التقديراللامعلمي لمصفوفة الاحتمالات الانتقالية باستخدام التحليل العنقودي

أ.د.احسان كاظم القرشي
جامعة القادسية

م. ماجد جاسب عبدالله
هيئة التعليم التقني

الملخص

التحليل العنقودي او العنقدة (Clustering Analysis) يعد واحدا من الاساليب الاحصائية التي تهتم بتقسيم وتصنيف عناصر بيانات المتغيرات الى عدة مجموعات جزئية تكون متجانسة داخل المجموعة الواحدة (العنقود الواحد) وتكون متباينة ومختلفة مع المجاميع الأخرى (العناقيد الأخرى). وتعرّف العمليات التصادفية بأنها عبارة عن مجموعة من المتغيرات العشوائية $\{x_t \ , \ t \geq 0\}$ مؤشرة بدلالة الزمن او اي مؤشر اخر، و سلاسل ماركوف من الانواع الشائعة الاستخدام من العمليات التصادفية.

فكرة البحث تعتمد على الربط بين التحليل العنقودي والعمليات التصادفية (سلاسل ماركوف) لتقدير مصفوفة الاحتمالات الانتقالية بطريقة لامعلمية .الجانب التطبيقي تمثل بالمحاكاة باستخدام برنامج (Matlab-7.10) لتوليد البيانات واجراء التحليل العنقودي وتقدير مصفوفة الاحتمالات الانتقالية.

المقدمة

التحليل العنقودي او العنقدة (Clustering Analysis) يعد واحدا من التحليلات الاحصائية التي تهتم بتقسيم وتصنيف عناصر بيانات المتغيرات الى عدة مجموعات جزئية تكون متجانسة داخل المجموعة الواحدة (العنقود الواحد) وتكون متباينة ومختلفة مع المجاميع الأخرى (العناقيد الأخرى) وبسبب هذه الخاصية التي يمتلكها التحليل العنقودي . اما العمليات التصادفية فتعرّف بأنها عبارة عن مجموعة من المتغيرات العشوائية $\{x_t \ , \ t \geq 0\}$ مؤشرة بدلالة الزمن او اي مؤشر اخر، و سلاسل ماركوف تعد من الانواع الشائعة الاستخدام من العمليات التصادفية.

تعتمد فكرة البحث على الربط بين التحليل العنقودي والعمليات التصادفية لتقدير مصفوفة الاحتمالات الانتقالية بطريقة لامعلمية بالاعتماد على التغير في تصنيف المشاهدات في التحليل العنقودي ، فالمشاهده يمكن ان تصنف من مجموعه معينه(عنقود) لكن بعد مده من الزمن تصنف ضمن مجموعه اخرى(عنقود اخر) نتيجة التغيرات التي تطرأ على الظاهره المدروسه ، وهذا السلوك يكون مشابه لسلوك العمليات التصادفية(سلسلة ماركوف) التي تتغير مع الزمن من خلال عَد العناقيد في التحليل العنقودي كحالات لفضاء الحاله وانتقال المشاهدات بين العناقيد هي انتقالات بين حالات النظام وبذلك يمكن تقدير مصفوفة الاحتمالات الانتقاليه لسلسة ماركوف.

هدف البحث

يهدف البحث الى تقدير مصفوفة الاحتمالات الانتقالية من خلال الربط بين اسلوبين من الاساليب الاحصائية هما:

١ - التحليل العنقودي.
٢ - سلاسل ماركوف.

التحليل العنقودي The Cluster Analysis [3],[4]

إن التحليل العنقودي أو العنقدة احد الاساليب الاحصائية وهي طريقة أنموذجية لتجميع البيانات (العناصر) ضمن مجموعات. فيتم تقسيم مجموعة من البيانات الى عدد من المجاميع الجزئية أو العناقيد بالإعتماد على تشابه العناصر ، حيث تملك العناصر داخل العنقود الواحد درجة عالية من التشابه بينما تملك العناصر المنتمية الى عناقيد أخرى ومختلفة درجة عالية من عدم التشابه ، فمثلاً إذا توفرت لدينا مجموعة من البيانات $[X = (x_1,...., x_n)]$ فإن الهدف هو تجزئة أو تحديد المجاميع الجزئية أو العناقيد من العناصر المتشابهه . فالتحليل العنقودي اسلوب يستخدم التصنيف البيانات في مجاميع فرعية و لا يستند الى العلاقة بين متغيرات معتمدة ومستقلة.

