استخدام خوارزمية (K-Means) للعنقدة في تنقيب البيانات (Data Mining) مع واقع تطبيقي

أ.م.د. قتيبة نبيل نايف/ كلية الادارة والاقتصاد / جامعة بغداد
الباحث/ محي الدين خلف أيوب

المستخلص

ان التقدم العلمي الكبير أدى الى الانتشار الواسع للمعلوماتية بحيث اصبحت المعلومات تتراكم بشكل هائل في قواعد بيانات كبيرة ، وهنا تكمن اهمية البحث في محاولة تنقيح وتبويب هذا الكم الهائل من البيانات وتصنيفها بحيث تؤدي الغرض المطلوب في استخراج المعلومات المخفية او في تصنيف البيانات بموجب علاقاتها ببعضها بغية الافادة منها لأغراض تقنية .

وان العمل بمصطلح التنقيب في البيانات (Data Mining) يعد ملائم في هذا المجال لأهمية البحث في استخدام خوارزمية (K-Means) لتصنيف البيانات بأسلوب تقني في واقع تطبيقي مع ما يمكن من ملاحظة التأثير في المتغيرات (v) من خلال تغيير حجم العينة (n) وكذلك عدد العناقيد (K) واثرهما في عملية العنقدة في مراحل الخوارزمية ، من خلال تكوين عناقيد مثالية بحيث تحقق مجموعة بيانات جديدة ومفيدة تجيب عن كل الاستفسارات وبحسب صفات البيانات (Object) العائدة لخوارزمية البحث وبحسب متغيرات البحث (V) المطبقة في برنامج الجانب التطبيقي.

المصطلحات الرئيسية للبحث/ العناصر ، تنقيب البيانات ، العنقدة ، التعليم الالي ، الخوارزمية.

البحث مستل من رسالة ماجستير

مجلة العلوم
الاقتصادية والإدارية
العدد 91 المجلد ٢٢
الصفحات ٣٨٩-٤٠٦

١- المقدمة Introduction:

إن التطور الكبير في مجالات العلوم المختلفة وما صاحبه من تطور في تكنولوجيا المعلومات في عصر الاتصالات الحديثة والانترنت، أدى إلى زيادة كمية البيانات الرقمية إذ لم تعد وسائل التحليل الإحصائي التقليدية قادرة على التعامل معها بغية استخلاص معلومات تفي باتخاذ قرارات صائبة ودقيقة.

كذلك فإن وجود كميات كبيرة من البيانات في قواعد البيانات (Data Base) وفي مخازن البيانات (Data warehouse) زادت الحاجة إلى تطوير أدوات تحليل البيانات لغرض استخلاص واكتشاف المعرفة (Knowledge Discloser) والافادة منها.

ومن هنا ظهر اسم تنقيب البيانات (Data Mining) كأسلوب تقني يهدف إلى استخراج المعلومات المخفية (Hidden information) بأدوات تقنية وإحصائية حديثة.

ومما تجدر الإشارة إليه أن أساليب تنقيب البيانات (Data Mining) تركز أيضاً على بناء التنبؤات المستقبلية في سلوك واتجاهات (Trends) المتغيرات مما يساعد على اتخاذ قرارات أكثر دقة وفي الوقت المناسب .

ومن هنا تكمن اهمية البحث في استخدام احدى طرائق العنقدة وهي خوارزمية (K-Means) في تنقيب البيانات DM في واقع تطبيقي من اجل استخلاص الفائدة والمعرفة من البيانات المطبقة، وبحسب صفات (Object) متغيرات البحث (V) للتجارب المبينة في الجانب التطبيقي .

٢- هدف البحث:

من الانظمة المعلوماتية الجديدة انظمة التنقيب في البيانات ضمن قواعد البيانات (Data Base) ومخازن البيانات (Data Warehouse) ، ومن ادوات وسائل التنقيب في البيانات طريقة (K-Means) إذ يهدف البحث الى استخدام خوارزمية (K-Means) للعنقدة في واقع تطبيقي للوصول الى المعلومة من خلال تصنيف صفات (Object) لمتغيرات (V) البحث التطبيقي فضلا عن معرفة طبيعة العلائق الموجودة بين صفات البيانات والحصول على النتائج المرجوة لأفضل تصنيف للبيانات بالسرعة المطلوبة .

المبحث الاول/ الجانب النظري:

١- مفهوم تنقيب البيانات DM -(Data Mining):

ظهر مصطلح تنقيب البيانات في منتصف الستينات في الولايات المتحدة الأمريكية[2]وهو مصطلح يجمع بين الإحصاء وتكنولوجيا المعلومات وقواعد البيانات (Data Bases)، الذكاء الاصطناعي (Neural Intelligent)، التعليم الآلي (Machine Learning)، وظهرت عدة تعاريف لمفهوم تنقيب البيانات DM منها ((الاستكشاف الآلي أو المؤتمت لأنماط شائقة وغير جلية مخفية في قاعدة البيانات))[3] .

وعُرف ((عبارة عن تحليلات لكمية كبيرة من البيانات لغرض ايجاد قواعد وأمثلة ونماذج يمكن أن تستخدم لتقود وتدل صاحب القرار، وتتنبأ بالسلوك المستقبلي))[4]، وهناك تعريف آخر لمفهوم تنقيب البيانات ((تحليل لمجموعات كبيرة الحجم من البيانات المشاهدة للبحث عن علاقات محتملة، وتلخيص البيانات في أشكال جديدة لتكون مفهومة ومقيدة لمستخدميها)) [26].

وتسمى أحياناً (اكتشاف المعرفة) وهي عملية تحليل البيانات من منظورات مختلفة واستخلاص العلاقات فيما بينها وتلخيصها إلى معلومات مفيدة من خلال استكشاف قواعد جديدة من قواعد البيانات الكبيرة وكذلك اكتشاف نماذج (Model) جديدة وصولاً إلى معلومات ذات قيمة وذلك باستعمال مجموعة من الأدوات والتقنيات الحديثة أو الخوارزميات وغيرها من أدوات الإحصاء الاعتيادية والرسوم البيانية. ومن خلال التعريفات السابقة يمكن القول أن تنقيب البيانات (Data Mining) هي عملية اكتشاف المعرفة من قواعد البيانات (Data Base)، وتسمى أحياناً اكتشاف المعرفة (Knowledge discovering from Data- KDD).

٢- أهمية تنقيب البيانات:

أن تنقيب البيانات (DM) أسلوب تمكن من خلاله الوصول إلى المعلومات المخزونة في مستودع البيانات (Data Warehouse) ويتضمن استخدام التحليل الإحصائي (Statistical Analysis) لاكتشاف العلاقات الخفية في البيانات(Romney and Baul, 2009)[32]، كما يعتبر تنقيب البيانات أحد تكنولوجيا الذكاء الاصطناعي (Artificial Intelligent) فضلاً عن الأنظمة الخبيرة والشبكات العصبية، ويهدف إلى تمكين النشاط أو المنظمة من الاستغلال الأمثل لبياناتها فهي تحاول ايجاد المعلومات في مجاميع البيانات الكبيرة التي لا تعلم المنظمة أو النشاط بوجودها، كذلك ايجاد العلاقات وعمل التنبؤات .

