جامعة محمد خيضر (بسكرة)
كلية العلوم الاقتصادية والتجارية والتسيير
قسم العلوم التجارية

السنة الجامعية: 2025/2024
المستوى: أولى ماستر تسويق
مقياس: الأساليب الكمية في التسويق 2

المحاضرة الأولى: أدوات تقسيم السوق -التحليل العنقودي-

Cluster Analysis (Clustering)

1. تمهيد:

يؤدي التحليل العنقودي Cluster Analysis دورا مهما في الكثير من الميادين مثل: علم النفس، إحصاء، العلوم الاجتماعية، بيولوجيا، ... الخ، وفي التسويق تقوم الشركات بجمع بيانات كثيرة عن العملاء الحاليين والمحتملين، وهنا يساعدها التحليل العنقودي في تقسيم العملاء إلى عدد صغير من المجموعات، بهدف استخدامها في تحليلات إضافية وأنشطة التسويق، فمثلا تصنيف الأشخاص، المناطق والمنتجات إلى مجاميع عنقودية ذات صفات متشابهة، يساعد في تحديد قطاعات السوق والشركات المنافسة.

إذن التحليل العنقودي هي أداة إحصائية تستخدم في مجال الأبحاث التسويقية لوضع المستهلكين الذين يتشابهون في نفس الصفات ضمن مجموعات محددة، مما يسهل على المؤسسة تصميم المنتجات بشكل يتناسب مع كل مجموعة، كما يسهل عليها وضع الخطط التسويقية المناسبة لكل مجموعة.

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

2. تعريف التحليل العنقودي:

التحليل العنقودي هو: "أحد الأساليب الاحصائية الرياضية لتقسيم عناصر المجتمع المدروس إلى عدة مجموعات متعالقة ومتجانسة داخليا (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".

التحليل العنقودي عبارة عن: "إجراءات تهدف إلى تصنيف مجموعة حالات Cases (أو متغيرات Variables) بطرق معينة وترتيبها داخل عناقيد Clusters، بحيث تكون الحالات المصنفة داخل عنقود معين متجانسة فيما يتعلق بخصائص محددة، وتختلف عن حالات أخرى موجودة في عنقود آخر".

3. خصائص التحليل العنقودي:

- أسلوب تكافلي، حيث إنه لا يفرق بين المتغيرات التابعة والمستقلة، وبدلاً من ذلك، يبحث في التشابه بين العناصر، ويفعل ذلك عبر دراسة وفحص العلاقات المتشابكة بين الأفراد والعناقيد تبادليًا.
- **العنقود** Cluster عبارة عن مجموعة من المفردات، الحالات أو الملاحظات المتجانسة نسبيًا، وليس من الضروري أن تتساوى العناقيد من حيث الحجم.
- تتميز العناصر المكونة للعنقود الواحد بأنها متشابهة مع بعضها وفق المتغيرات المستخدمة، بينما تكون عناصر العناقيد المختلفة أقل تجانساً.
- حجم العينة لا علاقة له بالاستدلال الإحصائي، وذلك لأن الهدف ليس تقدير إلى أي مدى النتائج المحصل عليها يمكن تمتد لتشمل المجتمع، وإنما تكون نتائج التحليل العنقودي أصل العينة أي بمتغيرات الدراسة فقط

4. فوائد التحليل العنقودي في تقسيم العملاء:

يقدم التحليل العنقودي العديد من الفوائد في تقسيم السوق (تجزئة العملاء):

أ. تعزيز التسويق المستهدف:

إحدى الفوائد الرئيسية لاستخدام التحليل العنقودي في تجزئة العملاء هي القدرة على تعزيز جهود التسويق المستهدفة. من خلال تجميع العملاء في مجموعات متميزة بناءً على متغيرات ديموغرافية أو سلوكية أو نفسية مختلفة، يمكن للشركات الحصول على فهم أعمق لقاعدة عملائها. يتيح ذلك للمسوقين تصميم رسائلهم التسويقية وعروضهم الترويجية لتناسب شرائح محددة من العملاء.

ب. تحسين تطوير المنتج:

يمكن أن يلعب التحليل العنقودي أيضاً دوراً حاسماً في توجيه استراتيجيات تطوير المنتج، من خلال تحليل بيانات العملاء، يمكن للشركات تحديد مجموعات من العملاء ذوي التفضيلات والاحتياجات المماثلة، ويمكن لهذه الرؤية أن تساعد في إنشاء منتجات جديدة أو تحسين المنتجات الحالية لتلبية احتياجات هذه الشرائح المحددة من العملاء بشكل أفضل.