يعرف العنقود بأنه مجموعة من العناصر القريبة أو الأكثر تشابهاً مع بعضها داخل العنقود الواحد والبعيدة او اقل تشابهاً مع عناصر العناقيد الاخرى. علما ان هناك تعاريف اخرى استنادا الى المركز او الجوار او الكثافة. ومن مقاييس التشابه وعدم التشابه (الاختلاف) مايلي:

المسافة الأقليدية Euclidean Distance

يعدللمقياس الأقليدي أكثر المقاييس شيوعاً وإستخداماً لحساب المسافات بين العناصر في فضاء متعدد الأبعاد . ويتم حساب هذه المسافة على النحو الآتي :

$$d_E(x_i, x_j) = \|x_i - x_j\| = \sqrt{\sum_{d=1}^{p} (x_{id} - x_{jd})^2}$$

لكل $i, j \in n$ إذ أن x_j, x_i يمثلان العنصرين ith و jth في البعد (P) .

٢.معامل الارتباط Correlation Confection

$$r_{xy} = \frac{\sum_{d=1}^{p} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{d=1}^{p} (x_i - \bar{x}_j)^2 \sum (y_i - \bar{y})^2}}$$

ويستخدم هذا المقياس لقياس التشابه بين العناصر في فضاء متعدد الأبعاد $\forall\, i, j \in n$

علماً ان مقياس عدم التشابه هو 1-r2 x,y or 1- |rx,y | ، علما ان هناك مقاييس اخرى للتشابه والاختلاف بين العناصر في التحليل العنقودي منها ((المسافة الأقليدية المربعة Squared Euclidean Distance - مسافة مهنلوبس Mahalanobis Distance- مسافة Chebychev (Maximum Distance) - مسافة القوة Power Distance - مسافة مانهاتن (City-Block) Distance Manhattan - مسافات مقياسL-S العامة (Minkowski Distance)))

طرائق التحليل العنقودي [8],[4] Cluster Analysis Methods

طرائق التحليل العنقودي بصورة عامة على نوعين متداخلة وغير متداخلة وكل منها تحتوي على عدة طرائق، فالعنقدة المتداخلة تهدف الى الحصول على الشكل الهرمي للعناقيد المسماة بالمخطط الشجري (Dendrogram) الذي يبين عملية إرتباط العناقيد بعضها مع بعض من خلال سلسلة متداخلة من التجزيئات بإدماج البيانات في العناقيد الصغيرة بصورة متكررة الى عناقيد أكبر (فتسمى طرائق التجميع) أو بفصل عناقيد كبيرة الى عناقيد أصغر (فتسمى طرائق التقسيم) .على الرغم من أن معظم الاساليب المتداخلة تتضمن إرتباط عنقودين أو فصل عنقود الى عنقودين ، فإن بعض الأساليب المتداخلة تربط أكثر من عنقودين في خطوة واحدة أو تفصل عنقود الى أكثر من عنقودين جزئيين

وهناك أسلوبان أساسيان لإنشاء العنقدة المتداخلة هما :

1- طرائق العنقدة المتداخلة التجميعية أو قوانين الربط Linkage Rules (Bottom-Up)- Agglomerative

٢ - طريقة العنقدة المتداخلة التقسيمية Divisive (top-down)

اما طرائق العنقدة غير المتداخلة Nonhierarchical Clustering Methods

والتي تسمى بـ (العنقدة التجزيئية) (Partitional Clustering) فتتم العنقدة غير المتداخلة بتجزئة مجموعة البيانات الى مجاميع جزئية وذلك بتحديد فئة من المجاميع غير المتداخلة التي تحتوي على علاقات غير متداخلة فيما بينها.ولاتملك طرائق العنقدة غير المتداخلة أشكال تشبه الشجرة (Tree) حيث تتشكل العناقيد بإتباع إسلوب العنقدة التجزيئية (Partitional Clustering) للحصول على تجزئة مفردة لمجموعة من العناصر الى عناقيد بالإعتماد على التوافق بين عناصر البيانات وعملية تجزئتها .

تعتمد طرائق التجزئة التي تستخدم مربع الخطأ (squared error) الى تحديد عدد العناقيد (K) وتمثيل كل عنقود بنموذج أولي (Prototype) ومن ثم العمل على تصغير دالة الهدف التي تمثل مجموع كل عناصر البيانات لمربع المسافة بين العنصر والنموذج الأولي للعنقود المحدد لها .وغالباً ما تمثل النماذج الأولية هذه مراكز العناقيد (cluster centers) كما في (K-Means) ، و(K-Medoid), .