ويتضح من ذلك أن أهمية تنقيب البيانات ناجمة عن عملية استكشاف وتحليل كميات كبيرة من البيانات لغرض الحصول على علاقات ونماذج خفية تساعد على استخلاص المعلومات المفيدة والساندة لاتخاذ قرارات عمل استراتيجية كفيلة بزيادة أداء المنظمة أو النشاط.

١- صياغة المتغيرات وتحويلها (Variables Construction and Translation): حيث يجب أن تصاغ المتغيرات الجديدة لبناء النماذج الفعالة التي تفي بالجانب التطبيقي .

٢- تكامل البيانات (Data Integration: اذ إن مجاميع البيانات في دراسة التنقيب عن البيانات من الممكن خزنها في قواعد البيانات متعددة الأغراض والتي تكون بحاجة إلى توحيدها في قاعدة بيانية واحدة.

٣- تصميم وتنسيق البيانات (Data Formating and consisting): في هذه الخطوة إعادة ترتيب حقول البيانات وفقاً لنموذج التنقيب في البيانات.

3- أساليب تنقيب البيانات Techniques of Data Mining[11,23,26]:

تستخدم عملية تنقيب البيانات تقنيات وأساليب عديدة تتمكن من خلالها اكتشاف الاتجاهات والنماذج الخفية في مقادير كبيرة من البيانات ويمكن استخدام واحدة أو أكثر من هذه الأساليب، وهي كالآتي:

اولاً: التصنيف (Classification):

يتم في التصنيف تحليل مجموعة البيانات لتكوين مجموعة من القواعد المتجمعة التي يمكن أن تستخدم لتصنيف بيانات التشغيل إلى مجموعات بحسب صفات معينة، إي: أيجاد المعلومات التي تتعلق بالخصائص المشتركة، وللتصنيف أدوات عديدة مثل:

أولاً: شجرة القرار (Decision Tree).
ثانياً: المجاور الأقرب (Nearest Neighbor).
ثالثاً: الانحدار (Regression).

ثانياً: الاقتران (Association):

وهي القاعدة التي تتضمن علاقات اقتران ثابتة بين مجموعة من الأشياء في قاعدة البيانات، أي الاقتران بين حدوث حدث ما، وحدوث حدث آخر وهي غالباً ما تسمى سلسلة السوق (Market Basket Analysis).

ثالثاً: التحليل المتسلسل (Sequential Analysis):

وهو يشبه الاقتران ويوضع تحت مسمى تحليل الربط (Link Analysis) لكونه مرتبطاً بالزمن، فيبحث عن نماذج تحدث بالتتابع، أي: يتعامل مع البيانات التي تحدث في حالات منفصلة.

رابعاً: التجميع او العنقدة (Clustering):

وهي تقنية تجمع الكيانات المتشابهة سوية وتفصلها عن البيانات غير المتشابهة وفي مجموعات مختلفة، وتعتمد بصورة أساسية على قياس المسافة وتعد تقنية المجاور الأقرب (Nearest Neighbor) شكلاً آخر للتجميع، إذ من الممكن أن تكون هناك مفاتيح مختلفة لاثنين من أدوات التنقيب على البيانات نفسها. يختلف التجميع (Clustering) عن التصنيف (Classification) بأنه في الأول لا تعرف ما ستكون عليه التجمعات عند البدء أو بأية صفة ستتجمع البيانات.

وتستخدم في التجميع أدوات مثل:

أ- متوسط (K means).

ب- الشبكات العصبية (Neural Networks).

٤- أدوات أساليب تنقيب البيانات (Tools of Data Mining Methods)[15,25,27]:

هناك العديد من الأدوات (Tools) التي تستخدم في تقنيات تنقيب البيانات وهذه الأدوات متنوعة ولكل منها دور يخدم غرض معين ومن هذه الأدوات (Tow Crows, 1999: 10-15)[25]:

اولاً: أشجار القرار (Decision Trees):

وهي مشتقة من الإحصاء والذكاء الاصطناعي حيث تستخدم الارتباط في البيانات وتستخدم الاستدلال الإحصائي على قواعد العمل وتعد أساس بناء النموذج التنبؤي كما يمكن أن تستخدم في الشبكات العصبية كمدخلات.

ثانياً: الشبكات العصبية (Neural Networks):

وهي قريبة من أشجار القرار لكنها أصعب فهماً وتقدم نماذج ذات قوة تنبؤية أفضل وتتكون من طبقات المدخلات (Layers) وترتبط بعقد الطبقات المخفية (Hidden Layers) التي ترتبط بدورها بعقد طبقة مخفية أخرى وحتى طبقة المخرجات (Out Put Layer) حيث تضم واحد أو أكثر من المتغيرات التابعة.

ثالثاً: الانحدار (Regression):

يستخدم الانحدار في التنبؤ بالقيم الجديدة بالاعتماد على القيم الموجودة ويستخدم الانحدار الخطي للحالات البسيطة، أما الحالات المعقدة التي يصعب التنبؤ بها يستخدم الانحدار النسبي لأنها تعتمد على تفاعلات معقدة لمتغيرات متعددة.

رابعاً: التنبؤ (Predictor):

وتنبأ بالقيم المستقبلية غير المعروفة بالاعتماد على سلاسل تغير الزمن للمتنبئات، فيؤخذ بالحسبان الخواص المميزة للزمن كتدرج المدد والموسمية.

خامساً: استنتاج القاعدة (Rule Induction):

فيها يتم اشتقاق مجموعة من القواعد المستقلة وعلى خلاف أشجار القرار فهي لا تأتي من شجرة، وقد لا تغطي القواعد الممكنة كل الحالات الممكنة كما أنها قد تتعارض في تنبؤاتها.

سادساً: المجاور الأقرب (K- Nearest Neighbor):

أن فكرة المجاور الأقرب تؤسس على أن حل المشاكل الجديدة يكون عن طريق ملاحظة ومعرفة حلول المشاكل المشابهة والتي تم حلها مسبقاً.

سابعاً: التحليل التمييزي (Discriminate Analysis):

وهي أداة تصنيف تجد السطوح المتعددة التي تفصل الفئات ويكون النموذج الناتج سهل التغيير لأن كل ما على المستخدم هو تحديد جانب الخط الذي تقع عليه النقطة.

ثامناً: الإسناد (Boosting)[33]:

وتعتمد على أخذ عينات عشوائية متعددة من البيانات وبناء نموذج التصنيف لكل منها، ويتغير وضع البناء بالاعتماد على نتيجة النماذج السابقة ويكون التصنيف الأخير هو الفئة الأكثر تخصيصاً من قبل النماذج (Dasid H.A Firet,1996)[33].

تاسعاً: الخوارزميات الجينية (Genetic Algorithms):

سميت بذلك لأنها تتبع نموذج نشوء الإحياء الذي يتنافس فيه أعضاء النشوء الواحد من النماذج لتتقدم خصائصها في النشوء اللاحق من النماذج إلى أن يتم إيجاد النموذج الأفضل. أن المعلومات موجودة في الكروموسومات التي تضم خواص بناء النموذج، فالخوارزميات الجينية تعمل أساساً كطريقة لإنجاز البحث الموجه عن النماذج الجيدة.

