم مقياس للمسافات بين النقطتين j و k في المستوى هو المسافة الاقليدية، والتي تتمثل في الضلع الوتر في المثلث القائم $(j\ h\ k)$، لذلك نحسب مربع ذلك الوتر حسب نظرية (فيثاغورث) وبدلالة إحداثيات النقطتين $k(x_{1k}\ x_{2k})$ و $j(x_{1j}\ x_{2j})$ كمايلي:

$$d_{jk}^2 = A^2 + B^2 = \left(x_{1j} - x_{1k}\right)^2 + \left(x_{2j} - x_{2k}\right)^2 \qquad (5)$$

والشكل التالي يوضح ذلك:



الشكل (2): تمثيل المسافة الاقليدية d_{jk}

ومن العلاقة (5) نجد أن المسافة الاقليدية d_{jk} بين النقطتين j و k تساوي :

$$d_{jk} = +\sqrt{\left(x_{1j} - x_{1k}\right)^2 + \left(x_{2j} - x_{2k}\right)^2} \qquad (6)$$

ويمكن تعميم هذه العلاقة على عدة متغيرات $X_1\ X_2\ X_3 \dots . X_P$ وفي الفضاء R^P، فنجد أن المسافة الاقليدية بين أي نقطتين j و k من الفضاء R^P تعطى بالعلاقة :

$$d_{jk} = \sqrt{\left(x_{1j} - x_{1k}\right)^2 + \left(x_{2j} - x_{2k}\right)^2 + \left(x_{3j} - x_{3k}\right)^2 + \dots + \left(x_{Pj} - x_{Pk}\right)^2} \qquad (7)$$

وهي أقصر مسافة ممكنة بين k و j لذلك تسمى بالمسافة النظرية .

ولكن الحياة العملية لا تعتمد كثيراً على هذه المسافات، فمثلاً لا يمكننا الذهاب من شارع لآخر دون الالتفاف حول بعض المباني التي بينهما، وبناءً على ذلك تم استنباط مقياس آخر للمسافة يسمى مسافة المقاطع (City Block Distance) ويعرف في المستوى بالعلاقة التالية (انظر الشكل 2) :

$$d_{jk}^b = |A| + |B| = \left|\left(x_{1j} - x_{1k}\right)\right| + \left|\left(x_{2j} - x_{2k}\right)\right| \qquad (8)$$

ويمكن تعميم (8) على p متغيرا في الفضاء R^P بالعلاقة التالية:

$$d_{jk}^b = \sum_{i=1}^{P} \left|\left(x_{ij} - x_{ik}\right)\right| \qquad (9)$$

وأخيراً نشير إلى أن حساب المسافات من العلاقات (6) (7) (8) (9) السابقة، يشترط أن تكون المتغيرات $X_1\ X_2\ X_3 \ldots . X_P$ متغيرات معيارية، أو تكون ذات واحدات قياس موحدة .

ونورد فيما يلي جدولاً بأسماء وتعاريف أهم المقاييس المستخدمة لحساب المسافات للمتغيرات العددية الموحدة أو المعيارية . وذلك حسب طبيعة وشروط كل مسألة أو كل قضية بحثية وفي الفضاء R^P .

جدول (3): مقاييس حساب المسافات بين النقطتين (j و k) في R^P

اسم المقياس	الصيغة الرياضية للمقياس	ملاحظات
المسافة الاقليدية Euclidean	$d_e = \left[\sum_{i=1}^{P}\left(x_{ij} - x_{ik}\right)^2\right]^{\frac{1}{2}}$	**جذر مجموع مربعات الفروقات**
مسافة المقاطع City Block	$d_{cb} = \sum_{i=1}^{P}\left\|x_{ij} - x_{ik}\right\|$	**مجموع القيم المطلقة للفروقات**
مسافة (تشيبتشيف) Chebychev	$d_{ch} = Max\left\|x_{ij} - X_{ik}\right\|$	**أكبر القيم المطلقة للفروقات**
مسافة (مينكوفسكي) Minkowski	$d_m = \left[\sum_{i=1}^{P}\left(x_{ij} - x_{ik}\right)^m\right]^{\frac{1}{m}}$	**الجذر الـ (m) لمجموع الفروقات من المرتبة m' – تعميم الاقليدية**
المسافة التربيعية Q_s Mahalanobis	$d_q = \sum\left(X_{ij} - X_{ik}\right)^2 S^{-1}\left(X_j' - X_k\right)$	**متناظرة: S^{-1}**

مثال 2: لنأخذ بيانات المثال (1) ونقوم بحساب عناصر مصفوفة التباعد D بواسطة استخدام المسافة الاقليدية المعرفة بالعلاقة (6) .

