مجلة جامعة تشرين للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (39) العدد (5) 2017
Tishreen University Journal for Research and Scientific Studies - Engineering Sciences Series Vol. (39) No. (5) 2017

Study the Effect of Categorical Attributes Weight (Parameter(γ)) on Clustering Mixed Data Types in Data Mining.

Dr. Jabr Hanna*
Faten Alkrdy**

(Received 20 / 9 / 2017. Accepted 16 / 10 / 2017)

□ ABSTRACT □

Data Mining is considered as the most important research scopes by researchers over the world, the reason behind this importance comes from its usage in various areas such as research, scientific, economic and military. This techniques in fact is alternative way to the old traditional querying systems that was common in past, and introduce a powerful technique to discover a hidden knowledge from a large dataset, which wasn't clear before applying it.

In this article we study effect of the categorical attributes weight in measuring distances between objects which are clustered, the common clustering algorithm called K-prototypes is applied to the 'adult' dataset that contains six numerical attributes and nine categorical attributes describing a large amount of people whose ages, education levels, occupations, nations and salaries are different.

We choose two numerical attributes and three categorical attributes from the dataset to be clustered using K-prototypes algorithms, considering several values for (γ parameter) (0.25, 0.5, 0.75, 1) respectively, after that we applied the Rand Index criterion to check the quality of clustering operation in the four scenario.

Keywords: Clustering, Data Mining, Numerical Attributes, Categorical Attributes.

* **Professor Doctor in Department of computer and automatic control Engineering, Faculty of Mechanical and electrical Engineering, Tishreen University, Latakia, Syria.**
** **Postgraduate Student in Department of computer and automatic control Engineering, Faculty of Mechanical and electrical Engineering, Tishreen University, Latakia, Syria.**

مجلة جامعة تشرين للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (39) العدد (5) 2017
Tishreen University Journal for Research and Scientific Studies - Engineering Sciences Series Vol. (39) No. (5) 2017

دراسة تأثير وزن الصفات الفئوية (البارامتر (γ)) على عنقدة البيانات المختلطة في التنقيب عن البيانات

د. جبر مخائيل حنا*
فاتن فجر الكردي**

(تاريخ الإيداع 20 / 9 / 2017. قُبِل للنشر في 16/ 10 / 2017)

□ ملخّص □

يعتبر التنقيب عن البيانات من الأبحاث التي حازت على اهتمام عدد كبير من الباحثين في جميع أنحاء العالم، ويعود السبب في ذلك إلى استخدامها على نطاق واسع في شتى المجالات البحثية والعلمية والاقتصادية والعسكرية، جاءت الحاجة إلى التنقيب عن البيانات بسبب كم البيانات الهائل الذي نتعامل معه اليوم بسبب التطور السريع والمتزايد لتكنولوجيا المعلومات ونظم الاتصالات والانترنت، وقدمت حلاً بديلاً عن الطرق التقليدية السابقة والتي تعتمد على تخزين هذا الكم الهائل من البيانات ضمن قاعدة بيانات ومن ثم القيام بعمليات الاستعلام التي تتطلب وقتاً وجهداً كبيرين من قبل المبرمجين والحواسيب التي تطبق عليها هذه العمليات، علاوة على أن هذه التقنية يمكن من خلالها الكشف عن معرفة مخبأة لم يكن ليتسنى لنا الوصول إليها باستخدام الطرق التقليدية السابقة.

تم في هذا البحث دراسة تأثير وزن الصفات الفئوية على عنقدة البيانات المختلطة، حيث طبقت خوارزمية العنقدة K–prototypes على مجموعة بيانات Adult Dataset والتي تتألف من ست صفات عددية وتسع صفات فئوية، وتتضمن معلومات حول أشخاص من جنسيات مختلفة ولديهم وظائف متنوعة ومستويات ثقافية متتالية بالإضافة إلى المستوى المعيشي لهم.

تم اختيار صفتين عدديتين وثلاث صفات فئوية ثم تطبيق خوارزمية العنقدة K–prototypes على هذه الصفات مع اختيار لقيم وزن الصفات الفئوية (البارامتر (γ)) (0.25 ، 0.5 ، 0.75 ، 1) على الترتيب، ثم تم تطبيق المعيار Rand Index لقياس جودة العنقدة .

الكلمات المفتاحية: العنقدة، التنقيب عن البيانات، الصفات العددية، الصفات الفئوية.

*أستاذ – قسم هندسة الحاسبات والتحكم الآلي– كلية الهندسة الميكانيكية والكهربائية– جامعة تشرين– اللاذقية– سورية.
**طالبة دراسات عليا – قسم هندسة الحاسبات والتحكم الآلي– كلية الهندسة الميكانيكية والكهربائية– جامعة تشرين– اللاذقية– سورية.

مقدمة:

إن استخدام تقنيات التنقيب في البيانات في جميع المجالات يوفر للمؤسسات القدرة على استكشاف والتركيز على أهم المعلومات في قواعد البيانات، كما تركز تقنيات التنقيب في البيانات كذلك على بناء التنبؤات المستقبلية واستكشاف السلوك والاتجاهات مما يسمح باتخاذ القرارات الصحيحة في الوقت المناسب.

يتميز عصرنا الراهن (عصر الانترنت والاقتصاد الرقمي) بالكمالهائل والانتشار واسع النطاق للبيانات حتى أضحى من المستحيل على المحللين استخلاص معلومات ذات معنى باللجوء فقط إلى الطرق التقليدية للتحليل التمهيدي للبيانات أو الاستعلامات التقليدية.

مع وجود كميات كبيرة من البيانات المخزنة في قواعد البيانات ومخازن البيانات ازدادت الحاجة إلى تطوير أدوات تمتاز بالقوة لتحليل البيانات واستخراج المعلومات والمعارف منها، من هنا ظهر ما يسمى بالتنقيب في البيانات كتقنية تهدف إلى استخراج المعرفة من كميات هائلة من البيانات. وهي تقنية حديثة فرضت نفسها بقوة في عصر المعلوماتية، واستخدامها يوفر للشركات والمنظمات في جميع المجالات القدرة على استكشاف والتركيز على أهم المعلومات في قواعد البيانات، والتي تعتبر بدورها مرحلة من مراحل عملية أكثر تعقيدا هي استكشاف المعرفة في مثل هذه القواعد. حيث أن الكثير من الشركات والمنظمات الرائدة اليوم تستخدم عملية استكشاف المعرفة في قواعد البيانات بشكل منهجي ومنظم لتحديد التوجهات المناسبة وتحقيق المنافسة المطلوبة.