ج. تحسين الاحتفاظ بالعملاء وولاءهم:

يمكن أن يساهم التحليل العنقودي في تحسين الاحتفاظ بالعملاء وولائهم من خلال تمكين الشركات من تحديد وفهم شرائح العملاء ذات القيمة العالية، من خلال تحليل البيانات المتعلقة بسلوك العملاء، وتاريخ الشراء، ومستويات الرضا، يمكن للشركات تحديد مجموعات من العملاء الذين يتمتعون بالولاء العالي والذين من المرجح أن يقوموا بعمليات شراء متكررة، وتسمح هذه المعرفة للشركات بتنفيذ إستراتيجيات الاحتفاظ المستهدفة، مثل برامج الولاء الشخصية، أو العروض الحصرية، أو خدمة العملاء الاستباقية، لتعزيز الروابط مع هذه القطاعات القيمة.

د. التخصيص الفعال للموارد:

يمكن أن يساعد التحليل العنقودي أيضاً الشركات على تحسين تخصيص مواردها من خلال تحديد القطاعات ذات الربحية المتفاوتة أو إمكانات النمو، ومن خلال النظر في عوامل مثل القيمة الدائمة للعميل، ومساهمة الإيرادات، وإمكانات السوق داخل كل مجموعة، يمكن للشركات تخصيص مواردها بشكل أكثر كفاءة.

5. أهداف التحليل العنقودي

- **اكتشاف** أنماط معينة تنظم المشاهدات (العملاء أو المنتجات مثلا)، وتقسيمها الى مجموعات (عناقيد) تتمتع عناصرها بخواص أو صفات مشتركة؛
- **تصنيف** عينة المشاهدات (أفراد، أشياء ...) الى فئات مختلفة، وذلك بالاعتماد على تشكيلات من فئات المتغيرات (الخصائص التي يتم دراسة المشاهدات وفقا لها)؛
- **التشخيص** Identification عبر تحديد خصائص كل مجموعة من الأفراد (العملاء)، وبالتالي تحديد واختيار الاستراتيجية التسويقية والمزيج التسويق الأنسب لهم؛
- **التنبؤ** بتصرفات أو خواص أفراد أخرين، وذلك بالاعتماد على معرفة الفئات التي ينتمي لها هؤلاء الأفراد، وذلك إذا كانت عناصر تلك الفئات تشترك معا في نفس الخواص.

6. أنواع التحليل العنقودي:

بصورة عامة يتفرع التحليل العنقودي إلى نوعين أساسيين، كل منهما ينقسم إلى طرق فرعية عديدة، هذين النوعين هما:

أ. التحليل العنقودي الهرمي Hierarchical

هو أسلوب يبني تسلسلاً هرمياً للمجموعات عن طريق دمجها أو تقسيمها بشكل متكرر، على أن لا يتم ذلك في مراحل متتالية، وهذه الطريقة لا تتطلب المعرفة المسبقة لعدد العناقيد التي سيتم تجميع/ تقسيم العناصر على أساسها، وهي تناسب العينات الصغيرة نسبيا.

ب. التحليل العنقودي غير الهرمي Non- Hierarchical

في هذه الطريقة يكون عدد العناقيد محدد مسبقا أو معين كجزء من إجراءات التحليل العنقودي، ومن بين أهم أشكالها: "طريقة ك- متوسطات (K-means)، والتي تعتمد في التصنيف على تجميع العناصر المدروسة في k عنقود أولي، ثم حساب مراكز (النقط المتوسطة) هذه العناقيد، والمسافات بين العناصر والمراكز، ثم نقل العناصر إلى العناقيد أقرب إليها.

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 عنقودا، بواسطة دمج المفردات المتقاربة ضمن مجموعات متعالقة تسمى عناقيد، وبحيث يكون العنقود الأول C_1 أصغر وأبسط العناقيد، ويكون العنقود الأخير C_m أعقدها وأشملها (لأنه يضم جميع المفردات والعنقيد الجزئية)، على أن يتألف كل عنقود من عدة مجموعات متقاربة ومرتبطة مع بعضها بواسطة علاقات تحقق شروط التقارب المفضلة (حسب المتغير أو الصفة المدروسة).