٥- مقاييس المسافة [30,7] :

من الممكن أن نبين هنا أنواع البيانات أو المتغيرات والتي تستخدم في الأغلب في التحليل العنقودي وتبيان كيفية إجراء العمليات عليها في عمليات التحليل، على فرض أن هذه البيانات تحتوي على N من الصفات على هيئة أشخاص، بيوت، أقطار، وثائق...الخ

أن خوارزميات العنقدة تعمل على البيانات بالصيغتين الآتيتين:

١) مصفوفة البيانات Data Matrix ويمكن تمثيلها بـ n من الصفات و P من المتغيرات وبناء المصفوفة يكون n×p وكالآتي:

$$\begin{bmatrix} X_{11} \dots X_{if} \dots X_{ip} \\ \vdots \quad \vdots \quad \vdots \\ X_{i1} \dots X_{if} \dots X_{ip} \\ \vdots \quad \vdots \quad \vdots \\ X_{n1} \dots X_{nf} \dots X_{np} \end{bmatrix}$$

٢- مصفوفة غير متماثلة Dissimilarity matrix : وهذه تمثل بـ n من الصفات و n من الصفات ايضاً وبناء المصفوفة يكون n×n وكالآتي:

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & & \\ \vdots & \vdots & 0 & \\ d(n,1) & d(n,2) \dots . & & 0 \end{bmatrix} \qquad (1)$$

اذ أن: i , j ليست سالبة.

d(I , j) يقيس المسافة للصفات المختلفة بين i , j.

إذا كان I =j (أي متشابهان).

أي: ٠ = d(i , j) = d(j, i)

اولاً: المتغيرات المقاسة بالفترة (Interval scale variables):

أي عندما تكون المتغيرات مستمرة وكالآتي:

١- حساب معدل الانحرافات المطلقة يكون:

$$S(f) = \frac{1}{n}\left(|X_{1f} - m_f| + |X_{2f} - m_f| + \dots . |X_{nf} - m_f| \right) \qquad (2)$$

اذ ان M_f معدل قيم المتغيرات

$$M_f = \frac{1}{n}(X_{1f} + X_{2f} + \dots + X_{nf})$$

2- حساب مقياس الانحراف المعياري أو Z-score

$$Zif = \frac{x_{if} - m_f}{s_f} \qquad (3)$$

اذ ان S_f الانحراف المعياري
وأن قياس المسافة الاقليدية لكل زوج من الصفات يكون :

$$d(i,j) = \sqrt{(X_{i1} - X_{j1})^2 + (X_{i2} - X_{j2})^2 + \cdots + (X_{in} - X_{jn})^2} \quad (4)$$

حيث: $i = (X_{i1}, X_{i2}, \ldots, X_{in})$
$J = (X_{j1}, X_{j2}, \ldots, X_{jn})$
وهناك مقياس (Manhattan) لقياس المسافة ويكون بالشكل الاتي:

$$d(i,j) = |X_{i1} - X_{j1}| + |X_{i2} - X_{j2}| + \ldots + |X_{in} - X_{jn}| \quad (5)$$

وكلا المقياسين للمسافة Manhattan و Euclidean يحققان الآتي:

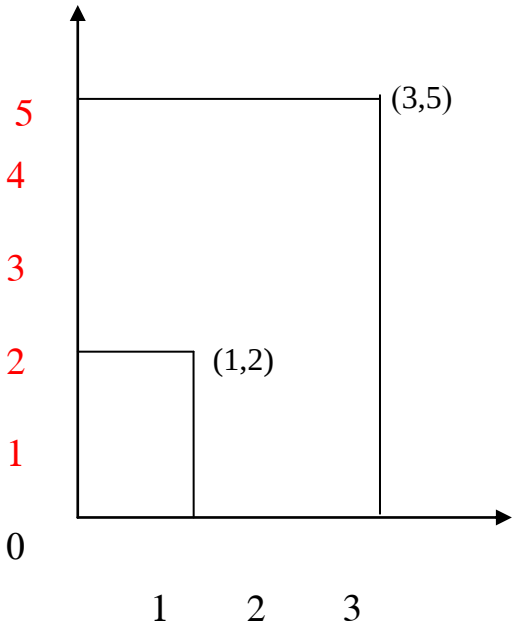
١- ليست سالبة $d(i,j) \geq 0$
٢- المسافة لنفس الصفة تساوي صفر $d(i,j) = 0$
٣- المسافة دالة متماثلة $d(i,j) = d(j,i)$
٤- $d(i,j) \leq d(i,h) + d(h,j)$، حيث h صفة أخرى.

ثانياً: مقياس المتغيرات الثنائية Binary variables:

ويستعمل هذا المقياس لحساب عدم التشابه بين صفات البيانات من خلال قياس المتغيرات الثنائية المتماثلة أو غير المتماثلة وكالآتي:

Euclidean distance = $(2^2+3^2)^{1/2}$ = 3.61
Manhattan distance = (2+3) =5

وإن الشكل (١) الاتي يبين قياس المسافة الاقليدية ومانهاتن (Manhattan) لـ ٢ من الصفات :



الشكل (1-١) يبين قياس المسافة لصفتين في خوارزمية العنقدة

يكون المتغير الثنائي يملك صفتين فقط هما 1 ,0 حيث : 0 يمثل أن المتغير غير موجود، و 1 يمثل أن المتغير حاضر.

ويوجد نوعان من المتحولات الثنائية:

البيانات المتناظرة (Symmetric) $$d(i,j) = \frac{r+s}{q+r+s+t}$$

اذ ان r , s عدد مرات الاختلاف.

و q عدد مرات التشابه (الغرضان يجتمعان بنفس الصفة) .

و t عدد مرات التشابه (الغرضان لا يجتمعان بنفس الصفة) .

البيانات الغير متناظرة (Asymmetric) $$d(i,j) = \frac{r+s}{q+r+s}$$

وبعد تمثيل البيانات نقوم بتشكيل مصفوفة عدم التشابه.

ثالثاً: قياس المتغيرات الرتبوية (Ordinal variables):

وهي تشبه المتحولات المطلقة (Absolute variables) ولكن في هذه الحالة يؤخذ الترتيب بعين العناية مثلاً درجات التقدير (دكتوراه، ماجستير، بكالوريوس، دبلوم، إعدادية) فتعطى الرتب (1، 2، 3، 4، 5) فيصبح لدينا مجال التصنيف { m،....،1} ، حيث m= 5 .

وفيما يأتي خطوات تشكيل المصفوفة:

١- نحول المجال { m , 1} إلى المجال {1 , 0} من خلال العلاقة الاتية:

$$Z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

٢- نقوم بمعالجة المتحولات الناتجة وكأنها متحولات المجال.

٣- نقوم بحساب المسافات.

٤- ثم نشكل مصفوفة عدم التشابه.