ومن طرائق العنقدة غير المتداخلة طريقة K من المتوسطات وطريقة النماذج الاولية -K وطريقة التوزيعات المختلطة وطريقة كثافة التقديروقد تم استخدام الطريقة الاولى في الجانب التطبيقي لذلك سنتطرق لها بشيء من التفصيل .

طريقة عنقدة K من المتوسطات K-Means Clustering Method

حيث تبدأ الطريقة مع K من العناقيد العشوائية ومن ثم تحريك (نقل) العناصر بين هذه العناقيد مع العمل على تصغير الإختلافات داخل العناقيد وتعظيم الإختلافات بين العناقيد ، بتحديد عدد ثابت من العناقيد وإختيار مراكز العناقيد الأولية عشوائياً من ثم يتم تعيين كل البيانات الى أقرب مركز عنقودي بدرجة عضوية مساوية الى (١) وعضوية (٠) الى العناقيد الأخرى ، بعد ذلك يتم إختيار مقياس المسافة الملائم . وتُعد هذه الطريقه من الطرائق اللامعلميه كثيرة الاستعمال ، و يمكن تلخيص خطوات خوارزمية هذه الطريقة على النحو الآتي :

بفرض أن $X = (x_{1p}, x_{2p}, ..., x_{np})$ تمثل مصفوفة مجموعة من البيانات ذات ابعاد(n*p)التي يتم تصنيفها الى k من العناقيد وتمثل هذه العناقيد موجهات (nx1)ممكن إختيارها عشوائياً من عناصر البيانات الكلية أو تمثل أول العينات k أو تمثل العينات المستندة على بعض المعلومات السابقة . ويتم تثبيت معيار التوقف ∈ (termination tolerance)لطريقة العنقدة مثلاً ٠.٠١ و ٠.٠٠١ وهكذا .

توليد مصفوفة التجزئة $U = \lfloor u_{ij} \rfloor$ لكل $(j = 1,...., n \ , i = 1,2,....., k)$ توليداً قياسياً منتظماً .

إختيار مراكز العناقيد $(C_1, C_2, ..., C_k)$ حيث :

يحدد كل عنقود بواسطة المركز المتوسط (Centroid) .

تكون عدد المراكز مساوية الى عدد العناقيد النهائية .

تعيين كل عناصر البيانات الى أقرب مركز متوسط (أقرب جوار)بإستخدام مقياس المسافة الملائم ، مثلاً في حالة إستخدام المسافة الأقليدية فإن عناصر xj للعنقود ci تحدد على أساس أن :

$$|x_j - c_r| < |x_j - c_i| \qquad for\ all\ i, r = 1,2,....,k$$

$$j = 1,2,...,n \ ; r \neq i$$

اعادة حساب المراكز لكل عنقود ، حيث يتم تحديد مراكز العناقيد الجديدة بحساب معدل عناصرالبيانات في كل عنقود.

إعادة تحديد أعضاء العنقود $U = \lfloor u_{ij} \rfloor$ لتصغير مربع الخطأ بين عناصر البيانات ومراكز العناقيد الحالية .

إعادة تكرار الخطوات السابقة حتى الوصول الى التقارب المتمثل بحالة عدم التغير في مراكز العناقيد بمعنى إذا تحقق : $C_i^{(l+1)} = C_i^{(l)} \quad ; \forall i = 1,....., k$

أو أن الخوارزمية تتوقف في حالة أن التغيرات في مصفوفة التجزئة تكون قليلة جداً أي عندما :

$$\prod_{j=1}^{n} \max \left\| U^{(l)} - U^{(l-1)} \right\| < \in$$

حيث أن l تمثل عدد التكرارات ويمثل ∈ معيار التوقف

مقياس صحة العنقدة الظلي Silhouette Validity Measure [3],[4]

هناك عدة مقاييس لصحة العنقدة تعتمد اساليب مختلفة منها (اجراء التحليل ياستخدام مقياس مسافة مختلف او باستخدام طريقة عنقدة مختلفة او تجزئة البيانات عشوائيا الى نصفين وتحليل كل نصف كحالة منفصلة).