ولنفترض أولاً أن كل من الطلاب الخمسة يشكل لوحده عنقوداً خاصاً، ثم نقوم بحساب المسافات الاقليدية بين كل زوج منهم (j , k) من العلاقة:

$$d_{jk} = \sqrt{\left(x_{1j} - x_{1k}\right)^2 + \left(x_{2j} - x_{2k}\right)^2}$$

وهكذا نجد أن المسافة بين الطالبين (1، 2) تساوي:

$$d_{12} = \sqrt{(2-8)^2 + (4-2)^2} = 6.325$$

وكذلك نجد أن المسافة بين الطالبين (1، 3) تساوي:

$$d_{13} = \sqrt{(2-9)^2 + (4-3)^2} = 7.071$$

وبمتابعة حساب هذه المسافات للأزواج المختلفة الأخرى، نحصل على مصفوفة التباعد D لهؤلاء الطلاب التالية:

$$D = \begin{array}{c} \text{الطلاب} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & 6,325 & 7.071 & 1.414 & 7.159 \\ & 0 & 1.414 & 7.616 & 1.116 \\ & & 0 & 8.246 & 2.062 \\ & & & 0 & 8.500 \\ & & & & 0 \end{bmatrix} \end{array}$$

ومنها نلاحظ أن أصغر عناصر هذه المصفوفة هو (1.116) وهو يقابل الطالبين 2 و5، لأنهما الأكثر تشابهاً في النفقات، لذلك يمكن أن ننشأ منهما العنقود الأول كما سنرى لاحقاً .

7-3 حساب مصفوفة التشابه أو التقارب (Similarity) في حالة المتحولات النوعية (المرتبة والاسمية) .

عندما تكون المتحولات X نوعية- مرتبة أو اسمية- فلا يكون لها قيم عددية، بل يكون لها حالات مرتبة أو فئات مختلفة (كمستوى التعليم أو الحالة الاجتماعية). لذلك لا يمكننا حساب المسافات بين المفردات من العلاقات الكمية المذكورة في الجدول (3) .

وفي مثل هذه الحالات نلجأ إلى استخدام التكرارات المطلقة المقابلة لحالات أو فئات تلك المتغيرات . لأن الحالات المتشابهة تكون تكرارتها متقاربة .

فإذا كان المتغير X ثنائياً – أي يتألف من حالتين فقط (نجاح وفشل)- فإننا نفترض أنه يأخذ القيمة (1) إذا تحققت حالة النجاح، ويأخذ القيمة (0) إذا تحققت حالة الفشل .

أما إذا كان المتغير X أكثر من حالتين، فإننا نعتبر كل حالة مستقلة عن الحالات الأخرى، ونعرف عليها متغيرات ثنائية جديدة ($X'_1\ X'_2\ ...\ X'_s$)، ثم نفترض أن كل متغير جديد X'_t يأخذ القيمة (1) عندما تتحقق الحالة المقابلة له، ويأخذ القيمة (0) عندما لا تتحقق تلك الحالة .

أي أنه لحساب الاختلافات في حالة المتغيرات النوعية، يجب أن تكون (أو أن نجعل) تلك المتغيرات متغيرات ثنائية وتأخذ إحدى القيمتين قيمة (1) عند التحقق وقيمة (0) عند عدم التحقق .

وعندها فإن بيانات العينة المأخوذة من n مفردة والمعتمدة على p متغيرا ثنائياً هي $X_1\ X_2\ X_3\ ....X_P$، تنظم وتوضع في جدول كالتالي:

جدول (4): بيانات المتغيرات النوعية (فرضية)

المتغيرات / المفردات	X_1	X_2	X_3	....	X_i	....	X_P
1	1	0	1	....	0	....	1
2	0	1	1	....	0	....	0
3	1	0	0	....	1	....	1
i	....	....	....	....	X_{ik}	....	....
n	1	1	0	....	1	....	0

ونلاحظ من الجدول (4) أن كل مفردة j من العينة تعطينا مقابل كل متغير من المتغيرات $X_1\ X_2\ X_3 \dots X_i \dots X_P$ إحدى القيمتين (1) أو(0) [وهي القيم التي في السطر المقابل للمفردة j] .