7. أساليب التحليل العنقودي الهرمي:

يستخدم التحليل العنقودي الهرمي لعنقدة مفردات العينة أسلوبين عمليين ما:

أ. أسلوب التجميع Agglomerative Technique

يفترض 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.

ب. أسلوب التقسيم Divisive Technique

يفترض هذا الأسلوب من البداية أن جميع مفردات العينة n (مفردة) تشكل عنقوداً واحداً شاملا، 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.

5. متطلبات التحليل العنقودي التجميعي:

إن أسلوبي التجميع والتقسيم يعتمدان على عدد عناصر العينة المدروسة (n مفردة أو مشاهدة)، وعلى طبيعة المتغيرات المستقلة $x_1, x_2, \ldots x_p$ المستخدمة في عملية العنقدة Clustering، على أن تنظم بيانات مفردات العينة حسب المتغيرات النظامية (معيارية أو ثنائية)، وتوضع في جدول مناسب كما يلي:

نموذج جدول البيانات اللازمة للتحليل العنقودي

المتغيرات/المفردات	X_1	X_2	X_3	X_i	X_p
1	X_{11}	X_{12}	X_{13}	$X_{1\,i}$	X_{1p}
2	X_{21}	X_{22}	X_{23}	X_{2i}	X_{2p}
.					
.					
J	X_{j1}	X_{j2}	X_{j3}	X_{ji}	X_p
.					
.					
n	X_{n1}	X_{n2}	X_{n3}	X_{ni}	X_{np}

إذا كانت المتغيرات $x_1, x_2, \ldots x_p$ كمية، فإننا نقوم بحساب عناصر مصفوفة التباعد Dissimitlarity(D) ، وهي عبارة عن المسافات التي تفصل بين مفردات العينة. أما إذا كانت المتغيرات نوعية أو مختلطة، فإننا نقوم بحساب عناصر مصفوفة أخرى مصفوفة التشابه أو التقارب Similarity(S)، وهي أوزان التكرارات التي تقابل المفردتين المتشابهتين (j, k).

5. حساب مصفوفة التباعد (حالة متغيرات كمية):

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 n…j … 3 2 1، نريد تصنيفها ضمن عناقيد حسب قيم المتغيرات المؤثرة عليها، والتي سنرمز لها بـ $x_1, x_2, \ldots x_p$، وسنستخدم قيم هذه المتغيرات لحساب عناصر مصفوفة التباعد أو مصفوفة المسافات، والتي سنرمز لها بـ D ونكتبها كمايلي:

$$D = \begin{array}{c} \text{المفردات} \rightarrow \\ \\ 1 \\ 2 \\ . \\ j \\ . \\ n \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & k & n \end{array} \\ \begin{bmatrix} d_{11} & d_{12} & d_{13} & d_{1k} & d_{1n} \\ d_{21} & d_{22} & d_{23} & d_{2k} & d_{2n} \\ . & & & & \\ d_{j1} & d_{j2} & d_{j3} & d_{jk} & d_{jn} \\ . & & & & \\ n_{n1} & d_{n2} & d_{n3} & d_{nk} & d_{nn} \end{bmatrix} \end{array}$$

المسافة بين المفردة k والمفردة j

حيث أن: العنصر d_{jk} هو المسافة بين المفردتين (k , j) وهنا نلاحظ أن هذه المصفوفة هي مصفوفة مربعة ومتناظرة (لأن: $d_{jk}=d_{kj}$)، وأن عناصر قطرها الرئيسي تساوي أصفارا (لأن المسافة بين النقطة j ونفسها: $d_{jj}=0$)، ولهذا يمكن كتابتها اختصارا على شكل مصفوفة مثلثية عليا كما يلي:

$$D=\begin{array}{c} \\ 1 \\ 2 \\ . \\ j \\ . \\ n \end{array}\begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & k & n \end{array} \\ \begin{bmatrix} 0 & d_{12} & d_{13} & d_{1k} & d_{1n} \\ & 0 & d_{23} & d_{2k} & d_{2n} \\ & & 0 & & \\ & & & 0 & d_{jk} \quad d_{jn} \\ & & & & 0 \\ & & & & 0 \end{bmatrix} \end{array}$$

ولكن حساب هذه المسافات يتعلق بطبيعة المتغيرات، فهل **هي مستمرة أم منقطعة؟** كما أنها تتعلق بطبيعة المسافة المراد حسابيا، **فهل هي المسافة النظرية أم المسافة الفعلية؟** لذلك فإننا نعرف هذ المسافات حسب هذه الحالات، فعندما تكون المتغيرات X كمية (عددية)، فإن المشكلة الوحيدة التي تعترض حساب المسافة بين أي مفردتين هي وحدات القياس لتلك المتغيرات (موحدة أم مختلفة).