أهمية البحث وأهدافه:

يهدف هذا البحث إلى دراسة تأثير وزن الصفات الفئوية على تابع حساب المسافة بين العناصر التي يتم عنقدتها باستخدام خوارزمية العنقدة المعروفة K–prototypes والمخصصة للاستخدام على البيانات المختلطة Mixed Data بنوعيها العددية Numerical والفئوية Categorical ومن ثم قياس أداء العنقدة في كل مرة تم تغيير قيمة هذا البارامتر فيها بغية معرفة مدى تأثيره، وأي مجال من القيم تكون هي الأكثر مناسبة لاستخدامه تبعاً لمجموعة البيانات المدروسة.

طرائق البحث ومواده:

يرتبط هذا البحث ببعض المفاهيم النظرية حول العنقدة والتنقيب عن البيانات وصولاً إلى مفهوم اكتشاف المعرفة مع شرح مختصر وواضح للخوارزمية المستخدمة في البحث ومعيار قياس دقة العنقدة.

مراحل اكتشاف المعرفة:

يمكن تلخيص المراحل الأساسية لاكتشاف المعرفة من البيانات بالخطوات التالية:

1. تكامل البياناتData integration: يتم في هذه المرحلة تجميع البيانات المتشابهة وذات الصلة من مصادر البيانات المتعددة ودمجها معا.

2. اختيار البيانات Data selection: في هذه المرحلة، يتم تحديد واسترجاع البيانات الملائمة من مجموعة البيانات بحيث نحصل على البيانات المطلوبة (target data).

3. تحويل البيانات Data transformation: في هذه المرحلة يتم تحويل البيانات إلى نماذج مخصصة ملائمة لإجراءات البحث والاسترجاع بواسطة تنظيف البيانات (data cleaning) كالتخلص من القيم الفارغة null والقيم المكررة Repeated Valuesوتقليل الأبعاد وغيرها من العمليات الضرورية كالتقييس على أشعة البيانات (normalization).

4. التنقيب عن البيانات Data mining: أي استخدام طرق ذكية تطبق لاستخلاص أنماط البيانات لاستخراج نماذج مفيدة قدر الإمكان وهنا تستخدم خوارزميات كثيرة لهذا الغرض.

5. تقييم النمط Pattern evaluation: يتم في هذه المرحلة تحديد الأنماط المهمة حقا والتي تمثل قاعدة المعرفة لاستخدام بعض المقاييس المهمة.

6. تمثيل المعرفة وتقديمها Knowledge presentation: وهي المرحلة الأخيرة من مراحل اكتشاف المعرفة في قواعد البيانات وهي المرحلة المهمة للجهة المنفذة لعملية التنقيب، وتعتبر هذه المرحلة أساسية كما أنها تستخدم الأسلوب المرئي لمساعدة الجهة المستفيدة في فهم وتفسير نتائج استخراج البيانات.[2]

مفهوم التنقيب في البيانات (Data Mining):

التنقيب في البيانات هو تقنية تهدف إلى استنتاج المعرفة من كميات هائلة من البيانات، تعتمد على الخوارزميات الرياضية والتي تعتبر أساس التنقيب عن البيانات وهي مستمدة من العديد من العلوم مثل علم الإحصاء والرياضيات والمنطق، والذكاء الاصطناعي، والنظم الخبيرة، وعلم التعرف على الأنماط، وعلم الآلة. وغيرها من العلوم والتي تعتبر من العلوم الذكية وغير التقليدية.

ظهر مصطلح التنقيب في البيانات في منتصف التسعينات في الولايات المتحدة الأمريكية، و توجد عدة تعاريف لهذا المفهوم، حيث يمكن تعريفه بأنه: " الاستكشاف الآلي أو المؤتمت لأنماط شائعة و غير واضحة مخفية في قاعدة بيانات معينة"[(1)] ،أو أنه: " طريقة تحليل دقيقة وذكية، تفاعلية و تسلسلية، تسمح للمطورين عند استخدام هذه الطريقة باتخاذ قرارات والقيام بأعمال ملائمة لصالح المسؤولين عن المنظومة المدروسة و المؤسسة التي يعملون بها"[(2)]، أو أنه: " عبارة عن تحليلات لكمية كبيرة من البيانات بغرض إيجاد قواعد و أمثلة و نماذج يمكن أن تستخدم لكي تقود و تدل أصحاب القرار، و تتنبأ بالسلوك المستقبلي"[(3)]، كما يمكن تعريفه كذلك بأنه: " تحليل لمجموعات كبيرة الحجم من البيانات المشاهدة للبحث عن علاقات محتملة و تلخيص للبيانات في أشكال جديدة لتكون مفهومة و مفيدة لمستخدمها" .[(4)]

من خلال التعاريف السابقة يمكن القول بأن التنقيب في البيانات عبارة عن طريقة لاستخراج أو اكتشاف معرفة مفيدة وقابلة للاستخدام الأمثلي من خلال مجموعة كبيرة من البيانات. حيث يساعد في استكشاف المعرفة المخفيّة والنماذج غير المتوقّعة، إضافةً إلى استكشاف قواعد جديدة موجودة في قواعد بيانات كبيرة.

وقد اجتذبت هذه التقنية الكثير من الاهتمام في الأوساط البحثية على مدى العقد الماضي، في محاولة لتطوير خوارزميات قابلة للتوسع والتكيف مع كميات متزايدة من البيانات لكي تساعد في البحث عن أنماط معرفية ذات معنى. وقد نمت مجموعة من الخوارزميات والبرمجيات وبشكل كبير خلال هذه الفترة، إلى حد أن التوسع قد جعل من الصعب على العاملين في هذا الحقل تتبع التقنيات المتاحة لحل مهمة معينة.

تقييس أشعة الصفاتNormalization:

التقييس هو طريقة من نوع خاص تهدف لاستخلاص الخصائص (feature extraction) فيما يلي بعض الأنواع المشهورة.

3-3-1Min- max normalization:

وهو أحد طرق تقييس الصفات (Larose 2005) والذي استخدم في كثير من الأبحاث وهو معرف بالعلاقة:[4]

$$x_{ij}^{*} = \frac{x_{i,j} - x_{min,j}}{x_{max,j} - x_{min,j}} \qquad (1)$$

النتائج هنا لا تعتمد على الوحدات الأصلية للبيانات. وهذا التوازن الخطي يحول البيانات إلى المجال [0,1] في كل الأحوال فإن هذه الطريقة لم تحقق مساواة في متوسطات الصفات، وبذلك فإن تطبيق هذا التحويل من أجل تقييس مجموعات بيانات العالم الحقيقي والعنقدة باستخدام المسافات الاقليدية ومقياس منكويسكي سوف لن يعطي مساهمات متساوية للصفات في مقياس التشابه لأن المتوسط الحسابي للصفات المقيسة لا يكون نفسه.