كما نلاحظ من جهة أخرى أن كل متغير X_i يأخذ إحدى القيمتين (1) و(0) مقابل مفردات العينة [وهي القيم التي في العمود المقابل لـ X_i] .

ولدراسة التقارب بين أي مفردتين (j وk) نقوم بإيجاد جدول التوافق بينهما، ونحسب التكرارات المتقابلة لتوافقهما ولتعارضهما في السطرين (j وk) من الجدول (4) فنحصل على الجدول التالي :

جدول (5): جدول التوافق للمفردتين (j وk) فقط

البيان		قيم المفردة K		المجموع
		1	0	
قيم المفردة j	1	$a\ (1,1)$	$b\ (1,0)$	$a + b$
	0	$c\ (0,1)$	$d\ (0,0)$	$c + d$
المجموع		$a + c$	$b + d$	$P = a + b + c + d$

إن الرموز في الجدول (5) تعني أن:

a : هو عدد تكرارات الأزواج (1−1) ، b: هو عدد تكرارات الأزواج (0 و1) .

c : هو عدد تكرارات الأزواج (1−0) ، d: هو عدد تكرارات الأزواج (0 و0) .

ولتقدير عناصر مصفوفة التقارب S بين جميع مفردات العينة. يجب علينا أن نقوم بإيجاد جميع جداول التوافق لجميع الأزواج المختلفة، التي يمكن تشكيلها من العينة ذات الحجم n، والتي يبلغ عددها $C_n^2 = \frac{n(n-1)}{2}$ زوجاً مختلفاً . وهناك عدة مقاييس للتقارب بين هذه الأزواج تحسب من القاعدة التالية:

$$S_{jk} = \frac{\text{عدد الأزواج المتشابهة}}{P\ \text{عدد المتغيرات المؤثرة}} = \frac{a+d}{P} \qquad (10\quad)$$

وبعد إيجاد تلك الجداول التوافقية لكل مفردتين (j وk) نقوم بحساب عناصر مصفوفة التقارب S_{jk} في حالة المتحولات الثنائية باستخدام أحد المقاييس المعرفة في الجدول التالي:

جدول (6): مقاييس التقارب للمتغيرات الثنائية

رقم المقياس	المقاييس الرياضية لـ S_{jk}	ملاحظات
1	$\frac{a+d}{P}$	نسبة تكرارات الأزواج المتشابهة (1 و1) و(0 و0) بأوزان متساوية
2	$\frac{2(a+d)}{2(a+d)+b+c}$	بمضاعفة أوزان الأزواج المتشابهة (1 و1) و(0 و0)
3	$\frac{a+d}{a+d+2(b+c)}$	بمضاعفة أوزان الأزواج غير المتشابهة (0 و1) و(1 و0)
4	$\frac{a}{P}$	بحذف تكرار الأزواج (0 و0) من البسط: Russel + Rao
5	$\frac{a}{a+b+c}$	بحذف تكرار الأزواج (0 و0) من البسط والمقام Jaccard (لأنها خارج الموضوع)

رقم المقياس	المقاييس الرياضية لـ S_{jk}	ملاحظات
6	$\frac{2a}{2a+b+c}$	بحذف تكرار الأزواج (0 و 0) من البسط والمقام ومضاعفة تكرارات الأزواج (1 و 1) czekanowcki
7	$\frac{a}{a+2(b+c)}$	بحذف تكرارات الأزواج (0 و 0) من البسط والمقام، ومضاعفة تكرار الأزواج غير المتشابهة
8	$\frac{a}{b+c}$	نسبة الأزواج المتشابهة من الشكل (1 و 1) إلى الأزواج غير المتشابهة (1 و 0) و (0 و 1) مع حذف الأزواج (0 و 0)

ولكن استخدام هذه المقاييس يتعلق بطبيعة المسألة المطروحة وبالهدف الذي يقصده الباحث، علماً بأن هذه المقاييس تعطينا قيماً مختلفة للعناصر غير القطرية S_{jk} في مصفوفة التقارب S، وهذا قد يؤدي بنا إلى الحصول على نتائج مختلفة .