فإذا كانت وحدات القياس لجميع المتغيرات X كمية موحدة (كدخل الأسرة أو نفقاتها على الغذاء، اللباس، السكن، النقل الاتصالات... الخ)، فإننا نقوم بحساب المسافة بين أي مفردتين حسب المسافة الإقليدية:

$d_{jk}^2= (x_{1j} - x_{1k})^2 + (x_{2j} - x_{2k})^2$: حالة متغيرين فقط X_1 وX_2 فقط، و j و k هما المفقردتان.

$d_{jk}^2= (x_{1j} - x_{1k})^2 + (x_{2j} - x_{2k})^2 + \dots + (x_{pj} - x_{pk})^2$: حالة p متغير، و j و k هما المفقردتان.

أما إذا كانت وحدات القياس للمتغيرات X كمية مختلفة (كالدخل وعدد أفراد الأسرة ومساحة السكن... الخ)، فإننا نقوم بتحويل هذه المتغيرات إلى متغيرات معيارية Z وفق العلاقة التالية: $z_i= (x_i-\bar{x}_i)/\sigma_i$

فنحصل على المتغيرات المعيارية: $Z_1, Z_2, \dots Z_p$، التي تتميز بأن متوسط كل منها يساوي 0، وتباينها (وانحرافها المعياري) يساوي 1، ثم نتعامل معها كالعادة.

وبعد حساب المسافات بين كل المفردات d_{jk}، وتشكيل مصفوفة التباعد D، نقوم بتجميع المفردات حسب الأقرب فالأقرب، ورياضياً تعتبر النقطة j أقرب إلى النقطة k من أية نقطة أخرى ℓ إذا كانت المسافة بينهما d_{jk} تحقق العلاقة:

$$d_{jk} < d_{j\ell} \qquad \ell = 1 \dots n,\ \ell \neq k$$

8. حساب مصفوفة التشابه أو التقارب (حالة متغيرات نوعية: اسمية أو ترتيبية):

عندما تكون المتغيرات X نوعية ترتيبية أو اسمية كمستوى التعليم أو الحالة الاجتماعية، فإنه لا يمكننا حساب المسافات بين المفردات بطريقة المسافة الإقليدية، لذا نلجأ إلى استخدام التكرارات المطلقة المقابلة لحالات أو فئات تلك المتغيرات، لأن الحالات المتشابهة تكون تكرارتها متقاربة.

- فإذا كان المتغير X ثنائيا، 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 للمتغير X أكثر من حالتين، فإننا نعتبر كل حالة مستقلة عن الحالات الأخرى، ونعرف عليها متغيرات ثنائية جديدة ($X_1', X_2' \dots X_P'$)، نفترض أن كل متغير جديد X_t' ، يأخذ القيمة (1) عندما تتحقق حالة معينة، ويأخذ القيمة (0) عندما لا تتحقق تلك الحالة(أي كل الحالات الأخرى). ولحساب نسبة التقارب S_{jk} بين أي مفردتين (j و k) نقوم بتطبيق الصيغة التالية:

$$S_{jk}= \frac{\text{عدد الأزواج المتشابهة}}{\text{عدد المتغيرات P}} = \frac{a+d}{P}$$

حيث a عدد المتغيرات X'_i التي تحصل عندها المفردتين j وk على نفس القيمة 1، وb هو عدد المتغيرات X'_i التي تحصل عندها المفردتين j وk على نفس القيمة 0 في نفس الوقت.

وأخيرا نحصل على عناصر مصفوفة التقارب S ونكتبها كما يلي:

$$S=\begin{array}{c} \\ 1 \\ 2 \\ . \\ j \\ . \\ n \end{array}\begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & k & n \end{array} \\ \begin{bmatrix} S_{11} & S_{12} & S_{13} & S_{1k} & S_{1n} \\ S_{21} & S_{22} & S_{23} & S_{2k} & S_{2n} \\ . & & & & \\ S_{j1} & S_{j2} & S_{j3} & S_{jk} & S_{jn} \\ . & & & & \\ S_{n1} & S_{n2} & S_{n3} & S_{nk} & d_{nn} \end{bmatrix} \end{array}$$