Z -Score Standardization:

وهي تقنية أخرى مشهورة للتقييس والتي تقيس المتغيرات بأخذ الفرق بين قيمتها وقيمة المتوسط الحسابي لها (mean) وموازنة هذا الفرق بأخذ الانحراف المعياري للمتغير (Larose 2005, Jain and Dubes 1988) وهو يعطى بالعلاقة

$$x_{ij}^{*} = \frac{x_{ij} - \bar{x_j}}{s_j(x)} \qquad (2)$$

من أجل البيانات العددية وعندما يتم حساب المسافات الاقليدية فإن المنهج العام لهذه الطريقة من التقييس z-score هو تحويل الصفة A_j^n إلى متغير عشوائي له متوسط zero mean وفرق unit variance كمايلي:[4]

$$x_{ij}^{*} = \frac{x_{ij} - \mu_j}{\sigma_j} \qquad (3)$$

حيث μ_j هو meanالمتوسط و σ_j هي standard deviation الانحراف المعياري للبعد j للصفة A_j^n المعتبرة هذا التقييس يقدم مساهمات متساوية للصفات في مقياس التشابه الإقليدي.

أدوات التنقيب عن البيانات:

هناك بعض الأدوات المستخدمة في التنقيب عن البيانات منها:

1. التصنيفclassification: يعد التصنيف من أهم الأدوات المستخدمة في التنقيب عن البيانات فهو يستخدم من أجل توزيع السجلات المخزنة في قاعدة البيانات على مجموعة من الصفوف المحددة مسبقاً. وهناك الكثير من الخوارزميات المستخدمة كمصنفات منها 3id والمصنف 4.5c ومصنفات أشجار القرار وغيرها.

2. العنقدةclustering: وهي أيضا من الأدوات الهامة للتنقيب عن البيانات يتم فيها تجميع الأغراض في عناقيد مختلفة تبعاً لمقدار تشابهها واختلافها ومن أشهر أنواعها العنقدة التقسيمية والعنقدة الهرمية، ولكل نوع عدد كبير من الخوارزميات المستخدمة، يوجد بالإضافة إلى ما سبق عدة أدوات أخرى مثل قواعد الارتباط وتحليل السلاسل الزمنية ولكل منها خوارزميات خاصة.

ونظراً لوجود هذا الكم الكبير من الخوارزميات التي تقوم بعمليات التنقيب وجدت أن اختيار أي واحدة من هذه الخوارزميات هي واحدة من أصعب المهام في عملية التنقيب بحيث تكون الخوارزمية المختارة مناسبة لنوع البيانات المتاحة.

العنقدة Clustering:

وهي تقنية مهمة ومشهورة جداً في التنقيب عن البيانات يتم فيها تجميع مجموعات ضخمة من البيانات في عناقيد تحوي مجموعات أصغر ببيانات متشابهة بحيث يكون العنقود مفيداً وذو معنى وذلك باستخدام تقنيات مختلفة. العنقود هو مجموعة من أغراض البيانات التي تكون متشابهة في نفس العنقود (قد يكون التشابه كبير أو صغير) ومختلفة عن الأغراض الموجودة في العناقيد الأخرى وحيث أن حالة التشابه تقاس بالمسافة بين الأغراض بحيث تكون المسافة أصغرية بين أغراض العنقود الواحد (الأغراض المتشابهة) وتكون كبيرة نسبياً بين الأغراض المنتمية إلى عناقيد مختلفة (الأغراض غير المتشابهة).

معظم خوارزميات العنقدة تبدأ ببناء النموذج من خلال تكرار مجموعة من العمليات (الإجراءات) وتتوقف عندما يصل النموذج إلى الحالة المطلوبة (أي حدود العناقيد تثبت). وهناك عدة ميزات للعنقدة كالقدرة على التعامل مع أنواع مختلفة وكبيرة من الصفات. بالإضافة إلى اكتشاف عناقيد لم تكن ظاهرة من قبل، كما أنها تحتاج متطلبات قليلة حول معرفة البيانات المراد عنقدتها لتحديد بارامترات الدخل، وهي قادرة على التعامل مع الضجيج والعناصر الخارجة.

أنواع العنقدة:

يمكن تصنيف العنقدة إلى عدة أصناف، نذكر منها:

1. العنقدة المعتمدة على نماذج البيانات Model_based clustering:

أي خوارزمية تعليم مراقب أو (supervised learning) يجب تزويدها بمجموعة بيانات تدريب تقدم معلومات عن بعض الأجزاء في مجموعة البيانات. هدف التعليم المراقب هو بناء مصنف يولد لكل سجل دليلاً لصنف كخرج له (class lable) باستخدام قواعد مجموعة بيانات التدريب يتم توقع دليل lable للأمثلة الجديدة والتي هي غير موجودة في training set أي مجموعة التدريب.[3]

2. العنقدة المعتمدة على التابع الوصفي objective function_based clustering:

مجموعة كبيرة من خوارزميات العنقدة تهتم ببناء العناقيد لمجموعة البيانات بالاعتماد على قيمة مؤشر أداء يسمى التابع الوصفي objective function ويعرف أيضا باسم تابع التكلفة cost function وهو مرتبط مع مشكلة التحسين. حيث يشكل العنصر الأفضل من جزء مجموعة البدائل المتاحة والذي يتم اختياره ليصغر أو يكبر من قيمة هذا التابع. قيمة هذا التابع تحدد مدى جودة الحل الذي تم اختياره، وهناك خوارزميات عنقدة تختلف حول العنقدة المعتمدة على التابع الوصفي لأنه من غير المجدي عملياً أن نجد الحالة المثلى لتابع وصفي عبر الأخذ بعين الاعتبار لكل العناصر.[3]

3. العنقدة التقسيمية:

تطبيقات التنقيب عن البيانات تحتاج إلى تقسيم البيانات إلى عناقيد متجانسة والتي يمكن فيها اكتشاف مجموعات مثيرة للاهتمام من هذه البيانات، مثل معرفة الأشخاص الذين يحملون بوليسة تأمين سيارة، أو مجموعة العملاء في قاعدة بيانات مصرفية لديهم استثمارات عقارية، أو المنتجات التي تباع معا في متجر ما وغيرها من المعلومات في مختلف المجالات.[3]

للقيام بهذا التحليل هناك مشكلتين يجب العمل على حلهما، الأولى هي كفاءة تقسيم مجموعة كبيرة من البيانات إلى مجموعات متجانسة. والثانية هي فعالية تفسير ما تقدمه هذه المجموعات. نحن في هذا البحث نقدم حلا للمشكلة الأولى ونقترح حلا للثانية. العديد من خوارزميات العنقدة يتم توظيفها لحل المشكلة الأولى.