وأخيراً نحصل على عناصر مصفوفة التقارب S ونكتبها كما يلي :

$$S_{P*P} = \begin{array}{c} \text{المفردات} \\ 1 \\ 2 \\ 3 \\ \cdots \\ n \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & \ldots & n \end{array} \\ \begin{bmatrix} S_{11} & S_{12} & S_{13} & \ldots & S_{1n} \\ S_{21} & S_{22} & S_{23} & \ldots & S_{2n} \\ S_{31} & S_{32} & S_{33} & \ldots & S_{3n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ S_{n1} & S_{n2} & S_{n3} & \ldots & S_{nn} \end{bmatrix} \end{array} \qquad (11)$$

حيث أن العناصر S_{jk} تحسب من أحد المقاييس المذكورة في الجدول (6) السابق للمفردتين j و k.

ومن خواص هذه المصفوفة إنها مصفوفة مربعة من المرتبة $n*n$، ومتناظرة لأن المقاييس المستخدمة لحسابها متناظرة (أي أن $S_{jk} = S_{kj}$) كما أن عناصر القطر الرئيسي فيها $(S_{jj} = 1)$ لأن جدول التوافق للمفردة j مع نفسها يكون من الشكل التالي:

j	1	0	1	0	1	0
j	1	0	1	0	1	0

$\Rightarrow$

j	1	0
1	3	0
0	0	3

ومنه نجد أن قيمة العنصر S_{jj} يساوي حسب المقياس (1) ما يلي:

$$S_{jj} = \frac{a+d}{a+b+c+d} = \frac{3+3}{3+0+0+3} = 1$$

وهكذا نجد أن معظم المقاييس الأخرى تعطينا نفس النتيجة (ماعدا المقياسين (4) و (8)).

لذلك يمكننا كتابة مصفوفة التقارب S على شكل مصفوفة مثلثية سفلى (لتمييزها عن مصفوفة التباعد D) كما يلي:

$$\begin{array}{c c} & \text{المفردات} \\ S = & \begin{array}{c} 1 \\ 2 \\ 3 \\ j \\ n \end{array} \begin{bmatrix} 1 & & & & \\ S_{21} & 1 & & & \\ S_{31} & S_{32} & 1 & & \\ & & & & \\ S_{n1} & S_{n2} & S_{n3} & \dots & 1 \end{bmatrix} \end{array} \qquad (12\quad)$$

ملاحظة هامة: إذا كانت المتغيرات $X_1\ X_2\ \dots X_i\dots X_P$ مختلفة (كمية ونوعية) فإننا نقوم بتحويل المتغيرات الكمية فيها (مثل X_1 و X_2) إلى متغيرات ثنائية، وذلك بتقسيم مجال تحول كل منها إلى مجالين فقط . وننشأ منها متحولات ثنائية جديدة معرفة على هذين المجالين كما يلي:

$$X_1' = \begin{cases} 1: & X_1 \le x_{01} \\ 0: & X_1 > x_{01} \end{cases} \qquad X_2' = \begin{cases} 1: & X_2 \le x_{02} \\ 0: & X_2 > x_{02} \end{cases} \qquad (13-7)$$

حيث أن x_{01} و x_{02} هما النقطتان الفاصلتان بين هذين المجالين .

أما المتغيرات النوعية المتبقية فنقوم بتحويلها أيضاً إلى متغيرات ثنائية كما وضحنا ذلك أعلاه . وبعد إجراء هذه التحويلات نقوم بجمع البيانات ووضعها في جدول كالجدول (4) . ثم نقوم بإيجاد جداول التوافق ككل زوج من المفردات (j , k)، فنحصل على $\frac{n(n-1)}{2}$ جدولاً مشابهاً للجدول (5) السابق .

وبعد كل ذلك نقوم بحساب عناصر مصفوفة التقارب S_{jk} وذلك باستخدام أحد المقاييس المعرفة في الجدول (6) السابق .