ومن خواص هذه المصفوفة إنها مصفوفة مربعة من المرتبة n×n، ومتناظرة لان المقاييس المستخدمة لحسابها متناظرة (أي أن: $S_{jk}=S_{kj}$)، كما أن عناصر القطر الرئيسي فيها ($S_{jj}=1$)، لذلك يمكننا كتابة مصفوفة التقارب S على شكل مصفوفة مثلثية سفلى (لتمييزها عن مصفوفة التباعد D) كما يلي:

$$S=\begin{array}{c} \\ 1 \\ 2 \\ . \\ j \\ . \\ n \end{array}\begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & k & n \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ S_{21} & 1 & 0 & 0 & 0 \\ . & & & & \\ S_{j1} & S_{j2} & 1 & 0 & 0 \\ . & & & 1 & 0 \\ S_{n1} & S_{n2} & S_{n3} & \dots & 1 \end{bmatrix} \end{array}$$

ملاحظة (1): كل عناصر مصفوفة التقارب ـ عناصر مصفوفة التباعد D التي تأخذ قيمها في المجال $[0, +\infty[$.

ملاحظة (2): يمكن حساب عناصر مصفوفة التقارب S_{jk} انطلاقا من عناصر مصفوفة التباعد d_{jk} من خلال العلاقة:

$$S_{jk} \quad \frac{1}{1+ d_{jk}} \qquad 0\leq S_{jk}\leq 1$$

ملاحظة (3):

يمكن معالجة المتغيرات النوعية أو المختلطة دون تحويلها إلى متغيرات ثنائية، حيث نستبدلها بأرقامها وندخلها إلى الجدول الأساسي لبيانات العينة، مثلا: النوع X_1 (0 أنثى، 1 ذكر)؛ X_2 الدخل (1 منخفض، 2 متوسط، 3 مرتفع)؛ X_3 الحالة الاجتماعية (1 أعزب، 2 متزوج، 3 مطلق، 4 أرمل)؛ X_4 العمر بالسنوات (1 أقل من 30، 2 لـ [30,50[، 3 لـ 50 فأكبر)؛ ...

ثم نعرف على بيانات المفردتين (j وk) متغير جديد Z، يأخذ القيمة (1) إذا كانت إجابة المفردة j مساوية لإجابة المفردة k، ويأخذ القيمة (0) إذا كانت إجابة j مختلفة عن إجابة k، ثم نقوم بحساب عناصر مصفوفة التقارب s_{jk} من خلال حساب المتوسط الحسابي لقيم المتغير Z بالمنسبة لكل مفردتين i و k بتطبيق العلاقة التالية:

$$s_{jk}= \frac{\sum_i z_{jk}}{p} \qquad i= 1 \dots P \; ; \quad J, k= 1.. n$$

9. خطوات التحليل العنقودي الهرمي التجميعي:

إن عملية العنقدة الهرمية التجميعية لـ n مفردة (مشاهدات أو متغيرات)، تتألف من الخطوات التالية:

1- تبدأ العنقدة الهرمية التجميعية من اعتبار كل مفردة تشكل عنقوداً خاصاً، ومن مصفوفة متناظرة للمسافات (أو للتشابه)، نرمز لها بـ: $D= [d_{jk}]$

2- نبحث في مصفوفة المسافات D عن أصغر عنصر فيها (أصغر مسافة)، ومنه نحدد العنقودين الأكثر قرباً أو تشابها من بين العناقيد المدروسة، وذلك من أجل دمجهما وتشكيل عنقود جديد منهما.

3- ندمج العنقودين u وv في عنقود واحد، ونرمز له بـ: (uv)، ثم نقوم بتحديث العناصر المتقاطعة في مصفوفة المسافات D كما يلي:

أ. نحذف العمودين والسطرين المقابلين للعنقودين u وv من المصفوفة.

ب. نضيف سطرا وعمودا جديدين، ونضعهما في مكان u أو v، أو في آخر أو أول المصفوفة D.

ج. نقوم بحساب شعاع عناصر العنقود الجديد (uv) من عناصر العنقودين u وv، ونرمز لذلك الشعاع بـ: $d_{(uv)}$، وهو عبارة عن المسافات المتقاطعة بين العنقود الجديد (uv) المتبقية في المصفوفة، ثم نضعها في العمود والسطر المخصصين لـ (uv) في المصفوفة، علما بأن عناصر العنقود الجديد (uv) تحسب باستبدال كل عنصرين متقابلين في العنقودين u وv بأصغرهما أو بأكبرهما أو بمتوسطهما، وذلك حسب الربط المستخدم في عملية العنقدة (المنفرد أو التام أو المتوسط على الترتيب).