اذ ان f تملك M_f حالة مرتبة من M_f ,..., 1

و f تملك i th من الصفات لـ x_{if}

$$r_{if} \in \{1, \dots, m\, f\}$$

٦- توزيع (F)[9] :

اذا كان هناك متغيرات (Q,W) كل منهما يتوزع بصورة مستقلة توزيع (X^2) مربع كاي ، واذا قسم كل منهما على درجة الحرية المناظرة له فان النسبة الناتجة من قسمة احدهما عل الاخر تسمى نسبة التباين ، وان التوزيع لها هو توزيع (F) وفي حالة اختبار تساوي الاوساط فإننا قد اوجدنا نسبة مجموع المربعات بين ومجموع المربعات داخل اي ان صيغة :

$$F = \frac{\frac{1}{t-1}\sum_j n(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2}{\frac{1}{n-t}\sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2}$$

اذ أن البسط والمقام هما تقديران غير متحيزين لتباين المجتمع σ^2 وأن هذه النسبة ايضاً لها توزيع F ويمكن تبسيط هذه النسبة كما يأتي:

$$F = \frac{\text{تقدير التباين بين المجاميع}}{\text{تقدير التباين داخل المجاميع}}$$

وعندما تكون اوساط (المجموعات) غير متساوية فان تقدير التباين من مجموع مربعات (بين) مساوياً الى $\sigma^2 + c$ حيث C> Zero وبالنتيجة فان نسبة التباين الى قيمة F ستكبر ومن ثم سنكون في شك في تساوي الاوساط المجموعات او العينات .
ويمكن صياغة مجموع المربعات الكلية ومجموع المربعات بين المجموعات ومجموع المربعات داخل المجموعات بالشكل الاتي:

$$\Sigma_i \Sigma_j (y_{ij} - \bar{y}..)^2 = \Sigma_i \Sigma_j {y_{ij}}^2 - \frac{(\Sigma\Sigma y_{ij})^2}{N} \qquad (1\text{-}1)$$

$$\sum_i \sum_j (y_{ij} - \bar{y}..)^2 = \sum_i \sum_j {y_{ij}}^2 - cf$$

$$\sum_i \sum_j (\bar{y}_{i\cdot} - \bar{y}..)^2 = \frac{\Sigma {y_{i\cdot}}^2}{ni} - \frac{{y..}^2}{N}$$

$$\Sigma_i \Sigma_j (y_{ij} - \bar{y}_{i\cdot})^2 = \Sigma(\Sigma {y_{ij}}^2 - \frac{{y_{ij\cdot}}^2}{ni}) \qquad (2\text{-}1)$$

ويبين الجدول (1-1) تقدير التباين يبين المجموعات وداخل المجموعات وتقدير F لمصدر التباين(S.O.V):

مصادر التباين (S.o.v)	درجة الحرية DF	مجموع المربعات SS	متوسط المربعات MS	F
تبين المجموعات Between groups	t-1	$SSG = \sum_{i=1}^{t} \frac{{y_{i\cdot}}^2}{ni} - \frac{{y..}^2}{N}$	$MSG = \frac{SSG}{t-1}$	$\frac{MSG}{MSW}$
داخل المجموعات Within groups	N-t	$SSW = \sum \left\| \sum {y_{ij}}^2 - \frac{{y_{i\cdot}}^2}{ni} \right\|$	$MSW = \frac{SSW}{N-t}$	
الكل total	N-1	$\Sigma_i \Sigma_j {y_{ij}}^2 - \frac{(\Sigma\Sigma y_{ij})^2}{N}$		

جدول (1-1) : تحليل التباين لمعيار واحد

6- **خوارزمية** K-Means[31]:

التعريف البسيط لعنقدة K-Means هو تصنيف البيانات إلى مجموعات من الأشياء تؤسس على صفات (Attribute) في K عدد من المجموعات، حيث K عدد صحيح موجب.

أن خوارزمية K-Mean للعنقدة تؤسس على المركز (Center Based) وهو خوارزمية تقنية تقوم بحل مشاكل العنقدة بمراحل (Steps) متعددة.

وتعرف K-Mean بأنها معدل (Centroid) النقاط في المجموعات وتطبق بين الأبعاد (Dimension) ذات المجال (Space) المستمر.

وإن خوارزمية K-mean يظهر لنا أجزاء (مجموعات) ذات كفاية معقولة من التباين لبياناتها ضمن الصف الواحد ومن الممكن تأمين ذلك بالتحليل الرياضي والتجارب العملية، أضف إلى ذلك أن K-Means ينتج برامج سهلة وهي اقتصادية من الناحية الحسابية ولذلك فأنها سهلة في العمليات والبرامج التي تستخدم العينات الكبيرة.

وخوارزمية K-Mean هي أحدى الخوارزميات التي يكون فيها تحليل البيانات يحقق مجموعة بيانات جديدة ومفيدة بسبب بساطة خوارزميتها، كونها متينة نسبياً، كفاية جيدة، تجيب على كل الاستفسارات لمجموعات البيانات المختلفة.

أن تحليل خوارزمية K-Mean في العنقدة تحدد من خلال:

١- تكوين عناقيد مثالية (Clustering membership optimal): أي كل نقطة هي عضو (member) في العنقود وتكون الأقرب إلى المركز.

٢- بلوغ الأمثلية للمحتوى: أي أن كل نقاط العناقيد تحقق الأمثلية إلى المركز (Centroid) بالتشابه والتقارب بينها.

أساسيات خوارزميات K-Means[31,7]:

أن تقنية العنقدة لـ K-Means هي أحدى الخوارزميات البسيطة والتي تتبع عدة مراحل (Steps) وكما في النقاط أدناه :