مثال 3: لنفترض أننا نريد تصنيف (5) طلاب ضمن عناقيد متشابهة وذلك حسب (6) متغيرات مختلطة هي: الوزن X_1 ، الطول X_2 ، لون العينين X_3 ، لون الشعر X_4 ، اليد المستخدمة X_5 ، والجنس X_6 ، وبعد استجوابهم حصلنا منهم على البيانات التالية:

جدول (7): البيانات الأولية لـ (5) طلاب حسب (6) متغيرات

المتحولات / رقم الطالب	الوزن كغ X_1	الطول سم X_2	لون العينين X_3	لون الشعر X_4	اليد المستخدمة X_5	الجنس X_6
1	68	140	أخضر	أشقر	اليمنى	أنثى
2	73	185	بني	أسود	اليمنى	ذكر
3	67	165	أزرق	أشقر	اليمنى	ذكر
4	64	120	بني	أسود	اليمنى	أنثى
5	76	210	بني	أسود	اليسرى	ذكر

للاستفادة من هذه البيانات نقوم بتحويل كل من هذه المتغيرات إلى متحولات ثنائية، ونعرف منها المتغيرات الثنائية التالية: $X_1', X_2', X_3', X_4', X_5'\ , X_6'$ كما يلي:

$$X_1' = \begin{cases} 1: & X_1 \ge 72 \\ 0: & X_1 < 72 \end{cases} \qquad X_4' = \begin{cases} 1: & X_4 = \text{أشقر} \\ \\ 0: & X_4 = \text{غير ذلك} \end{cases}$$

$$X_2' = \begin{cases} 1 : X_2 \geq 150 \\ 0 : X_2 < 150 \end{cases} \qquad X_5' = \begin{cases} 1 : X_5 = \text{اليمنى} \\ 0 : X_5 = \text{اليسرى} \end{cases}$$

$$X_3' = \begin{cases} 1 : X_3 = \text{بني} \\ 0 : X_3 = \text{غير ذلك} \end{cases} \qquad X_6' = \begin{cases} 1 : X_6 = \text{أنثى} \\ 0 : X_6 = \text{ذكر} \end{cases}$$

وبعد تفريغ قيم هذه المتغيرات في جدولٍ كالجدول (4) السابق، نحصل على جدول جديد المتغيرات الثنائية يأخذ الشكل التالي :

جدول (8) نتائج استجواب (5) طلاب بدلالة المتغيرات الثنائية:

المتغيرات / رقم الطالب	X_1'	X_2'	X_3'	X_4'	X_5'	X_6'
1	0	0	0	1	1	1
2	1	1	1	0	1	0
3	0	1	0	1	1	0
4	0	0	1	0	1	1
5	1	1	1	0	0	0

ولإنشاء مصفوفة التقارب S لهؤلاء الطلاب، علينا أولاً أن نقوم بإيجاد جداول التوافق لكل زوج منهم (وعددها $C_5^2 = 10$ أزواج) . وذلك بناء على بيانات الجدول (8) فنجد مثلاً أن جدول التوافق للطالبين (1) و(2) يأخذ الشكل التالي:

		قيم الطالب (2)		المجموع
		1	0	
قيم الطالب (1)	1	1	2	3
	0	3	0	3
المجموع		4	2	6

ومنه نحسب مقياس التقارب بينهما، وذلك من خلال المقياس الأول المعرف في الجدول (6) بالعلاقة التالية:

$$S_{12} = \frac{a+d}{P} = \frac{1+0}{6} = \frac{1}{6}$$

وكذلك نجد أن جدول التوافق للطالبين (1) و(3) يأخذ الشكل التالي:

		قيم الطالب (3)		المجموع
		1	0	
قيم الطالب (1)	1	2	1	3
	0	1	2	3
المجموع		3	3	6

ومنه نحسب عنصر التقارب بين الطالبين (1) و(3) وذلك من خلال نفس المقياس الأول من الجدول (6) المعرف بالعلاقة :

$$S_{13} = \frac{a+d}{P} = \frac{2+2}{6} = \frac{4}{6}$$

وبمتابعة حساب بقية عناصر مصفوفة التقارب S_{jk} بين هؤلاء الطلاب وباستخدام نفس المقياس، نحصل على مصفوفة متناظرة من المرتبة 5 * 5 وتأخذ الشكل التالي:

$$S = \begin{array}{c} \text{الطلاب} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 1 & & & & \\ \frac{1}{6} & 1 & & & \\ \frac{4}{6} & \frac{3}{6} & 1 & & \\ \frac{4}{6} & \frac{3}{6} & \frac{2}{6} & 1 & \\ 0 & \frac{5}{6} & \frac{2}{6} & \frac{2}{6} & 1 \end{bmatrix} \end{array}$$