د. نكرر الخطوات (2) و(3) حتى (n-1) مرة حتى تصبح جميع المفردات في عنقود واحد وفي كل مرحلة نقوم بتسجيل مواصفات وخواص العناقيد المدمجة ومستوياتها الجديدة (عناصر المسافات أو التشابه) في الأماكن المخصصة للدمج.

هـ. نقوم برسم مخطط الأغصان لكل مرحلة من هذه المراحل، وأخيرا ندمجها في المخطط العنقودي العام.

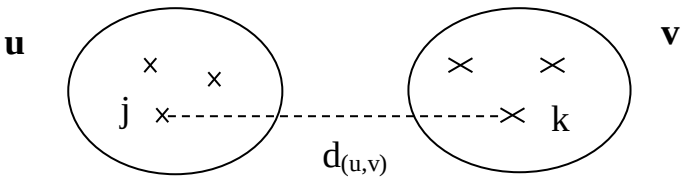
10. طرق الربط في التحليل العنقودي الهرمي:

يستخدم التحميل العنقودي الهرمي التجميعي عدة طرائق تسمى طرائق الربط وهي:

أ. طريقة الربط المنفرد Single Linkage

تبدأ من اعتبار كل مفردة من المفردات تشكل عنقوداً خاصا، وتعتمد على مصفوفة المسافات D (أو التشابه S) بين أزواج المفردات المدروسة، ويتم تشكيل العناقيد فيها من دمج العناقيد الأكثر تقاربا (الجوار الأقرب).

ولتحديد العنقودين الأكثر تقارباً نقوم بدراسة عناصر المصفوفة D، ونحدد أصغر عنصر فيها، ولنفترض أنه كان يقابل العنقودين u وv، لذلك نقوم بدمج هذين العنقودين في عنقود جديد ونرمز له بـ: (u , v)، ولحساب شعاع المسافات للعنقود الجديد (u, v)، نستبدل كل عنصرين متقابلين من (u) و(v) بأصغرهما (بأقربهما)، أي نطبق العلاقة: $d_{(u,v)=} \min[d_{jk}] \quad j\in u, k\in v$؛ حيث j ينتمي إلى u و k ينتمي إلى v.



ولحساب شعاع المسافات بين العنقود الجديد وأي عنقود آخر(أو مفردة w) نطبق العلاقة التالية:

$$d_{(u,v)w} = \min [d_{uw} , d_{vw}]$$

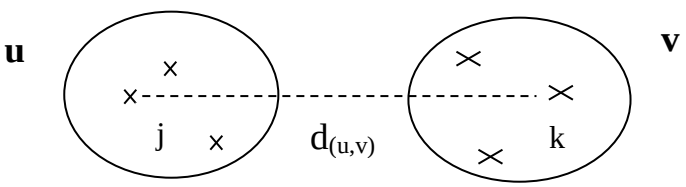
حيث أن: d_{uw} و d_{vw} هما المسافتان بين العنقود الأكثر تقاربا u مع العنقود w، والعنقود الأكثر تقاربا v مع العنقود w على الترتيب.

ب. طريقة الربط التام Complete Linkage

وتبدأ هذه الطريقة أيضاً من اعتبار كل مفردة من المفردات تشكل عنقوداً خاصا، وتعتمد على مصفوفة المسافات D (أو التشابه S) بين أزواج المفردات المدروسة، كما إن عملية تشكيل العناقيد فيها تتم من خلال دمج العناقيد الأكثر تقارباً، وذلك حتى نحافظ على التشابه داخلياً ضمن العنقود الجديد، ولكن عملية حساب عناصر شعاع المسافات للعنقود الجديد، تتم بواسطة استبدال كل عنصرين متقابلين من العنقودين المدمجين بأكبرهما (الجوار الأبعد)، لهذا فإن عملية الحساب تبدأ بالبحث عن أصغر عنصر في المصفوفة D، ثم تحديد العنقودين الأكثر تقارباً، ونرمز لهما بـ u و v، ثم نقوم بدمجهما في عنقود جديد نرمز له بـ (u , v)، ولحساب شعاع المسافات للعنقود الجديد (u , v) نستبدل كل عنصرين متقابلين في و بأكبرهما(بأبعدهما)، أي نطبق العلاقة التالية:

$$d_{(u,v)=} \max[d_{jk}] \quad j\in u, k\in v$$

حيث j ينتمي إلى u و k ينتمي إلى v.