عندما يتم معرفة القليل عن توزع البيانات فإن خوارزميات العنقدة يتم استخدامها، ومع ذلك فإن العنقدة في مجال التنقيب عن البيانات تختلف عنها في التطبيقات الأخرى. كما أن التعامل مع البيانات الكثيرة الأبعاد التي تحوي الاف أو حتى ملايين من السجلات مع عشرات أو مئات الصفات) يجعل استخدام خوارزميات العنقدة العادية في تطبيقات التنقيب عن البيانات أمراً صعباً أو غير ممكن، بالإضافة إلى أن البيانات في تطبيقات التنقيب عن البيانات تحتوي غالباً على قيم عددية وفئوية. فالطريقة التقليدية للتعامل مع الصفات الفئوية كصفات عددية لا تعطي غالباً نتائج جيدة أو نتائج ذات معنى، لأن مجالات الصفات الفئوية لا تكون مرتبة.

على الرغم من أن خوارزميات العنقدة الهرمية القياسية تستطيع أن تتعامل مع البيانات ذات الأنواع المختلطة (العددية والفئوية) 5و8 فإن التكلفة الحسابية لها تجعلها غير مقبولة لعنقدة البيانات الضخمة.

الخوارزمية المستخدمة في البحث:

تم استخدام خوارزميةK−prototypesوالتي تقوم على تكبير درجة التشابه بين أغراض العنقود الواحد كما في خوارزمية K−means ولكن درجة التشابه للغرض يتم معرفتها من خلال كل من الصفات العددية والفئوية وليس العددية فقط، وعندما يتم تطبيقها على بيانات عددية فإنها تصبح خوارزمية k−means المشهورة.

بتجريب هذه الخوارزمية على مجموعة بيانات حقيقية أظهرت قدرتها على تقسيم هذه البيانات إلى عناقيد تحوي مئات أو آلاف من الأغراض.

ويمكننا أن نبين باختصار بعض المفاهيم الرياضية لخوارزمية K−prototypes فإذا كان لدينا مجموعة من الأغراض$X = \{X_1, X_2, \ldots . X_n\}$ يرمز إلى n غرض وكل غرض$X_i = [X_{i1}, X_{i2}, \ldots . X_{im}]$يملك m قيمة للصفات. وبفرضk هو عدد صحيح موجب، فإن مهمة عنقدة X هو أن يتم إيجاد التقسيم الذي يجعل جميع الأغراض في X تتوزع إلى k عنقود مختلف.

ويعطى تابع التكلفة الأكثر انتشاراً يعطى بالعلاقة التالية:[7]

$$E = \sum_{l=1}^{K} \sum_{i=1}^{n} y_{il} d(x_i, Q_l) \quad (4)$$

$Q_l = [q_{l1}, q_{l2}, \ldots q_{lm}]$ممثل نموذج (prototype) شعاع مركز العنقود l . وحيث y_{il} هو عنصر مصفوفة التقسيم Y_{n*l}و d هو مقياس التشابه وغالباً يتم تعريفه بمربع المسافة الإقليدية. أما y فإنها تملك الخاصيتين التاليتين:$\sum_{l=1}^{k} y_{il} = 1$و $0 \le y_{il} \le 1$

فإننا نسمي التقسيم قاسي (hard partition) إذا كان لدينا:

$y_{il} \in \{0,1\}$وما عدا ذلك يسمى التقسيم ليناً (fuzzy partition) . في التقسيم القاسي إذا كانت $y_{il} = 1$فإن ذلك يدل على أن الغرض X_i تم نسبه إلى العنقود l . وفي هذ البحث ستتم دراسة العنقدة القاسية فقط.

من أجل العنقود l يمكن أن نكتب:

$$E_l = \sum_{i=1}^{n} y_{il} d(x_i, Q_l) \quad (5)$$

وهذه العلاقة تعبر عن التكلفة الكلية لنسب X إلى العنقود l . عندما تملك X صفات فئوية، يمكن تقديم مقياس تشابه كما في العلاقة التالي:

$$d(X_i, Q_l) = \sum_{j=1}^{m_r}\left(X_{ij}^r - q_{lj}^r\right)^2 + \gamma_1 \sum_{j=1}^{m_c} \delta\left(X_{ij}^c, q_{lj}^c\right) \qquad (6)$$

حيث $\delta(p,q) = 0$ عندما p=q و $\delta(p,q) = 1$ عندما p≠q . X_{ij}^r, q_{lj}^r هي قيم الصفات العددية للغرض ومركز العنقود، و X_{ij}^c, q_{lj}^c هي قيم الصفات الفئوية للغرض i ومركز العنقود l . أما m_r, m_c هي عدد الصفات العددية والفئوية.

γ_1 هي وزن الصفات الفئوية للعنقود l ويمكن أن نكتب E_l بالشكل:

$$E_1 = \sum_{i=1}^{n} y_{il} \sum_{j=1}^{m_r}(X_{ij}^r - q_{lj}^r)^2 + \gamma_1 \sum_{i=1}^{n} y_{il} \sum_{j=1}^{m_c} \delta\left(X_{ij}^c, q_{lj}^c\right) \qquad (7)$$

$$E_1 = E_1^r + E_1^c \qquad (8)$$

حيث E_1^rهي التكلفة الكلية لجميع الصفات العددية، و E_1^c هي التكلفة الكلية لجميع الصفات الفئوية.

وبشكل عام يمكن تلخيص الخطوات الرئيسية لعمل هذه الخوارزمية كما يلي:

1. اختيار k مركز ابتدائي للعناقيد (initial prototype) من مجموعة البيانات X بشكل عشوائي، مركز واحد لكل عنقود.
2. نسب كل غرض من مجموعة البيانات X إلى عنقود بحيث أن التشابه بين الغرض ومركز العنقود أقل التشابه بين هذا الغرض وبقية مراكز العناقيد الأخرى.
3. بعد نسب جميع الأغراض إلى مراكز عناقيد، يتم تحديث مراكز العناقيد بعد كل عملية توضع للأغراض.
4. نعيد اختبارات التشابه بين كل غرض ومراكز العناقيد الجديدة، إذا وجد غرض أقرب مركز له ينتمي إلى عنقود آخر مغاير لعنقوده، تتم إعادة نسب الغرض إلى هذا العنقود ذي المركز الأقرب.
5. نعيد تحديث مراكز العناقيد.
6. نعيد الخطوات السابقة حتى لا يعود أي غرض إلى تغيير العنقود الذي ينتمي إليه.