ومن خلال هذه المصفوفة نلاحظ مباشرة أن أكبر عنصر في المصفوفة S هو العنصر التكراري $\frac{5}{6}$، الذي يقابل الطالبين (2) و(5)، وهذا يعني أن هذين الطالبين أقرب إلى بعضهما (حسب المتغيرات المستخدمة) من أي طالبين آخرين، لذلك يمكننا أن نشكل منهما مجموعة أولى تمثل العنقود الأول .

كما نلاحظ أن أبعد طالبين عن بعضهما هما الطالبين (1) و(5) لأنهما يقابلان أصغر عنصر في المصفوفة هو $(S_{15} = 0)$، وهناك أزواج تقع بين هاتين الحالتين .

وإذا أردنا تقسيم الطلاب إلى مجموعتين جزئيتين متجانستين نسبياً بناء على بيانات مصفوفة التقارب، يمكننا أن نشكل مجموعتين جزئيتين مؤلفتين من هؤلاء الطلاب : $G_1 = (2\ ,5)$ و $G_2 = (1\ ,3\ ,4)$.

ملاحظة هامة: إن عناصر مصفوفة التقارب S_{jk} ترتبط مع عناصر مصفوفة التباعد d_{jk} من خلال العلاقة التالية

$$\tilde{S}_{jk} = \frac{1}{1+d_{jk}} \quad : \quad 0 < \tilde{S}_{jk} \le 1 \qquad (14\ \ )$$

ولكن حساب d_{jk} من S_{jk} يحتاج إلى توفر شرط (Gower) وهو أن تكون المصفوفة S غير سالبة التحديد $(X' * S * X \ge 0)$. وعندما يكون التشابه أعظمياً ومعيراً بـ $(\bar{S}_{ii} = 1)$، فإن المسافة d_{jk} ترتبط مع S_{jk} بالعلاقة التالية :

$$d_{jk} = \sqrt{2(1-\bar{S}_{jk})} \qquad (15\ \ )$$

4 أسلوب (جور Gower) لمعالجة بيانات المتغيرات النوعية أو المختلطة دون تحويلها إلى متغيرات ثنائية:

لنفترض إننا نريد تصنيف عينة من الموظفين إلى عناقيد متشابهة حسب المتغيرات النوعية والكمية المرفقة حسب حالاتها المختلفة وأرقامها الرمزية التالية:

X_1 – المستوى الاقتصادي: منخفض= (1)، متوسط= (2)، مرتفع= (3) .

X_2 – الحالة الاجتماعية: أعزب= (1)، متزوج= (2)، مطلق= (3)، أرمل= (4) .

X_3 – لون العينين: سوداء= (1)، خضراء= (2)، زرقاء= (3) .

X_4 – العمر (بالسنوات): أقل من 30 =(1)، من(30–50) = (2)، أكبر من 50 = (3) .

X_5 – الجنس: ذكر = (1)، أنثى= (2) .

تعالج حالات هذه المتغيرات ونستبدلها بأرقامها وندخلها إلى الجدول الأساسي لبيانات العينة، ولنفترض أن بيانات مفردتين منه مثل (j و k) على هذه الحالات كانت كما يلي:

المتغيرات / رمز الموظف	X_1	X_2	X_3	X_4	X_5	
J	2	2	3	1	2	
K	3	2	3	1	1	
$S_i(jk)$	0	1	1	1	0	$\bar{S}_{jk}$

ثم نعرف على إجابات المفردتين (j و k) متغير جديد $S_{i(jk)}$ ونضعه في السطر الثالث، ونجعله يأخذ القيمة (1) إذا كانت إجابة المفردة j متوافقة مع إجابة المفردة k (مساوية لها)، ويأخذ القيمة (0) إذا كانت إجابة j مختلفة عن إجابة k . فنحصل على السطر الأخير في الجدول السابق .