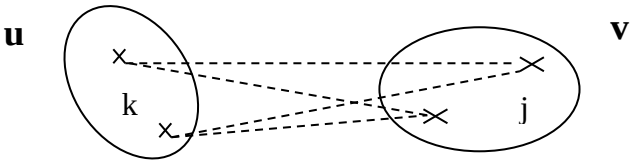
ثم نقوم بحساب شعاع المسافة بين العنقود الجديد وأي عنقود (أو مفردة w) من العلاقة التالية:

$$d_{w\,(u,v)} = \max\,[d_{uw}\,,\, d_{vw}]$$

وعند إجراء الحسابات نتبع نفس الإجراءات، التي أتبعناها في طريقة الربط المنفرد، مع اختلاف وحيد هو طريقة حساب المسافات الجديدة، حيث أن مسافات العنقود الجديد $d_{(u,v)}$ تحسب بأخذ أكبر أي عنصرين متقابلين من عناصر u وv، ويمكن تطبيق العلاقتين السابقتين على المصفوفة المثلثية D، كما يمكن تطبيق أسلوب المقارنة على المصفوفة الأساسية كما فعلنا في حالة الربط المنفرد.

ج. طريقة الربط المتوسط Average Linkage

إن هذه الطريقة تبدأ أيضاً من اعتبار كل مفردة من المفردات تشكل عنقوداً خاصاً، وتعتمد على مصفوفة المسافات D (أو التشابه S) بين أزواج المفردات المدروسة، كما إن عملية تشكيل العناقيد فيها تتم من خلال دمج العناقيد الأكثر تقارباً، ولكن عممية حساب عناصر شعاع المسافات للعنقود الجديد تتم بواسطة أخذ المتوسط الحسابي لكل عنصرين متقابلين من العنقودين المدمجين.



المسافة d_{jk} بين العنقودين k وj هي المتوسط بين المسافات الأربعة المنقطة.

ولتطبيق هذه الطريقة ندرس أولاً عناصر المصفوفة D ونحدد أصغر عنصر فيها، ولنفترض أنه يقابل العنقودين u وv، لذلك علينا أن نقوم بدمج العنقودين u وv في عنقود جديد ونرمز لو بـ (u , v) ولحساب شعاع المسافات المتوسطة لمعنقود الجديد (u , v) نأخذ متوسط كل عنصرين متقابلين في العمودين u وv ونضعهما في عمود وسطر مخصص للعنقود الجديد (u , v)، أي نقوم بحساب المسافات المتوسطة للعنقود (u , v) المسافات المتوسطة للعنقود:

$$d_{(u\,,\,v)\,j} = \tfrac{1}{2}\,[d_{u\,j} + d_{v\,j}]$$

كما يمكن حساب شعاع المسافة بين العنقود الجديد (u , v) وأي عنقود آخر w (أو مفردة) من العلاقة:

$$d_{(u\,,\,v)\,w} = \sum_{j=1}\sum_{k=1} d_{jk} \,/\, [n_{(uv)} * n_{(w)}]$$

حيث أن: $n_{(uv)}$ هو عدد المفردات الموجودة في العنقود (u , v)؛

وأن $n_{(w)}$: هو عدد المفردات الموجودة في العنقود (w)؛

وأن d_{jk}: هي المسافة بين العنصر j من (u , v) والعنصر k من (w).

11. التحليل العنقودي غير الهرمي Non hierarchical

يطبق التحليل العنقودي غير الهرمي على العينات الكبيرة، وهو يصلح لتجميع بيانات المفردات الكثيرة (أكثر مما يصلح للمتغيرات x)، وذلك ضمن k عنقوداً محدداً، وان عدد العناقيد k يمكن تحديده بشكل شخصي، أو من خلال أسلوب العنقدة.

ويعتمد التحليل العنقودي غير الهرمي على أسلوب التقسيم ويفترض ما يلي:

- وجود أو إيجاد تجزئة أولية للمفردات ضمن مجموعات محددة بأي شكل كان.
- اعتبار أن المجموعات الأولية تشكل عناقيد أولية.
- إعادة تقسيم المجموعات الأولية وتشكيل عناقيد أصغر فأصغر حتى مرحلة التوقف.