تقييم أداء العنقدة:

يوجد عدة طرق لتقييم جودة العنقدة منها مقياس rand أو Rand index هو مقياس تشابه بين طريقتي عنقدة للبيانات وهو معرف من قبل (Rand 1971) وهو معرف كمايلي :

إذا كان لدينا مجموعة S من N عنصر وجزئين {C={c1,c2,.....ck و {D={d1,d2,.....dk من مجموعة البيانات الحساب Rand index يجب أولا حساب الأرقام التالية :

a عدد أزواج العناصر في S والتي هي موجودة في نفس الوقت في C و D معا.

b عدد أزواج العناصر في S والتي تكون موجودة في مجموعات مختلفة عن C و D .

c عدد أزواج العناصر في S التي تكون موجودة في المجموعة C وهي غير موجودة في D .

d عدد أزواج العناصر في S التي تكون موجودة في المجموعة D وهي غير موجودة في C .

عندئذ فإن Rand Index يتم حسابه كمايلي:

$$R = \frac{a+b}{a+b+c+d} \qquad (9)$$

أي يمكن أن نعتبر أن a+b يمثل عدد التوافقات بين C و D و c+d يمثل عدد حالات عدم التوافق بين C و D. وهذا الشكل تم استخدامه من قبل عدة باحثين مثل (Huang، 2005) لتقييم خوارزمية العنقدة بأوزان الصفات. ونحن يمكن أن نستخدم العلاقة (9) لأن cluster lables لنقاط البيانات التجريبية تكون معروفة.

ملاحظة: Rand Index لديه قيمة بين 0 و 1 وكلما كانت قيمة R أكبر تكون العنقدة أفضل.

تحليل مجموعة البيانات المستخدمة في البحث:

تم استخدام مجموعة البيانات Adult Dataset وهي عبارة عن إحصائية لمجموعة من السكان الذين ينتمون إلى جنسيات متعددة، حيث تتضمن مجموعة البيانات معلومات حول العمر والمستوى التعليمي وطبيعة العمل والحالة الاجتماعية والبلد الأم والمستوى المعيشي، ويبين الجدول (1) ملخص عن مجموعة البيانات المستخدمة.

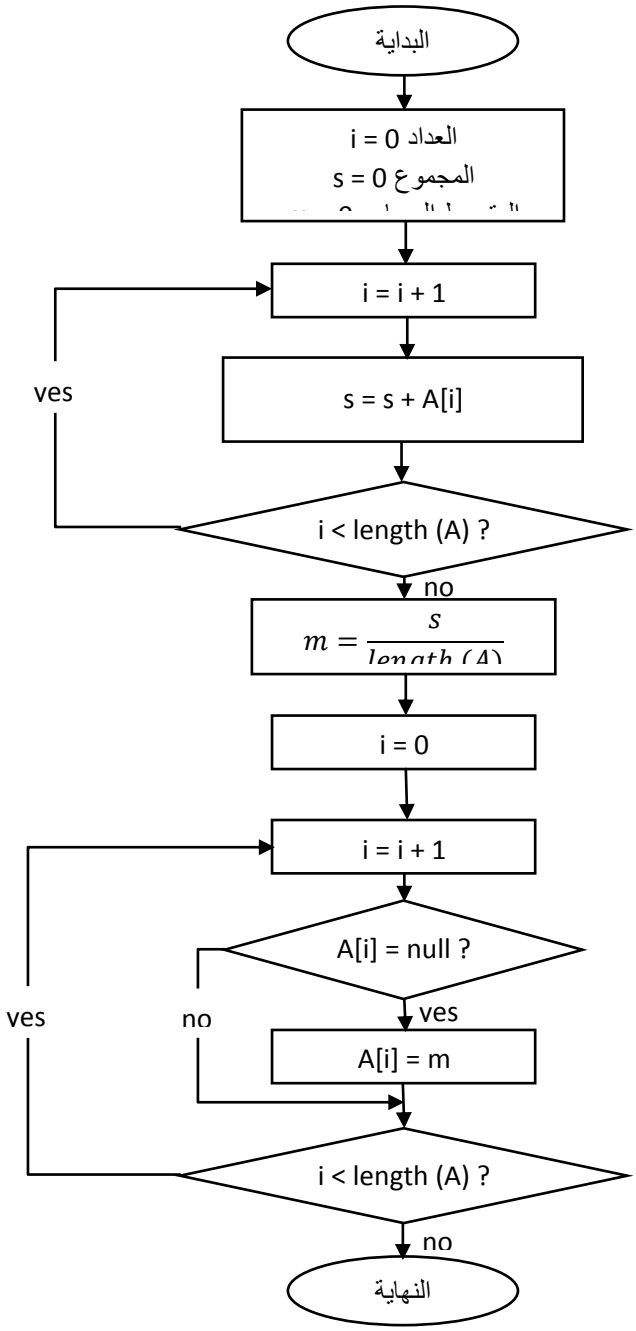
الجدول (1): ملخص عن مجموعة البيانات المدروسة

اسم العمود	صفة البيانات في العمود
العمر Age	عددية
صنف العمل Work Class	فئوية
الوزن الاستطلاعي Survey Weight	عددية
الثقافة Education	فئوية
الثقافة (كرقم)(Education (Number	عددية
الحالة الاجتماعية Martial Status	فئوية
العمل Occupation	فئوية
العلاقة Relationship	فئوية
العرق Race	فئوية
الجنس Sex	فئوية
الربح الأعظمي Capital Gain	عددية
الخسارة العظمى Capital Loss	عددية
عدد ساعات العمل في الأسبوع Hours Per Week	عددية
البلد الأم Native Country	فئوية
صنف الراتب Salary Class	فئوية

تم اتباع مجموعة من العمليات والتي تمثل مراحل المعالجة الأولية لمجموعة البيانات وتتضمن هذه المراحل مجموعة من الخطوات الغاية منها هي تهيئة مجموعة البيانات قبل تطبيق خوارزمية العنقدة، فهي تشمل عمليات ملء البيانات المفقودة بالقيم المناسبة بحسب كل عمود، ثم تأتي مرحلة التطبيع Normalization والهدف من هذه العملية هي تحويل قيم الأعمدة الرقمية لتصبح كلها ضمن المجال [0,1]، لكي لا تؤثر قيم كبيرة موجودة في أحد الأعمدة على قيم الأعمدة الأخرى أثناء القيام بعملية العنقدة Clustering، وفيما يلي نستعرض طرق المعالجة الأولية لمجموعة البيانات.