لكن سوف يتم اختيار مقياس صحة العنقدة الظلي لتوفره في البرنامج الذي تم استعماله في التطبيق فهو يحسب العرض الظُلي (Silhouette Width)لكل عنصر، ومعدل العرض الظُلي لكل عنقود والمعدل الكلي للعرض الظُلي لمجاميع البيانات . يستخدم هذا المقياس لحساب صحة أو قانونية العنقدة ولمقارنة العدد الأمثل من العناقيد المختارة من خلال تمثيل كل عنقود بما يسمى بالظُلّية التي تعتمد على مقارنة إدماج وفصل العناقيد

والذي يمثل مؤشر أو دليل ثقة لعضوية العنصر jth في العنقود المحدد S_i .

يمكن حساب مقياس العرض الظُلي $SR(j)$ للعنصر jth في العنقود ith حسب الصيغة الآتية :

$$SR(j) = \frac{b(j) - a(i)}{\max\ [a(j), b(j)]}$$

حيث أن $a(j)$ تمثل معدل المسافة بين العنصر jth وجميع العناصر الاخرى في العنقود s_i، وتمثل $b(j)$ أصغر معدل مسافة بين العنصر jth وكل العناصر الأخرى المعنقدة في $(i \neq r\ ;\ r = 1,....,\ k)s_r$.

وأن قيمة $SR\ (j)$ تكون بين ١+ و ١- ؛ $(-1 \le S(j) \le +1)$ ففي حالة إقتراب قيمة $S(j)$ من (١+)فيستنتج من هذا أن العناصر تكون معنقدة جيداً وأنها معينة الى العنقود الملائم . وعندما تقترب قيمة $SR\ (j)$ من (٠) فيعني هذا أن العنصر يقع الى حد ما بين عنقودين ومن الممكن تعيينه الى أقرب عنقود . أما في حالة إقتراب قيمة $SR\ (j)$ من (١-) فيعني هذا أن العنصر غير مصنف جيداً(Misclassified) . وتدل القيم الكبيرة لمقياس $SR(j)$ على العنقدة الجيدة وعلى وجود عدد مثالي من العناقيد .

العمليات التصادفي Stochastic Processes [2],[5]

تعرف العملية التصادفية بأنها عبارة عن ظاهرة تتغير بتغير دليل معين (كالزمن أو السُمك او أي دليل اخر) والتي لايمكن إيجاد قيمها النظرية بدقة تامة، بل لها مدى معين من القيم الممكنة والمتعلقة بتوزيع احتمالي يصف قيمتها عند كل قيمة من الدليل. وتعرّف كذلك بأنها عبارة عن مجموعة من المتغيرات العشوائية $\{x_t\ ,\ t \ge 0\}$ مؤشرة بدلالة الزمن (او اي دليل اخر)، كما يمكن تعريفها بأنها عبارة عن متسلسلات من المتغيرات العشوائية تولدت بواسطة القوانين الاحتمالية . يرمز للعملية التصادفية بالرمز (xn) ، حيث (n) تشير إلى الزمن المتقطع (n = 0 , 1 , ...) ، أو يرمز لها بـ (xt)عندما يكون الزمن مستمراً $(t \ge 0)$.

وعليه فإن العملية (xt) تسمى العملية التصادفية والقيم التي تفترض بواسطة العملية تسمى الحالات (States) ، ومجموعة القيم الممكنة تسمى فضاء الحالة (State Space) ، وإن مجموعة القيم الممكنة للمعلمة (T) المؤشرة (Index Parameter) تسمى فضاء المعلمة (Parameter Space) والتي يمكن أن تكون مستمرة أو متقطعة ، وتدل المعلمة (T) عادة على الزمن .وابسط مثال هو عدد الحوادث المرورية والذي يعد متغيرعشوائي لكن عند قياسه بدلالة دليل مثل الزمن او طول الطريق عند ذلك يصبح لدينا عملية تصادفية تمثل عدد الحوادث بدلالة الزمن او طول الطريق.

عمليات ماركوف (Markov Processes)

يقال للعملية التصادفية ذات المعلمة المتقطعة {Xt; t=0,+1,+2,...} أو ذات المعلمة المستمرة {X(t), t> 0} بأنها عملية ماركوف إذا حققت الخاصية الآتية :

$$P\{X_{(t+1)} = j \mid X_{(t)} = i, X_{(t-1)} = i_{t-1}, X_{(t-2)} = i_{t-2}, ..., X_{(t_o)} = i_o\} = P\{X_{(t+1)} = j \mid X_{(t)} = i\}$$

عند دراستنا لأية ظاهرة أو نظامفأننا نسعى غالباً إلى استنتاج ومعرفة حالة النظام عند نقطة زمنية معطاة، ولتكن t1، عن طريق معرفتنا بحالة هذا النظام عند أية نقطة زمنية مبكرة سابقة لها to ولا يُعتمد على جميع تاريخ النظام (System History) قبل النقطة الزمنية to. وبعبارة أخرى فان حالة النظام عند النقطة الزمنية الحالية t1 تكون مستقلة (Independent) عن جميع حالات النظام الأخرى قبل النقطة الزمنية to، بل تعتمد فقط على حالة النظام السابقة لها، أي عند النقطة الزمنية to. فالعمليات التصادفية التي تحقق هذا الشرط تسمى بعمليات ماركوف (Markov Processes).