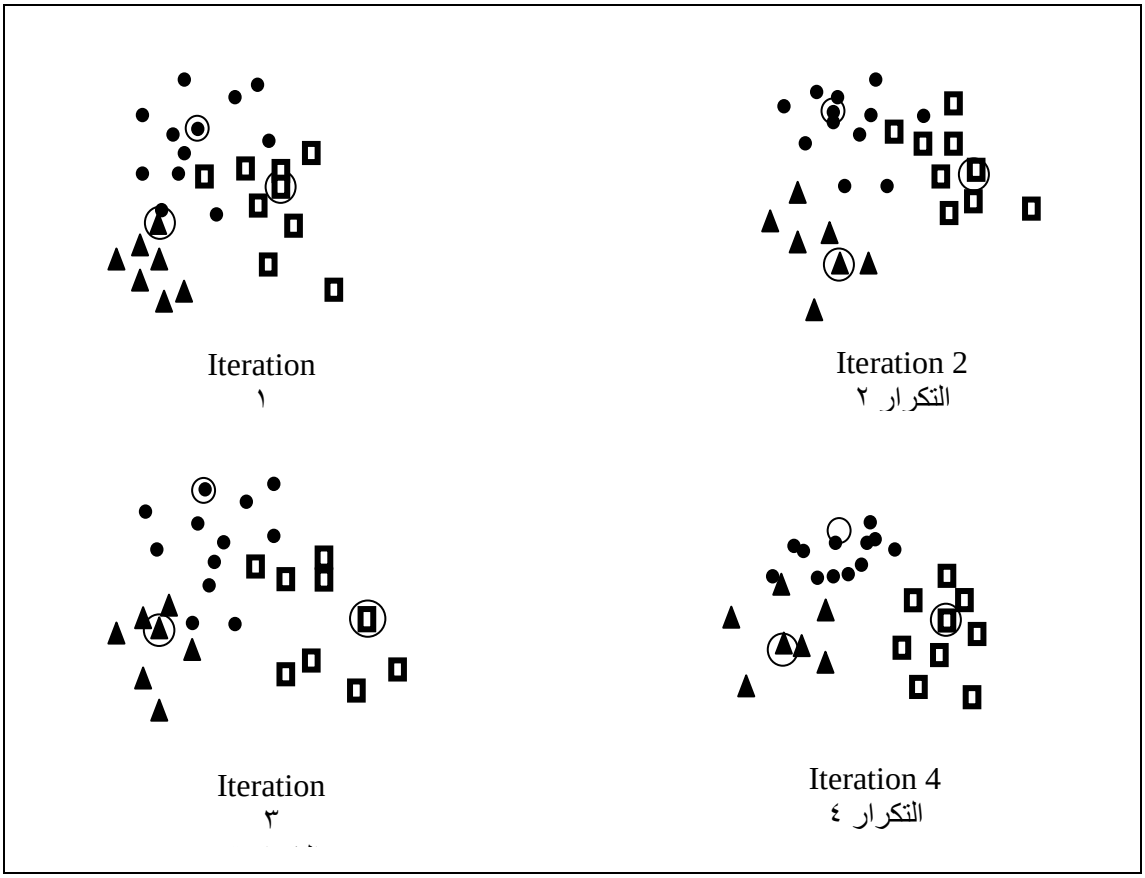
Step 1: اختيار k من النقاط والتي تمثل المراكز (Centroid) الأولية حيث K تمثل أيضاً عدد العناقيد المطلوبة.

Step2: تخصيص النقاط وحسب الصفات (Object) إلى أقرب مركز للعنقود.

Step3: يتم احتساب مراكز العناقيد (K- Centroid) من جديد.

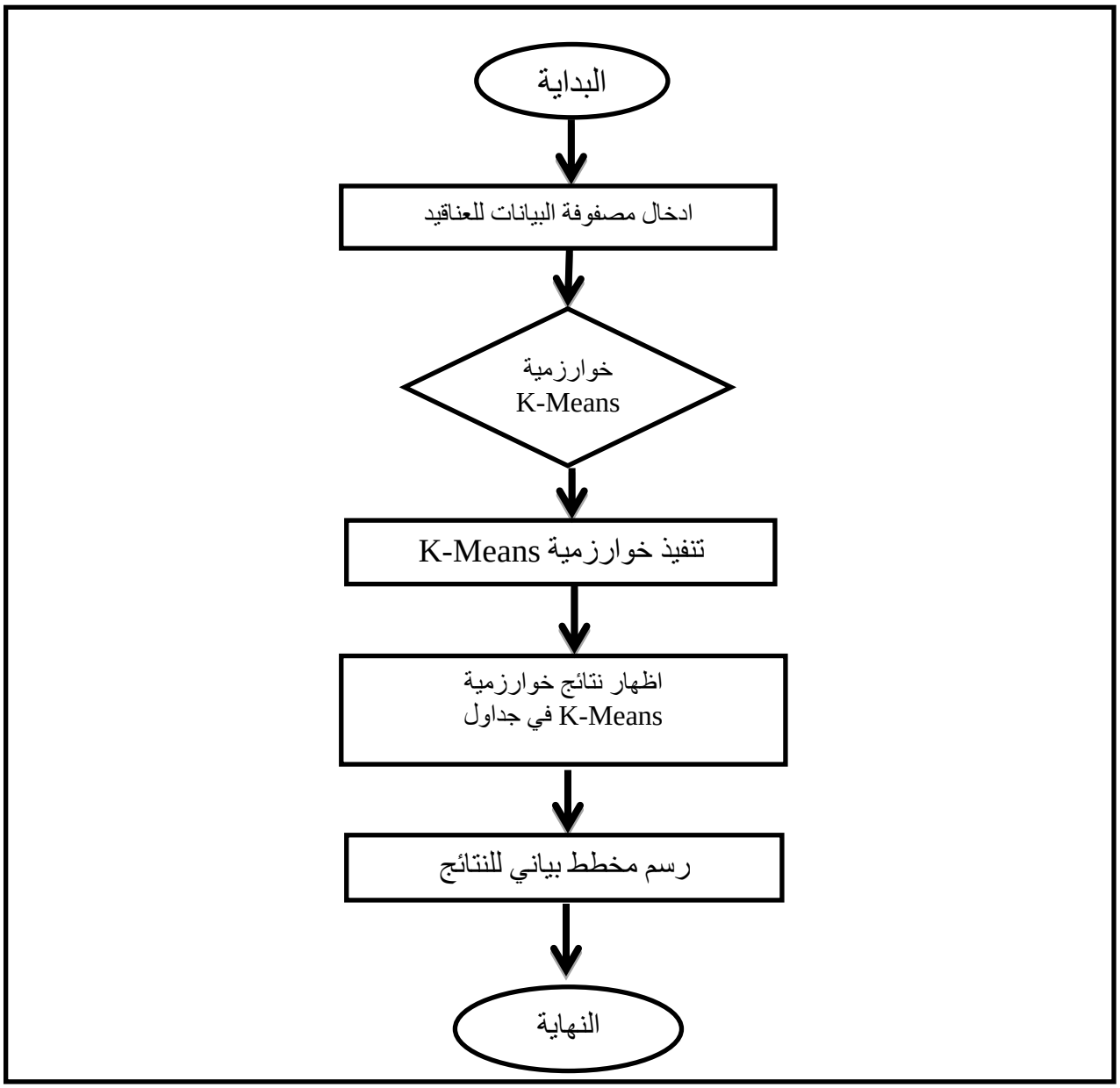
Step4: نعيد الخطوة (2) حتى تكون كل المراكز متماثلة أو متقاربة مع بعضها.

في الشكل (2-1) يبين عمل (K-mean) والذي يوضح عملية العنقدة مع ثلاثة مراكز (Centroid) حيث أن العنقود النهائي يكون رابع تكرار .



الشكل (2-1) يبين استعمال خوارزمية K-Means

ولغرض توضيح المخطط الانسيابي لخوارزمية (K-Means) يبين لنا الشكل (3-1) الخوارزمية وبحسب المخطط الانسيابي لها .



الشكل (3-1) يوضح المخطط الانسيابي لخوارزمية K-Means

المبحث الثاني/ الجانب التطبيقي:

١- المقدمة Introduction:

في الجانب التطبيقي تناولت الدراسة البيانات الخاصة بمتغيرات الدالة والتي تمثل الملاك التدريسي المتغيرات (V) او المشاهدات لتربية محافظة الانبار والمبينة في الملاحق .