ثم نقوم بحساب عناصر مصفوفة التقارب S_{jk} من خلال حساب المتوسط الحسابي لقيم المتغير $S_{i(jk)}$ ونرمز له بالرمز التالي:

$$\bar{S}_{jk} = \frac{\sum_{i=1}^{P} S_{i(jk)}}{P} \qquad j,k = 1\ 2\ 3 \dots n \qquad (16\quad)$$

ولكن (جور Gower) اقترح تثقيل هذا المتوسط بأوزان مناسبة مع أهمية المتغيرات $X_1\ X_2\ X_3 \dots X_P$، ورمز لها بالرموز المقابلة لها: $W_1\ W_2\ W_3 \dots W_P$، ثم حساب تقدير عنصر مصفوفة التقارب S_{jk} من العلاقة التالية:

$$\bar{S}_{jk} = \frac{\sum_{i=1}^{P} W_i S_{i(jk)}}{\sum_{i=1}^{P} W_i} \qquad (17\quad)$$

وهكذا نجد أن هذا الأسلوب يعطينا قيماً تقديرية لعناصر مصفوفة التقارب $\tilde{S}_{jk}$ ، وتأخذ قيمها المختلفة بين (0) و(1). وإنها تأخذ القيمة (0) عندما تكون جميع إجابات j مختلفة عن إجابات k (لأنه عندها

تكون جميع $(S_{i(jk)} = 0)$، وبالتالي فإن متوسطها $(\bar{S}_{jk} = 0)$، وتأخذ القيمة (1) عندما تكون جميع إجابات j متوافقة مع إجابات k (لأنه عندها تكون جميع $S_{i(jk)} = 1$، وبالتالي يكون متوسطاتها $(\bar{S}_{jk} = 1$.

وإن قيم $\bar{S}_{jk}$ الأخرى تأخذ قيماً عددية تقع في المجال $[0\,,1]$ أي أن يكون لدينا: $0 \le \bar{S}_{jk} \le 1$.

أي أن أسلوب $(Gawer)$ يحافظ على أن تأخذ عناصر المصفوفة S قيماً كسرية في المجال $[0\,,1]$.

5 تجميع المتغيرات X:

إن التحليل العنقودي يستخدم- بين الحين والآخر- لتجميع المتغيرات X بالاعتماد على المشاهدات. وهذه الحالة مطلوبة في تصميم الاستبيانات، حيث أن مسودة الاستبيان غالباً ما تتضمن بعض الأسئلة التي تحرص على تأمين معدل جيد للإجابات، وعندما يتم اختبار الاستبيان على عدد قليل من المستجوبين، يمكننا مباشرة أن نلاحظ أن الاجابات على مجموعات الأسئلة المتشابهة تكون مرتبطة بشدة، ولكن أهم تطبيق للتحليل العنقودي يمكن أن يتم على مجموعات أخرى من الأسئلة هي الأسئلة المتباعدة أو الغامضة حيث تكون أجوبتها غير مرتبطة، وبذلك تكون هذه الأسئلة نقطة ضعف في الاستبيان . وبعد تحليل الإجابات نقوم بتجميع أسئلة الاستبيان ضمن مجموعات متشابهة وعناقيد متباينة، ثم نقوم بإعادة صياغة الاستبيان . فنقوم بدمج بعض الأسئلة المتشابهة في سؤال واحد . ونعيد صياغة بعضها الآخر ونوضح معانيه ودلالاته .

مثال (4): لتوضيح ذلك نأخذ إجابات (5) أشخاص على (3) أسئلة في مسودة أحد الاستبيانات والتي كانت كما يلي:

جدول (7-9): بيانات المثال فرضية

الأسئلة / المستجوبين	Q_1	Q_2	Q_3
a	10	5.0	3.00
b	30	7.5	3.10
c	20	6.0	2.90
d	40	8.0	2.95
e	50	9.0	3.80