أي أن التوزيع الشرطي (Conditional Distribution) للمتغير X(t+1) معطى جميع قيمه الماضية الحاضرة تعتمد فقط على القيمة الحالية منه X(t) ولا تعتمد على أية قيمة أخرى من الماضي. وتعرف هذه الخاصية بخاصية ماركوف (Markov Property). وكما هو واضح فان هذه الخاصية تتضمن السببية

الاحتمالية (Probabilistic Causality)وتصنّف عمليات ماركوف تبعاً لفضاء الحالة (متقطع-مستمر) وفضاء المعلمة (متقطع-مستمر) .

سلسلة ماركوف (Markov Chain)

تسمى عملية ماركوف ذات فضاء الحالة المتقطع سلسلة ماركوف (Markov Chain) بغض النظر عن وصف فضاء المعلمة إنكان متقطعاً أم مستمراً، وغالبا ما تستخدم مجموعة الأعداد الصحيحة S={0,1,2,…} كفضاء حالة لتمثيل سلاسل ماركوف. وقد عرّفت سلسلة ماركوف بأنها عملية عشوائية متقطعة الزمن يتميز كل متغير عشوائي (Xt) فيها بارتباطه بالمتغير السابق له مباشرة (Xt-1) وبتأثيره على المتغير اللاحق (Xt+1) فقط، ومن هنا اطلق على هذه العملية اسم "سلسلة" لتعلق كل متغير بجواره المباشر فقط.

ويرمز لسلسلة ماركوف اختصاراً (MC)، وهي عبارة عن سلسلة من المتغيرات العشوائية (Sequence of Random Variables) {Xt; t=0,1,2,…} تحقق خاصية ماركوف وان الانظمة التي تمتلك تلك الخاصية تسمى بسلاسل ماركوف.

الاحتمالات الانتقالية (Transition Probabilities)

من المعروف أن سلسلة ماركوف (MC)بنيت أساساً على انتقال الظاهرة من حالة إلى أخرى ضمن فضاء الحالة تحكمها في ذلك قوانين احتمالية ،وتسمى الاحتمالات التي تصف الانتقال لسلسلة ماركوف من حالة إلى أخرى خلال فترة زمنية معينة، بالاحتمالات الانتقالية (Transition Probabilities)، ويرمز للاحتمال الانتقال من الحالة i عند أية لحظة زمنية t بغض النظر عن الحالة السابقة للحالة i (التاريخ المسبق) إلى الحالة jعند الزمن t+1 (بعد خطوة واحدة) بالرمز Pij المتمثل بالصيغة الاتية :

$$Pij=P\{Xt+1=j|Xt=i\}$$

ولجميع قيم i,j∈I حيث أن I هي مجموعة الأعداد الصحيحة (Integer Numbers). وهذا الوصف يلائم فضاء معلمة سلسلة ماركوف المتقطع الذي يمكن تمثيله بالأعداد الصحيحة {0,1,2,…}، وفي حالة كون فضاء المعلمة لسلسلة ماركوف مستمراً يُصبح الانتقال بين الحالات عبر الكثافات الانتقالية (Transition Densities)بدلاً من الاحتمالات الانتقالية.وبناءً على ذلك توضع الاحتمالات الانتقالية في مصفوفة تربيعية أبعادها n×n تسمى بالمصفوفة الاحتمالات الانتقالية، ويشترط في هذه المصفوفة تحقيق الشرطين الآتيين:

إن كل عنصر من عناصر هذه المصفوفة يجب أن لا يكون سالباً، أي Pij >0.

$$\sum_{j=1}^{n} p_{ij} = 1$$

إن مجموع عناصر كل صف فيها يجب أن يساوي الواحد الصحيح، أي .

والتمثيل الآتي هو لمصفوفة الاحتمالات الانتقاليه (Transfer Matrix) أبعادها n×n.