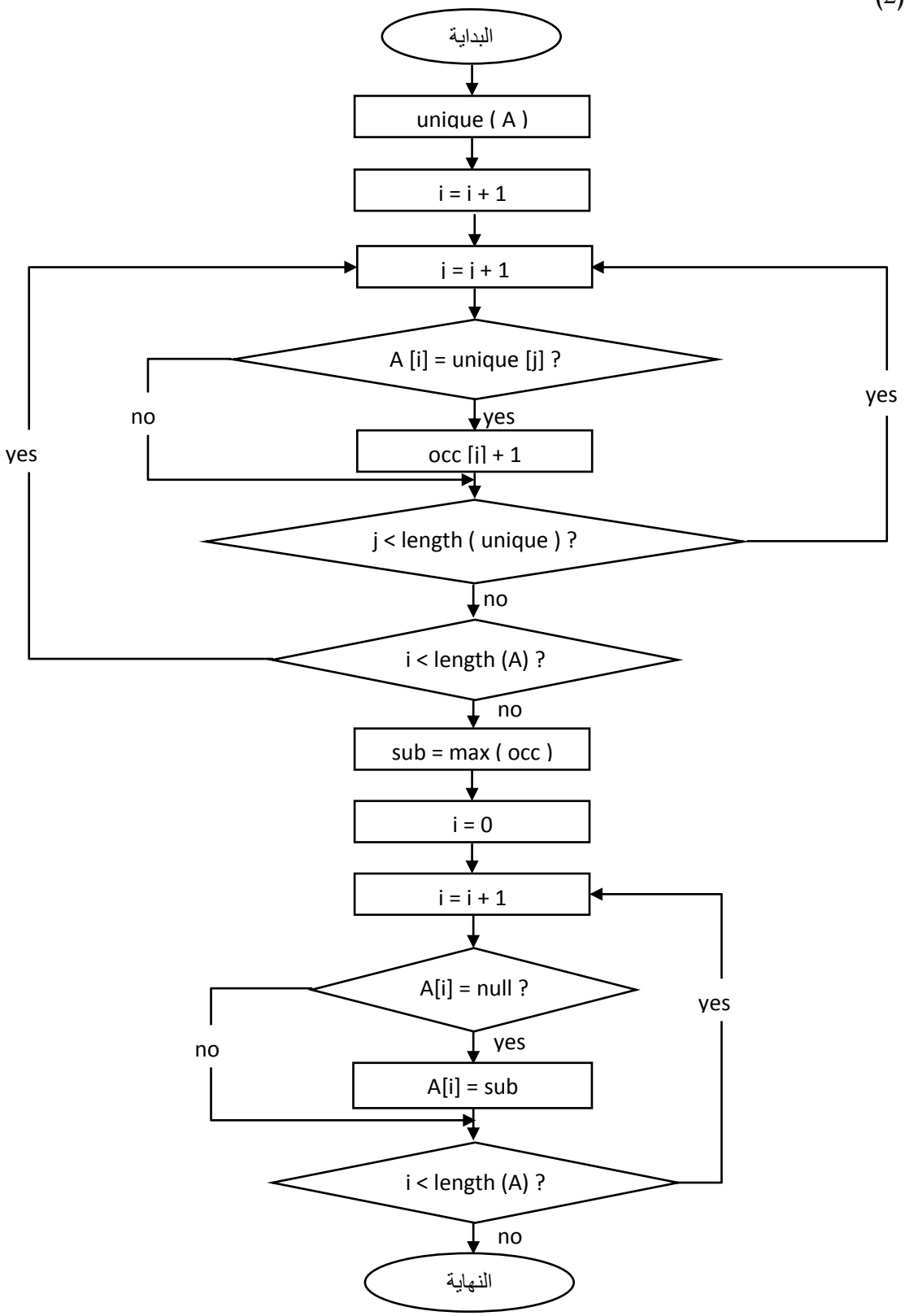
1. ملء القيم المفقودة:

يتم ملء القيم المفقودة في كل عمود بما يناسب القيم الموجودة في ذلك العمود، بالنسبة للأعمدة العددية مثلاً تم اختيار قيمة المتوسط الحسابي لقيم العمود لملء القيم الفارغة كما هو موضح في الخوارزمية المبينة في الشكل (1).



الشكل (1): المخطط التدفقي لملء القيم المفقودة للبيانات العددية

أما بالنسبة للأعمدة الفئوية فتم اختيار القيمة الأكثر تكراراً ضمن العمود لملء القيم الفارغة كما هو موضح في الشكل (2)



الشكل (2): يبين المخطط التدفقي لملء القيم المفقودة للبيانات الفئوية

مناقشة النتائج:

تم العمل في هذا البحث على أربعة سيناريوهات مختلفة من أجل تطبيق خوارزمية العنقدةK-prototypes على مجموعة البيانات Adult Dataset، وتختلف هذه السيناريوهات فيما بينها بقيمة البارامتر (γ) حيث يأخذ هذا البارامتر القيم (0.25 ، 0.5 ، 0.75 ، 1) على الترتيب، وهذا الاختلاف يؤدي إلى نسب مساهمة مختلفة للبيانات الفئوية في حساب تابع لمسافة تبعاً للوزن المطبق والذي يمثل قيمة هذا البارامتر.

تم في البداية اختيار قيم ابتدائية عشوائية لمراكز العناقيد Centroidsفي السيناريو الأول، ثم اعتمِدت هذه القيم كقيم ابتدائية بالنسبة للسيناريوهات الأخرى حتى تكون المقارنة دقيقة بين القيم عند تغير قيمة البارامتر (γ).

كما تم استخدام خمس صفات؛ اثنتان منها عددية وهي العمر Age وعدد ساعات العمل في الأسبوع Hours per Week، وثلاثة فئوية وهي المستوى الثقافي Education، والبلد الأم Native Country، وصنف الراتب Salary Class.

تم تطبيق عملية التطبيع Normalizationتحديدا Min- max normalizationعلى الصفات العددية لجعل مجالها محصور بين 0 و 1 .

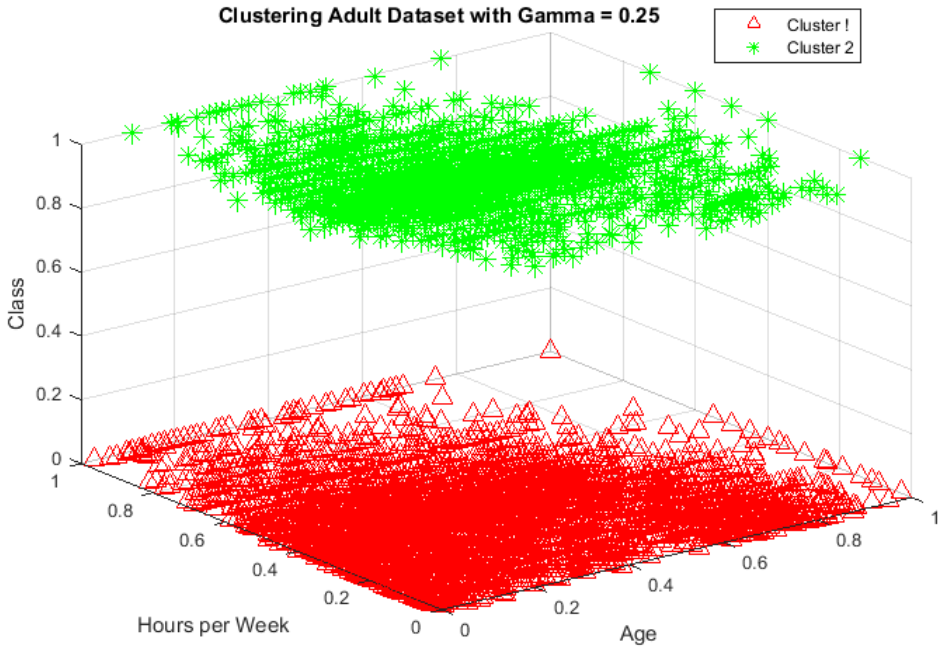
ويوضح الجدول (2) القيم الابتدائية لمراكز العناقيد Centroids المستخدمة في كل السيناريوهات.