تم سحب عينة مكونة من (١٠٠) مشاهدة اي ان حجم العينة (n=100) حيث $X=\{X_1,X_2,\ldots,X_n\}$ وان عدد المتغيرات التي وقع الاختيار عليها من عينة البحث هي m=5 علماً ان كل متغير من هذه المتغيرات V يمثل عدد من الصفات الخاصة بكل من متغيرات البحث للبيانات في الملحق.

في البرنامج التطبيقي تم تطبيق خوارزمية (K-Means) على عينة الدراسة واظهرت النتائج كما في الجدول (1-2) .

٢- المتغيرات والصفات:

اولاً: الجنس : ويتضمن المتغير صفتين هما (ذكر ، وانثى) وقد اعطي لهما الرمزين (٢,١) على الترتيب.

ثانياُ: الحالة العلمية الشهادة: ويشمل متغير الحالة العلمية الصفات (دكتوراه ، ماجستير، بكالوريوس، دبلوم ، اخرى) وقد تم تمثيلها بالرموز (٥,٤,٣,٢,١) على الترتيب.

ثالثاً: التخصص: هذا المتغير يشمل الاختصاصات التالية (اسلامية، تربية رياضية، تاريخ ،جغرافية ، انكليزية، عربية، فيزياء، رياضيات، كيمياء، علوم حياة، فنية) وقد مثلت بالرموز (١١,١٠,٩,٨,٧,٦,٥,٤,٣,٢,١) على الترتيب.

رابعاً: الدرجة الوظيفية : ويشمل متغير الدرجة الوظيفية الدرجات (الاولى ، الثانية ، الثالثة ، الرابعة ، الخامسة، السادسة ، السابعة، الثامنة) وقد تم تمثيلها بالرموز (٨,٧,٦,٥,٤,٣,٢,١) على التوالي.

خامساً: العنوان الوظيفي : ويشمل { مدرس اقدم اول (معلم جامعي اقدم اول)، مدرس اقدم (معلم جامعي اول)، مدرس اول (معلم جامعي)، مدرس ثاني (معلم اقدم اول)، مدرس ثالث (معلم اقدم ثاني)، مدرس رابع (معلم ثاني)، معلم ثالث ، معلم رابع ، معلم خامس ، مدرس، معلم } وتتمثل بالرموز من (١٠,٩,٨,٧,٦,٥,٤,٣,٢,١) على الترتيب.

٣- المقاييس المستخرجة:

تم استخراج بعض المقاييس الاحصائية من جداول البرنامج ومن اهمها:

اولاً: مقياس اختبار F-Test وهذا المقياس كما معلوم يقيس اختبار نسبة تباين مجموعتين او عينتين بدرجتي حرية n-1 لكل منهما اي لدرجتي حرية البسط والمقام، وان مقياس F-Test يبين تشتت او تباين البيانات في العناقيد المختلفة لبيانات العينة او متغيرات البحث.

ويشترط في هذا لمقياس ان تكون الاخطاء مستقلة من مفردة الى اخرى وتتوزع طبيعياً ($ei\sim N(0,\sigma^2)$) اما بالنسبة لصيغة F-test :

$$OR \quad F=\frac{MSC}{MSE} F=\frac{MSR}{MSE}$$

اذ ان $MSR=SSR/(n-1)$ $MSC=SSC/(n-1)$

ثانياً: مقياس MSE (Mean square error)

ان مقياس MSE له اكثر من صيغة وان الصيغة المطبقة في خوارزميات تنقيب البيانات موضوع الدراسة هي :

$$MSE=\frac{\sum_{i=1}^{n}(Xi-\ddot{X})^2}{(n-p)}$$

اذ ان n-p هي درجة حرية الخطأ

P عد الصفوف

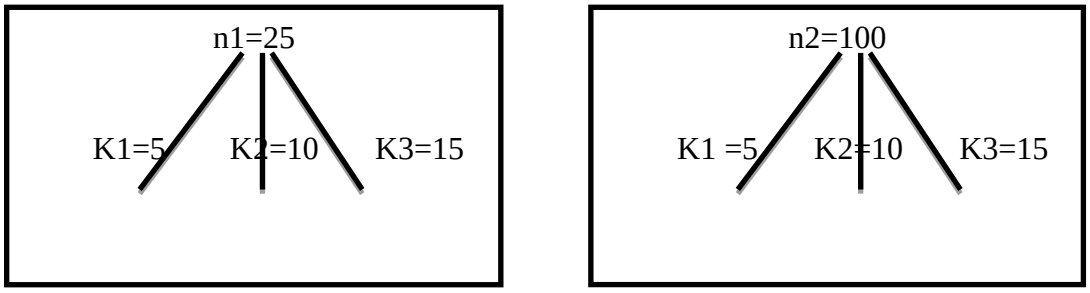
ثالثاً: عدد التكرارات (Iteration History) ويقصد بها عدد مرات اجراء العلميات للخوارزمية لحين الحصول على المتوسط الامثل وكلما كان عدد التكرارات اقل يكون التباين بين البينات اصغر، ومن ثم فان الوصول الى الامثلية يكون بطريق اقصر.

٤- اسلوب العمل:

تم اختيار حجوم عينات تكون ملائمة لعمليات تنقيب البيانات بحيث تتراوح من العينات الصغيرة الى العينات الكبيرة وهي(n1=25، n2=100) وقد اجريت عملية العنقدة وذلك باختيار عدد عناقيد البينات بشكل موازي لحجم العينات اي تكون العناقيد (صغيرة ، متوسط، كبيرة) ايضاً وكالاتي (K3=15,K2=10,K1=5) على الترتيب ولكل عينة، وذلك لملاحظة مدى تأثر عملية عنقدة البينات بكل من حجم البينات وعدد العناقيد.

٥- تحليل النتائج(Analyze the results):

تم تحليل النتائج لخوارزمية (K-Means) وكما فيما يأتي:
تم تطبيق خوارزميات العنقدة (K-Means) وبحسب العناقيد K لكل عينة n وكما فيما يأتي:



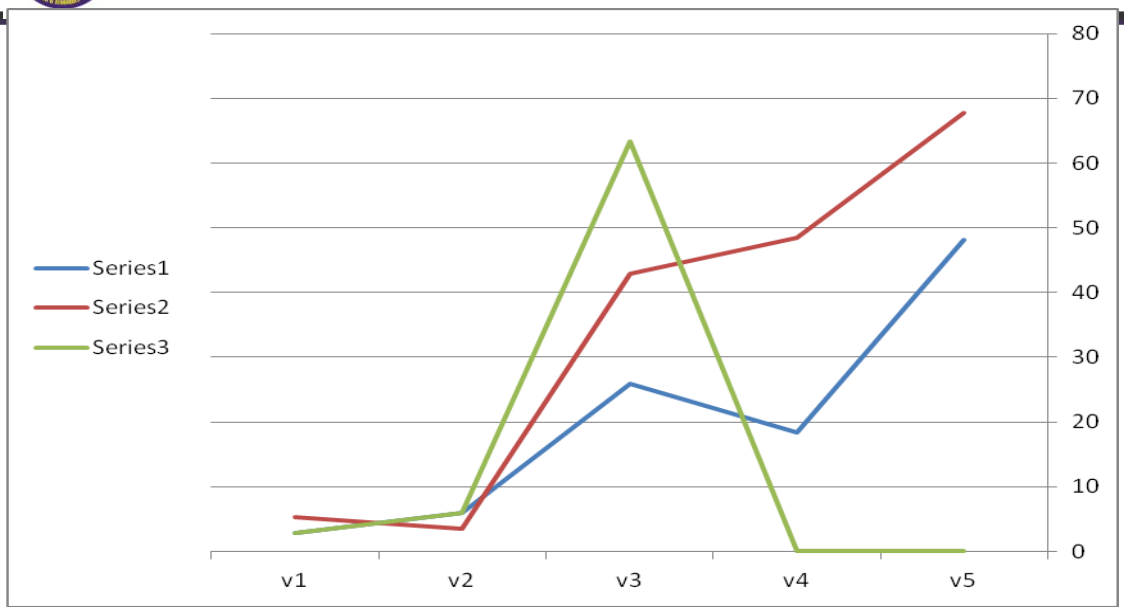
شكل (1-2) يبين توزيع العناقيد (K) على التجربتين

ومن ملاحظة النتائج في جدول التحليل (1-2) الاتي : يتبين لدينا مدى العلاقة بين حجم العنقود (K) لكل عينه (n) من عينتي البحث مع قيمة الاختبار (F-test) ولكل متغير (v) من متغيرات الدراسة :

F-test / حجم العينة	V1	V2	V3	V4	V5
25-5	2.836	5.918	25.928	18.338	48.095
25-10	5.333	3.489	42.924	48.523	67.807
25-15	2.806	5.914	63.351	0.000	0.000
100-5	9.341	45.747	77.670	117.451	161.996
100-10	5.663	936.660	64.152	227.274	280.926
100-15	4.600	24.874	102.294	250.906	334.852

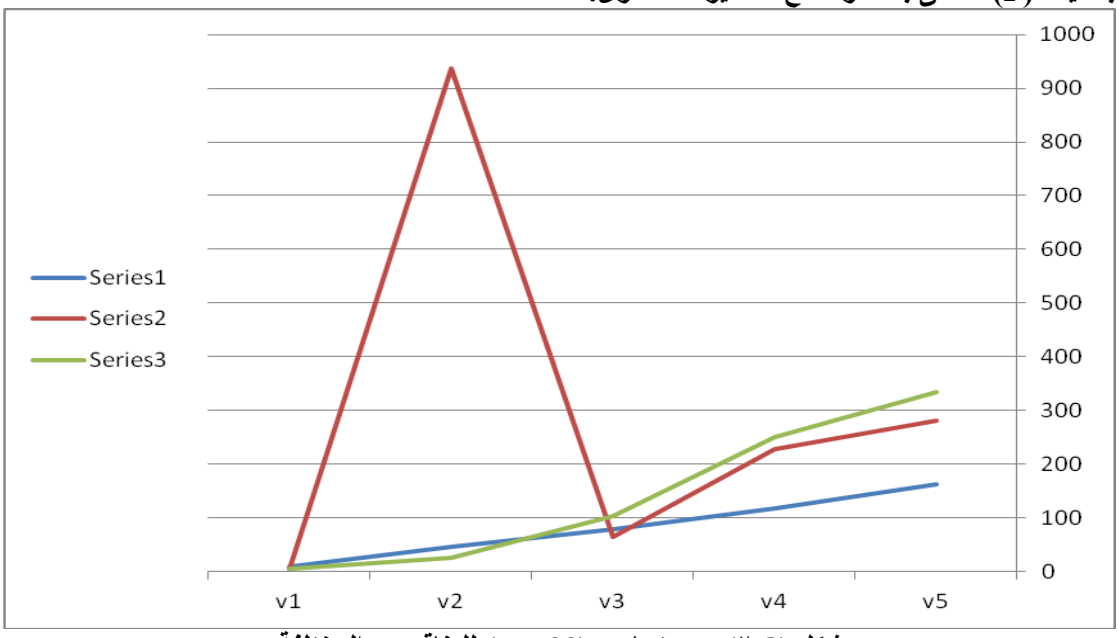
جدول (1-2) يبين قيمة F-Test للمتغيرات لكل عنقود ولحجوم العينات المختلفة

وعند ملاحظة جدول (1-2) يتبين معرفة تأثر F-Test بكل من (n) حجم العينة و (K) حجم العنقود وكذلك ملاحظة مدى تأثر كل من (F,n,K) لكل من المتغيرات الخمسة لأغراض العنقدة من خلال الاشكال المبينة بيانياً فيما يأتي:



شكل (2-٢) يبين تجارب (n=25) للعناقيد K المختلفة

من ملاحظة الشكل السابق تبين لنا ان قيمة (F) تتأثر بكل متغير من المتغيرات المأخوذة في البحث وللتجارب الثلاثة الاولى عند (n=25) ونلاحظ ان قيمة (F) الاعلى تعود للمتغير الخامس (V5) في حين ان قيم (F) الادنى تعود للمتغير الاول (V1) للتجربتين الاولى والثانية اما التجربة الثالثة فقد تميز المتغير الثالث بتقديمه (F) الاعلى بالمقارنة مع المتغيرات الاخرى.



شكل (3-٢) يبين تجارب (n=100) للعناقيد K المختلفة

من ملاحظة الشكل (3-٢) السابق تبين ان قيمة (F) اخذت قيماً متقاربة للتجارب الثلاث عند المتغير V3 في حين انها اخذت قيمة مرتفعة جداً عند المتغير V2 ثم انخفضت قيمة (F) لتتقارب مع بعضها عند المتغير الثالث V3 والمتغير الرابع V4 والمتغير الخامس V5 .

نستنتج من الاشكال السابقة أن قيمة (F) في المتغير الثالث V3 لحجوم العينات للتجارب (n=25) اخذت المتغير الثالث V3 كقيمة اعلى من بقية المتغيرات في حين ان في التجربة (n=100) تكون قيمة (F) هي الاعلى في المتغير الثاني V2 وهذا يفسر ان افضل عنقدة بالنسبة للتجربة (5-25) هي للمتغير V3 في حين ان المتغير V2 كان هو الافضل بالنسبة للتجربة (10-100).

وبالعودة الى النتائج في الجدول (1-2) والجداول السابقة ان خوارزمية العنقدة قد اختلفت بحسب ظروف التجربة المتمثلة بحجم العينة n وعدد العناقيد K وعدد التكرارات بغية الوصول الى العنقدة الامثل فضلاً عن عائدية كل مفردة او عنصر الى كل عنقود وهذه تختلف من مرحلة الى اخرى ومن متغير الى اخر.

الاستنتاجات :

١- تجب ملاحظة مدى تأثر العنقدة الاقل بظروف التجربة (عدد المفردات) فضلا عن عدد العناقيد المطلوبة K .