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1j} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2j} & \cdots & p_{2n} \\ \vdots & & & & & \\ p_{n1} & p_{n2} & \cdots & p_{nj} & \cdots & p_{nn} \end{bmatrix}$$

وكل عنصر من عناصر (Pij) يُعبر عن احتمال انتقال الظاهرة من الحالة i إلى الحالة j خلال وحدة زمنية واحدة (خطوة واحدة). ويمكن تعميم هذه الحالة، فعند إيجاد انتقال الظاهرة من الحالة i إلى الحالة j بعدد من الخطوات أو الوحدات الزمنية، لتكن k، فان رمز مصفوفة الانتقال هو P(k) ، ويمكن البرهنة على أن P(k)=Pk. وتعّد مصفوفة الانتقال إحدى المكونات الرئيسة لانموذج ماركوف.

أن الاحتمالات الانتقالية تمثل معلمات لذا يستوجب تقديرها ، فتستخدم طريقة الإمكان الأعظم لتقدير الاحتمالات الانتقالية(pij) للسلسلة الزمنية الجزئية (القصيرة) (Micro Data) حيث يكون انتقال الظاهرة لهذا النوع من السلاسل معلوم (ملاحظ) أو يمكن القول هي نوع من البيانات التي يمكن معرفة كيفية الانتقال (الانتقال المعلوم) من الحالة (i) في الوقت (t) إلى الحالة (j) في الوقت (t+1) ، أما عندما يكون الانتقال من حالة إلى أخرى في السلسلة الزمنية مجهول يتم استخدام طرق أخرى للتقدير مثل طريقة المربعات الصغرى الاعتيادية (OLS) . او. طريقة المربعات الصغرى العامة (GLS) .

وعلى افتراض أن هناك عينة مكونة من المشاهدات على شكل سلاسل ماركوف في حالة الثبوتية (ergodic) ، وعلى فرض أن العدد nI(0) يمثل العناصر المشاهدة في الحالة (i)عند الوقت (t = 0) ، حيث تشير العناصر المشاهدة إلى سلسلة من الحالات عند الوقت (t = 0 , 1 , 2 , ... , T) ، وعليه فإن عملية ماركوف في حالة الاستقرارية تعطى بالشكل الآتي :

$$p_r(x_0, x_1, ..., x_T) = p_r(x_0) \prod_t p_r(x_t / x_{t-1})$$

ليكن nij(t) تمثل عدد العناصر المشاهدة لكل (xt = sj , xt-1 = si) ، وإن :

$$n_{ij} = \sum_t n_{ij}(t)$$

$$\hat{p}_{ij} = \frac{n_{ij}}{\sum_j n_{ij}} \geq 0$$

إذ $\hat{p}_{ij}$ يمثل تقدير الإمكان الأعظم عندما تكون الاحتمالات الانتقالية مستقرة ، أما عندما تكون الاحتمالات الانتقالية غير مستقرة فنتبع الصيغة التالية:

$$\hat{p}_{ij}(t) = \frac{\sum_{t=1}^{T} n_{ij}(t)}{\sum_t \sum_j n_{ij}(t)}$$

الجانب التجريبي (المحاكاة) :[7] [6] [1]

تعرف المحاكاة بانها عملية تمثيل او تقليد للواقع الحقيقي باستخدام نماذج معينة، وكثيرا ما نجد في الواقع الحقيقي ان هناك عمليات تكون معقدة الفهم والتحليل لذلك فمن الافضل ان نوصف هذه العمليات بصورة مشابهة للصور الحقيقية بنماذج معينة، ففهم النموذج يحقق لنا قدرا من الادراك للعملية الاصلية او الواقع الحقيقي من خلال محاكاة النموذج، لقد تعددت اساليب المحاكاة ولاسيما بعد التطور السريع الذي حصل في استخدام الحاسبة الالكترونية ولكونها الاسلوب الفعال الذي يمكننا من ادارته بشكل تطبيقي واسع ذي مديات تفوق الامكانات المعقولة في التطبيق العملي .

ان اول مراحل استخدام اسلوب المحاكاة هو توليد المتغيرات العشوائية قيد الدراسة، كما ان أي تجربة محاكاة ماهي الى عبارة عن نوع معين من انواع المعاينة اذ تسحب هذه العينة من المجتمع الافتراضي الممثل للظاهرة المدروسة باحجام مختلفة من العينات وبتكراره بعدد كبير من المرات بدلا من ان تسحب من المجتمع الحقيقي. وبذلك فأن اسلوب المحاكاة يمكن ان يحقق للباحثين حلولا تحليلية وكذلك يؤمن قاعدة تجريبية تكون دليلا لهم مع القاعدة النظرية لاختيار الاسلوب الملائم او الطريقة الملائمة لتحليل ودراسة بيانات الظواهر التي يدرسونها من خلال مطابقة خصائصها مع الانواع التي طبقت المحاكاة عليها.