إن الاختيار الجيد للعنقيد أولية يجب أن يكون مستقلاً عن أي تحيز، وأن أحد أساليب الاختيار هو الأسلوب العشوائي لاختيار مفردات العناقيد الأولية من بين المفردات، أو بإجراء تجزئة عشوائية للمفردات ضمن مجموعات أولية، وتسمى هذه التجزئة بالتجزئة السريعة Quick.

ويستخدم في التحليل العنقودي غير الهرمي عدة طرائق رياضية لفرز المفردات إلى مجموعات أو عناقيد متشابهة، وتستند هذه الطرائق على مبدأ تصغير مجموع مربعات الانحرافات داخل العناقيد المشكلة في كل مرحلة من مراحل التشكيل، وأن أهم هذه الطرائق هي طريقة k- متوسطا (k- means)، التي يمر تطبيقها بالخطوات التالية:

1. تحويل البيانات الموجودة في المتغيرات إلى قيم معيارية، إذا كانت المتغيرات مختلطة (مقاسة بوحدات متنوعة)؛
2. تحديد عدد العناقيد المطلوب أن يجرى التصنيف على أساسها (k عنقود)؛
4. تحديد متوسط قيم العناقيد (Centroids) بشكل مبدئي؛
5. حساب المسافات بين كل مفردة ومراكز المتوسطات؛
6. نقل المفردة إلى العنقود الذي يكون متوسطه أقرب إليها (استخدام مقياس Euclidean Distances)؛
7. إعادة حساب متوسطات قيم العناقيد الجديدة؛
8. تكرار نفس الخطوات (3-6) حتى تتوصل إلى أنه ليس هنالك أية نقطة يمكن تحريكها إلى متوسط قيم العناقيد بشكل أكثر قرباً من الوضع الحالي.

12. سلبيات التحليل العنقودي

رغم أن استعمال التحليل العنقودي أثبت كبيرا في التسويق (تقسيم السوق/ العملاء، تطوير المنتجات، توزيع النشرات التجارية، تحديد مواقع المراكز التجارية ....)، إلا أنه لا يخلو من بعض الصعوبات في التطبيق وعدم الدقة في النتائج، ولعل من أبرزها:

1. التعقيد الحسابي: يمكن أن يكون التجميع الهرمي مكلفًا من الناحية الحسابية، وخاصة بالنسبة لمجموعات البيانات الكبيرة، لأنه يتطلب حساب المسافات بين جميع نقاط البيانات والمجموعات في كل خطوة.

2. الحساسية للضوضاء *Noise Sensitivity*: تنتج ضوضاء البيانات بفعل عدة عوامل (أخطاء القياس، أخطاء إدخال البيانات، قيم مفقودة، قيم شاذة، قيم، متغيرات غير ذات صلة ...)، والتجميع الهرمي حساس للضوضاء، والتي يمكن أن تؤدي إلى تعيين عناقيد (مجموعات) دون المستوى الأمثل إذا لم يتم التعامل مع الضوضاء بشكل صحيح.

3. عدم القدرة على الانعكاس Lack of Reversibility: بمجرد دمج المجموعات في التجميع الهرمي التجميعي، لا يمكن تقسيمها في نفس التحليل. من ناحية أخرى، يسمح التجميع الهرمي الانقسامي بتقسيم المجموعات ولكنه أقل استخدامًا.

4. الذاتية في قطع مخطط الشجرة Subjectivity in Dendrogram Cutting: يمكن أن يكون تحديد العدد الأمثل للعناقيد (المجموعات) عن طريق قطع مخطط الشجرة ذاتيًا، وقد يعتمد على معرفة المجال أو عتبات تعسفية، ويمكن أن يؤدي ذلك إلى تكوين مجموعات مختلفة بناءً على كيفية قطع الشجرة الشجرية.

5. مساحة كبيرة من الذاكرة: عند العمل مع مجموعات بيانات كبيرة، يمكن أن تستهلك التجميعات الهرمية قدرًا كبيرًا من الذاكرة في الحاسوب، بسبب تخزين مصفوفات المسافة والمخططات الشجرية.

6. الافتقار إلى الأمثلية الشاملة Lack of Global Optimality: يعتمد تطبيق برنامج التحيل العنقودي على اختيار مقياس المسافة، معايير الارتباط، ظروف البداية ... الخ، وهو ما يجعله برنامج جشع للكثير من المتطلبات التي قد لا تؤدي دائمًا إلى الحل الأمثل كليا.