الجدول (2): القيم الابتدائية لمراكز العناقيد

العمر Age	ساعات العمل في الأسبوع Hours per Week	المستوى الثقافي Education	البلد الأم Native Country	صنف الراتب Salary Class
42	38	'Bachelors'	'United-States'	'No'
35	40	'High-school'	'Yugoslavia'	'Yes'

-السيناريو الأول:

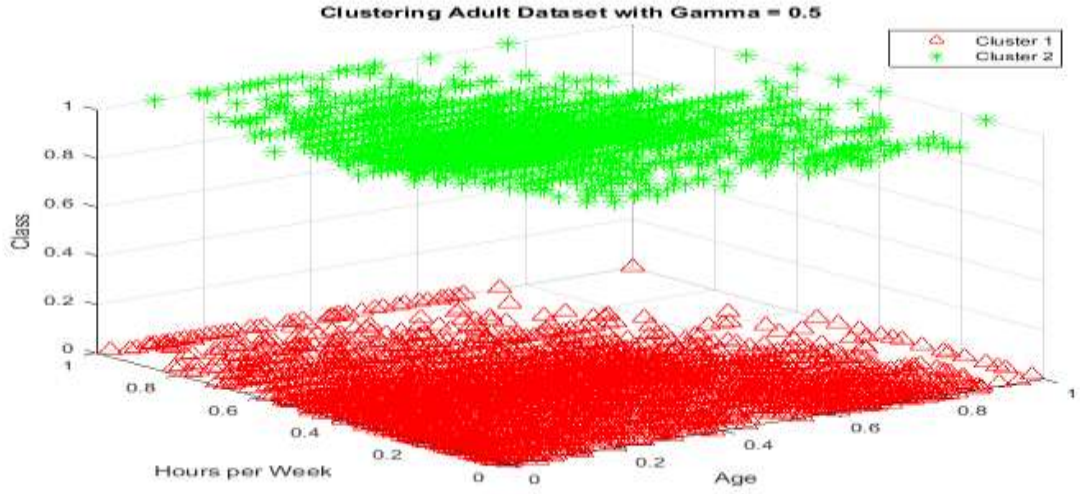
تم استخدام قيمة ($\gamma = 0.25$) على مجموعة البيانات وبالنتيجة حصلنا على عنقودين يبلغ عدد العناصر في كل منهما (7841،24720) للعنقودين الأول والثاني على الترتيب، ويوضح الشكل (3) نتيجة العنقدة المطبقة على مجموعة البيانات حيث يمثل العمر المحور (x) وعدد ساعات العمل في الأسبوع المحور (y) وصنف الراتب المحور (z)، وتمت معالجة صنف الراتب إذا كان أصغر أو يساوي (50000) فإن قيمة هذا البارامتر (0)، وإلا فإن قيمة هذا البارامتر (1) كما هو موجود في مجموعة البيانات، وتم اتباع الأسلوب ذاته في بقية السيناريوهات.



الشكل (3): يبين نتيجة العنقدة وفق السيناريو الأول

-السيناريو الثاني:

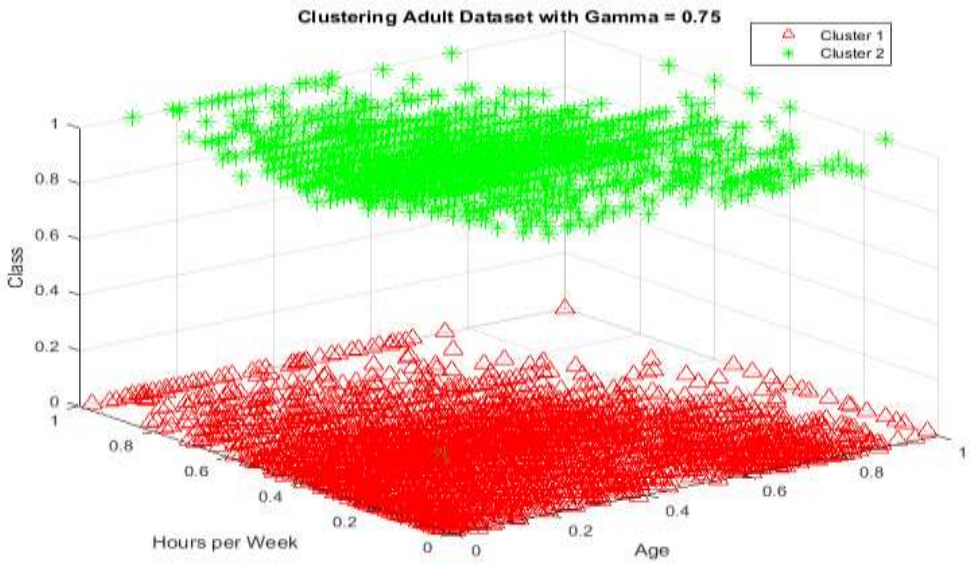
تم استخدام قيمة ($\gamma = 0.5$) على مجموعة البيانات وبالنتيجة حصلنا على عنقودين يبلغ عدد العناصر في كل منهما (7842،24719) للعنقودين الأول والثاني على الترتيب، ويوضح الشكل (4) نتيجة العنقدة المطبقة على مجموعة البيانات.