وصف تجربة المحاكاة الخاصة بالبحث:

١- برنامج المحاكاة كتب باستخدام البرنامج (Matlab-7.10) لتوليد المتغيرات العشوائية باستخدام الدالة (normrnd) والتي تولد متعدد متغيرات طبيعي بمتوسط(٥٠ 50) ومصفوفة تباين ($\begin{matrix} 10 & 2 \\ 2 & 10 \end{matrix}$) لتمثل بيانات ظاهرة معينة ولمدتين زمنيتين.

٢- اجراء التحليل العنقودي وتصنيف بيانات المدتين وتم اختيار التصنيف بثلاث عناقيد.

٣- بناء مصفوفة الانتقالات(Fij) بابعاد(٣*٣) لتمثل الانتقالات بين حالات التصنيف (الثلاث) خلال المدتين .

٤- بناء مصفوفة الاحتمالات الانتقالية(Pij).

بناء أنموذج ماركوف يتم اولاً بايجاد عدد تكرارات الانتقال بين حالات العملية من خلال حساب عدد الانتقالات من الحالة (i) الى الحالة (j) في خطوة واحدة اذ يتم بناء مصفوفة الانتقالات ذات الخطوة الواحدة :

$$F_{ij} = \begin{pmatrix} F_{11} & F_{12} & F_{13} \\ F_{21} & F_{22} & F_{23} \\ F_{31} & F_{32} & F_{33} \end{pmatrix}$$

ومن مصفوفة الانتقالات F يمكن تقدير مصفوفة الاحتمالات الانتقالية P ولكل صف بقسمة كل عنصر في الصف على مجموع الصف :

$$P_{ij} = \begin{cases} F_{ij} / \sum_{i=1}^{m} F_{ij} \, if \sum_{i=1}^{m} F_{ij} \rangle 0 \\ 0 \, if \sum_{i=1}^{m} F_{ij} = 0 \end{cases}$$

$$P_{ij} = \begin{pmatrix} P_{11} & P_{12} & P_{13} \\ P_{21} & P_{22} & P_{23} \\ P_{31} & P_{32} & P_{33} \end{pmatrix}$$

فتصبح لدينا مصفوفة الاحتمالات الانتقالية التالية:

نتائج البرنامج:

البيانات التي تم توليدها باستخدام المحاكاة كانت عبارة عن متجه ذو ابعاد (١٥*٢) تمثل متعدد متغيرات طبيعي، وكان تصنيف التحليل العنقودي بأختيار ثلاث عناقيد لهذه البيانات كما يلي:

2 2 2 2 2 2 3 3 3 3 3 1 1 1 1

1 1 1 1 1 2 2 2 2 2 3 3 3 3 3

وكانت معدلات مقياس صحة العنقدة الظلي للعناقيد كالاتي: SR(1)=0.62, SR(2)=0.43 , SR(3)=0.48

ورسم البياني المرفق يبين مقياس صحة العنقدةالظلي.مما يدل على صحة عملية العنقدة بصورة مقبولة.

فتكون مصفوفة الانتقالات كما يلي:

$$F = \begin{bmatrix} 0 & 0 & 4 \\ 5 & 1 & 0 \\ 0 & 4 & 1 \end{bmatrix}$$

ومنها يتم بناء مصفوفة الاحتمالات الانتقالية فتكون كما يلي :

$$P = \begin{bmatrix} 0 & 0 & 1 \\ 0.83 & 0.17 & 0 \\ 0. & 0.8 & 0.2 \end{bmatrix}$$

الاستنتاجات

١- امكانية استعمال وتوظيف طرائق التحليل العنقودي في تقدير الاحتمالات الانتقالية لسلاسل ماركوف.

٢-امكانية الربط بين اساليب تحليل متعدد المتغيرات واساليب العمليات التصادفية في التقدير.

٣- المحاكاة اسلوب يوفر بيانات تجريبية للباحثين بتوزيعات ومعلمات مختلفة تمكهنم من تصويرالواقع بشكل تقريبي.

التوصيات

١- استعمال طرائق اخرى من طرائق التحليل العنقودي في تقدير الاحتمالات الانتقالية لسلاسل ماركوف .