إن أفضل مقياس لتقارب الأسئلة $Q_1\,,Q_2,Q_3$ هو معاملات الارتباط الخطي بين كل زوج منها $Q_i\,,Q_j$ مأخوذة بالقيمة المطلقة، وبعد حسابها نحصل على مصفوفة التقارب التالية:

$$S = \begin{array}{c c} & \begin{array}{ccc} Q_1 & Q_2 & Q_3 \end{array} \\ \begin{array}{c} Q_1 \\ Q_2 \\ Q_3 \end{array} & \begin{bmatrix} 1 & |r_{12}| & |r_{13}| \\ |r_{12}| & 1 & |r_{23}| \\ |r_{13}| & |r_{23}| & 1 \end{bmatrix} \end{array} = \begin{bmatrix} 1 & & \\ 0.984 & 1 & \\ 0.076 & 0.230 & 1 \end{bmatrix} \qquad (18-7)$$

(الأسئلة)

وإن مقياس التباعد (حسب المسافات) بين أي سؤالين Q_i و Q_j يمكن أن يحسب من التحويل التالي :

$$d_{ij} = 1 - |r_{ij}|$$

وهكذا نحصل على المصفوفة الأولى للمسافات بين كل زوج من الأسئلة، والتي تأخذ الشكل التالي :

$$D = \begin{array}{c} \text{الأسئلة} \\ Q_1 \\ Q_2 \\ Q_3 \end{array} \begin{array}{c} \begin{array}{ccc} Q_1 & Q_2 & Q_3 \end{array} \\ \begin{bmatrix} 0 & 0.016 & 0.924 \\ & 0 & 0.770 \\ & & 0 \end{bmatrix} \end{array} \qquad (19 \quad)$$

وبذلك يمكننا أن نطبق على هذه المصفوفة أية طريقة من طرائق العنقدة وباستخدام الإجراءات المعتادة .

فمثلاً نجد أن أصغر قيمة في D هي: $d_{12} = 0.016$ وهذا يعني أن السؤالين Q_1 و Q_2 يشكلان عنقوداً واحداً، لأنهما متشابهان جداً، لذلك يمكن أن نختار أحدهما أو أن نصيغ منهما سؤالاً ثالثاً يعبر عنهما معاً. أما إذا كانت الأسئلة ثنائية (محولة من متحولات نوعية أو كمية) فإنه يمكننا تصنيف بياناتها المؤلفة من (1) و(0) ضمن جداول التوافق . وذلك باستبدال المفردات بالأسئلة Q . وبذلك نحصل مقابل كل زوج من الأسئلة $\left(Q_j , Q_k\right)$ على جدول للتوافق لهما يأخذ الشكل التالي:

جدول (10): جدول التوافق للسؤالين Q_j و Q_k

البيان		السؤال Q_k		
		1	0	**المجموع**
السؤال Q_j	1	a	b	$a + b$
	0	c	d	$c + d$
	المجموع	$a + c$	$b + d$	$n = a + b + c + d$

حيث أن: الرموز a , b , c , d هي عبارة عن أعداد التكرارات المطلقة المتقابلة للأزواج $(1,1)$ و$(0,1)$ و$(1,0)$ و$(0,0)$ على الترتيب، وإن أفضل مقياس للتقارب بين هذين السؤالين $\left(Q_j , Q_k\right)$ هو معامل الاقتران أو ارتباط المتغيرات الثنائية المعرف بالعلاقة التالية:

$$r_{jk} = \frac{a * d - b * c}{(a + b)(c + d)(a + c)(b + d)} \qquad (20 \quad)$$

وبعد حساب هذه المعاملات لجميع الأزواج، نأخذ القيمة المطلقة لها، ونضعها في مصفوفة خاصة للتقارب بين الأسئلة المدروسة ونرمز لها بـ

$$S = \begin{bmatrix} 1 & 0 & 0 \\ |r_{12}| & 1 & 0 \\ |r_{13}| & |r_{23}| & 1 \end{bmatrix} \Rightarrow \qquad (21 \quad)$$

ومنها يمكننا أن نحصل على مصفوفة التباعد D وتطبيق الإجراءات اللازمة عليها .

علماً بأن معاملات الاقتران أو الارتباط r_{ij} المعرفة في (7−20) ترتبط مع المتغير χ^2 بواسطة العلاقة $\left(r = \frac{\chi^2}{n}\right)$، وتستخدم هذه الخاصة عند اختبار الاستقلال لأي متحولين نوعيين . ومن أجل قيمة ثابتة لـ n فإن التشابه الكبير (أو الارتباط القوي) يتوافق مع انعدام الاستقلال .