الشكل (4): نتيجة العنقدة وفق السيناريو الثاني

-السيناريو الثالث:

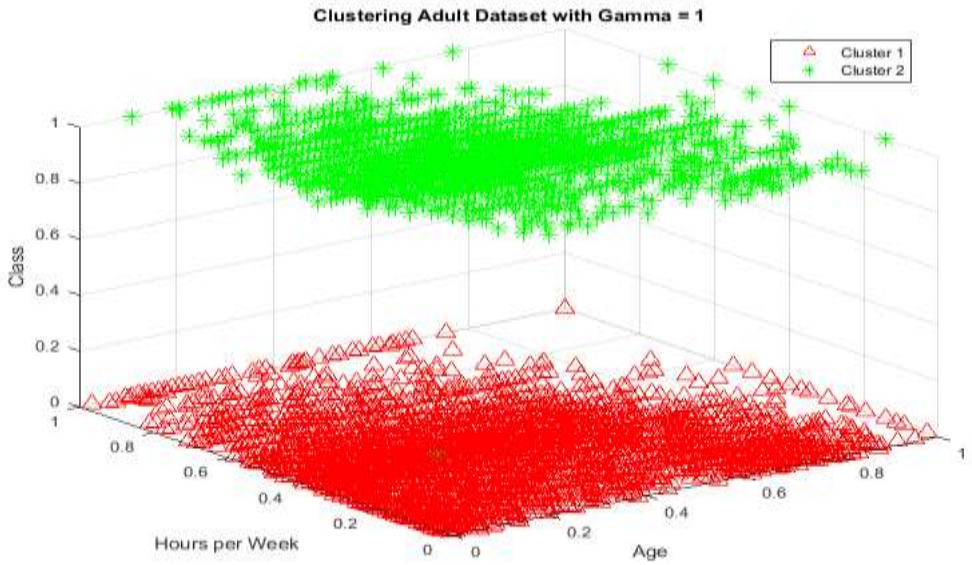
تم استخدام قيمة ($\gamma = 0.75$) على مجموعة البيانات وبالنتيجة حصلنا على عنقودين يبلغ عدد العناصر في كل منهما (7878،24683) للعنقودين الأول والثاني على الترتيب، ويوضح الشكل (5) نتيجة العنقدة المطبقة على مجموعة البيانات.



الشكل (5): نتيجة العنقدة وفق السيناريو الثالث

-السيناريو الرابع:

تم استخدام قيمة ($\gamma = 1$) على مجموعة البيانات وبالنتيجة حصلنا على عنقودين يبلغ عدد العناصر في كل منهما (7921،24640) للعنقودين الأول والثاني على الترتيب، ويوضح الشكل (6) نتيجة العنقدة المطبقة على مجموعة البيانات.



الشكل (6): نتيجة العنقدة وفق السيناريو الرابع

-حساب دقة العنقدة بطريقة Rand Index للسيناريوهات الأربعة:

يبين الجدول (3) حساب معيار دقة العنقدةRand Index عند تطبيقه على السيناريوهات الأربعة السابقة.

الجدول (3): حساب Rand Index للسيناريوهات الأربعة.

	$\gamma = 0.25$	$\gamma = 0.5$	$\gamma = 0.75$	$\gamma = 1$
$\gamma = 0.25$	1	0.9999	0.9977	0.9951
$\gamma = 0.5$	0.9999	1	0.9978	0.9952
$\gamma = 0.75$	0.9977	0.9978	1	0.9974
$\gamma = 1$	0.9951	0.9952	0.9974	1

الاستنتاجات والتوصيات:

الاستنتاجات:

يمكن من خلال النظر في جدول حساب دقة العنقدة استنتاج أن العلاقة بين معيار دقة العنقدة Rand Index وقيمة البارامتر (γ) هي علاقة عكسية؛ أي أن قيمة المعيار.

Rand Index تكون أصغر كلما كانت قيمة البارامتر (γ) كبيرة، وعلى العكس فإنها تكبر كلما كانت قيمة البارامتر (γ)، ويرجع السبب في ذلك إلى أنه كلما كبرت قيمة البارامتر (γ) فإن مساهمة الصفات الفئوية في حساب تابع المسافة تزداد، وهذا ما يتسبب في انتماء بعض العناصر التي تتبع لأحد العنقودين لكي تصبح تابعة للعنقود الآخر.

التوصيات:

-استخدام قيم (γ) كبيرة نسبياً لأن ذلك يزيد من مساهمة الصفات الفئوية في حساب تابع المسافة مما يعطينا دقة عنقدة أفضل.

-تطبيق خوارزمية العنقدةK-Prototypes على مجموعة البيانات نفسها بعد اختيار عدد أكبر من الصفات العددية والفئوية.

-اختيار قيم ابتدائية عشوائية متغيرة لمراكز العناقيد Centroids وحساب دقة العنقدة في النتائج الجديدة.

المراجع:

1. Hand David, HeikkiMannila, Padhraic Smyth, "Principal Of Data Mining", A Bradford Book The MIT Press Cambridge, Massachusetts Londonengland,Pp.322, 2010

2. Sanjeev, Manchanda, "Knowledge Discovery In Database Processing Using Improved Data Mining Techniques", Doctor Of Philosophy In Computer Science, School Of Mathematic And Computer Applications, Thapar University, India, Pp. 184, 2010.

3. Amir Ahmad, LipikaDey, "A K-Mean Clustering Algorithm For Mixed Numeric And Categorical Data", Sciencedirect, Pp.25, 2007

4. C.Saranya, G.Manikandan, "A Study On Normalization Techniques For Privacy Preserving Data Mining", International Journal Of Engineering And Technology (IJET), Pp.4, 2013

5. Maria Del Mar Suarez Alvarez, "Clustering Algorithms For Numerical, Categorical And Mixed Data", A Thesis Submitted To Cardiff University, Pp.224, 2010.

6. Cosmin Marian Poteras, Marian CristianMih˘aescu, MihaiMocanu, "An Optimized Version Of The K-Means Clustering Algorithm", Conference On Computer Science And Information Systems Pp. 695–699, 2014

7. ZHEXUE HUANG, "CLUSTERING LARGE DATA SETS WITH MIXED NUMERIC AND CATEGORICAL VALUES", CSIRO Mathematical And Information Sciences GPO Box 664 Canberra ACT 2601, AUSTRALIA, Pp.14, 2011.

8. MihaiLupu (IRF), Nathalie Steinmetz (UIBK), "State-Of-The-Art And Requirements On Clustering Techniques", Service Detective, October 15, 2009

9. ShaffyGoyal, NamishaModi, "A Review On Various Classification Algorithmsfor Online Shopping Data ",International Journal Of Computer Application (2250-1797), Volume 6– No.2, March- April 2016.

10. Usama Fayyad, *Gregory Piatetsky-Shapiro, Padhraic Smyth, "*From Data Mining toKnowledge Discovery in Databases", American Association for Artificial Intelligence,pp.18. 1996.

11. Maria Del Mar Suarez Alvarez, "Design and Analysis of Clustering Algorithms for Numerical, Categorical and Mixed Data ",thesis submitted to Cardiff University, United Kingdom, pp.224, 2010.