٢- استعمال اساليب اخرى لمتعدد المتغيرات في تقدير الاحتمالات الانتقالية لسلاسل ماركوف او انواع اخرى من العمليات التصادفية .

٣- المساهمة بنشر الوعي والادراك للتحليل التصادفي من خلال استعمال العمليات التصادفية في الجوانب التطبيقية نظرا لقلة استعمالها قياسا لمواضيع الاحصاء الاخرى.

المصادر

١- بقجه جي صباح الدين واخرون -١٩٩٨- بحوث العمليات - المركز العربي للتعريب والترجمة والتأليف والنشر - دمشق- سوريا.

٢ - الربيعي ، فاضل محسن وعبد ، صلاح حمزة -٢٠٠٠- مقدمة في العمليات التصادفية - مديرية دار الكتب للطباعة والنشر ، بغداد .

٣- الشيخ، وفاء سيد حسنين-٢٠٠٦-إستخدام التحليل العنقودي في عمليات التصنيف مع تطبيق عملي- اطروحة دكتوراه في الاحصاء- كلية الادارة والاقتصاد - الجامعة المستنصرية.

٤ -Everitt, B.S. and Dunn, G.- 2001-Applied multivariate data analysis- Oxford - university press Inc., New York-Usa .

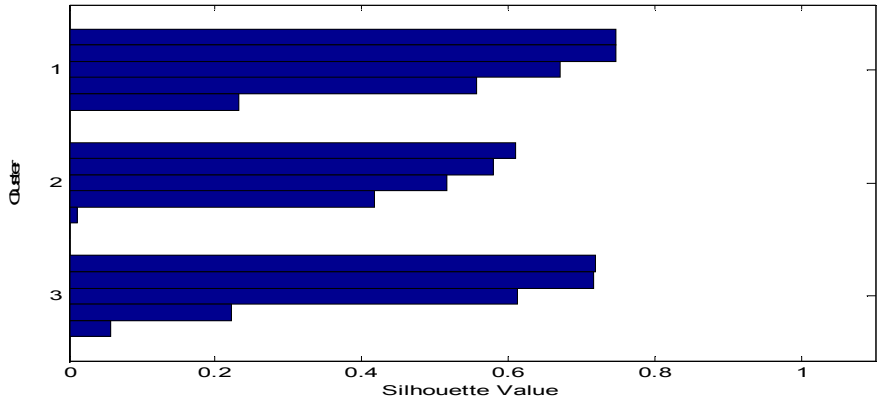
٥ -Ching W.-K.Michael K. N. -2006- Markov Chains: Models, Algorithms and Applications -Springer Science+Business Media, Inc-Usa.

٦ -Introduction to operations research/ ٢٠٠١- Hillier, F. S, Lieberman G. J. -7th ed. New York -Usa.

٧ - Matlab (7.10.0)-2010 -The Language of Technical Computing-The MathWorks,Inc- Usa.

8-Rencher ,A. C...-2002-Methods of Multivariate Analysis -2nd ed - John Wiley & Sons, Inc.-Usa.

ملحق-١-الرسم البياني



المخطط البياني لمقياس صحة العنقدة الظلي

الملحق -٢-البرنامج

```
clc
clear
 a=[50 50];
 m=[10 2;2 10];
for j=1:2;
15); t=mvnrnd(a,m,
if j==2 ,t=sort(t,'descend');
end
t= sort(t);
```

```
idx=kmeans(t,3);
a=idx;
silhouette(t,a,'Euclidean');
if j==1 ,  a1=a;
end
if j==2 , a2=a;
end
 [t  a];
end
[a1,a2]
x=zeros(3,3)
5;٢for j1=1:
 x(a1(j1),a2(j1))=x(a1(j1),a2(j1))+1;
end
 x
 s1=sum(x,2);
 for j2=1:3
 for j3=1:3;
x(j2,j3)=x(j2,j3)/s1(j2);
end
end
x
```

The Non-parametric Estimation of Transitions Probability Matrix by using Clustering Analysis

Ihsan K.SH.AL-Querashy
Al-Qadiseua University

Majed CH.A.AL-Equaly
Amarah Tech.Institute

Abstract

The aim of cluster analysis (clustering) is to classify a collection of data objects into clusters, such that objects within a cluster are more similar to each other than the objects in the other clusters. The change that happen on the state of the system with time become as behavior of Stochastic Process (Markov Chains)that changing with time ,we can represent the state of the system like Markov Chains and we can estimate the transitions probability matrix . We use the simulation to generating the data and apply the cluster analysis and estimate the transitions probability matrix.