٢- ان قيم (F) قد تغايرت فيما بينها من الاعلى الى الاقل فنلاحظ ان اعلى قيمة لـ (F) كانت للتجربة (١٠٠-١٠) للمتغير الثاني (V2) والتي تمثل اقل قيمة لـ MSE.

واما بالنسبة للتجربة (٢٥-٥) فان قيمة (F) اخذت اقل قيمة عند المتغير الاول (V١) والتي تكون فيها قيمة MSE اعلى من التجربة السابقة اي ان عنقدة البيانات بالنسبة للمتغير (V2) هي الافضل .

٣- يمكن للوصول الى العنقدة الامثل بالمرور بأعداد مختلفة من التكرارات بغية الوصول الى الامثلية المطلقة وبحسب صفات (Object) متغيرات البحث (V) .

٤- ان خوارزمية (K-Means) اظهرت الاشكال انه عند تغير ظروف التجربة تكون نتائج الخوارزمية متغايرة من حيث النوع (المفردات التي يمتلكها كل عنقود) والعدد (عدد المفردات داخل العنقود).

٥- في جدول خوارزمية (K-Means) المثلى يكون التغاير(F) بين العناقيد المختلفة في اقصى ما يمكن ويكون عندها MSE اقل ما يمكن ، بينما يكون التغاير (F) داخل كل عنقود في اقل ما يمكن.

التوصيات :

1- يمكن اعتماد نتائج خوارزميات العنقدة في التنقيب عن البيانات ولاسيما عندما تكون اعداد البينات كبيرة جداً حيث تتمكن النتائج من الوصول الى المفردة المطلوبة بعدد المرور بالطرق الاقصر لهذه النتائج.

2- بالإمكان زيادة عدد المتغيرات (V) للوصول الى عنقدة أمثل تحقق نتائج بحث وتصنيف في قواعد البيانات بأقصر وقت ولكل متغير من متغيرات البحث.

٣- يمكن استعمال خوارزمية العنقدة مع خوارزمية الذكاء الصناعي (Artificial Intelligence) وهذا ما يسهم في تقليل الجهد والوقت في الوصول الى المعلومة .

٤- يمكن استعمال خوارزميات عنقدة اخرى مثل (Ward) او (Seeded) لأغراض المقارنة والبحث عن الامثلية (Optimization) .

٥- ان تقنيات العنقدة يمكن ان تتعامل مع بيانات الحكومة الالكترونية والذكية وهذا ما يتطلب اعتماد نتائج العنقدة في الوصول الى قواعد البيانات الضخمة والتي من الممكن الدخول من خلالها في دائرة البحث اللامتناهية والتي تحتاج الى ادلة بحث وهذا ما تقدمه نتائج خوارزميات العنقدة .

المصادر العربية :

١- مصطفى؛ فؤاد عبيد؛ "تقنيات التنقيب في قواعد البيانات واستكشاف في المعلومات المخبأة فيها" *1995*.
٢- حسن؛ أ. طيار؛ شلاب ؛ عمار: "التنقيب في البيانات واتخاذ القرارات"، كلية العلوم الاقتصادية، جامعة لوت ، *20*؛*(1995)*.
٣- العلاق؛ بشير عباس: "الادارة الرقمية والتطبيقات"، مركز الامارات للدراسات والبحوث الذاتية *83*، *2005*.
٤-العلي؛ عبد القادر، قنديل؛ عامر ابراهيم، الطمري؛ غسان "المدخل الى دراسة المعرفة وادارة السيرة" عمان، *175 2006*.
٥-السامرائي؛لمياء عبد الصمد؛ " قواعد المعلومات"؛ العراق جامعة بغداد *20* (*1988*).
٦-العمران؛ حمد بن ابراهيم والعبيدي؛ هديل شوكت : "الوعي المعلوماتي والحكمة"، الرياض مكتبة الرشيد ص *88-89* (*2007*).
٧-رحيمة؛ وليد عبد الله : "استخدام التحليل العنقودي وتحليل الانحدار في تشخيص امراض القلب"، الجامعة المستنصرية؛ *1995*.
٨-صديق ؛ رضوان وعبد العزيز ؛ غيداء : " تقييم صحة العنقدة " مجلة الرافدين لعلوم الحاسبات والرياضيات المجلد *5* العدد*2* (*2008*)
٩- المشهداني، كمال علوان : "تصميم وتحليل التجارب-استخدام الحاسوب-" ، مكتب الجزيرة للطباعة والنشر، بغداد (٢٠١٠).

المصادر الاجنبية:

١٠-Houstion Andrea L.& others, “Medical Data Mining” on the internet Research on acancer Information system Artifical intelligence Review 2000.
١١-Edelstein, Herb, “Mining for gold” information week. April, 1997.
١٢- Remach an Ran M.puspa, “Mining for Gold wipro technologies” 2001
١٣- Luan Jing, “Data Mining And knowledge Management”inHigher Eduction Air Canada 2002.
١٤- Ahola Jussi & Rinta-Runsala Esa; “Data Mining case studies” in customer profiling 2001.
١٥- spitiopoulou, Myra & phole carsten: “Data Mining to measure and Improve the success of web sites” 2002.
١٦- Friedman I.H “Data Mining and Statistics” what in the connection the 29th sum passion on the information couputing science and Statistics, key note speech Houston (1997).
١٧- Little R.Rubin D, “statistical analysis with missing data widely”: New York (1987).
١٨- Hoes J.kauper M. “Data mining concept and technique” morgan kaufrnonn, san Francisco .A (2006).
١٩- Efrom B, Boot strap method Another Look at the Jacknife Amals of statistics 7, 1-26 (1979).
٢٠- Efrom B,cous G: Alies Lrely Look at the Boost strap the Jackife and cress-ualidalion American statistic 37,36-48 (1988).
٢١- Fland, D. uanni laH. Smuth R. “principle of data Mining” Mit press London 2001.

٢٢- source, Ranach and ran M. psuhpa, "Mining for gold wipro Technologies" Decemper 2001.
٢٣- Ramch M. push: "Data Mining for gold" wipro technologies 2001.
٢٤- Soni, Sanjay and Tang zhohui and young Jim performance study of Microsoft "Data Mining Algorithins Microsoft" 2001.
٢٥-Two crows corporation, "Introduction to Data Mining and knowledge discovery" third Edition 1999.
26- Atre shaka, Detining todays "Data mining" Excutive update Business intelligence Advisory service 2001.
27- Brand Estelle & Gerristen, Rob "Data Mining" solution Magazine 1998.
28- "Discovering knowledge in Data Mining" By Daniel T.Larose 2005.
29- "Data Mining" k- Dustering problem By Eham karousi university of Agder 2002.
30- "Data Mining concept and techniques" second Edition By Jiawi Han university of lionois at urbanchampaign Michline kamber.
31-Halkid, y.batistakis M.vozor giannis "clustering Algorithm and usability measures 2001.
32- Romney_ Marshall B. and Baul John Standart "Accounting information system 9th ed" .upper saddle river prentice Hall (2009).
33- Dasid H.A Firet occurrence of common Jerm in mathematical statistic the American, 121-133 199 ٦.

User (K-Means) for clustering in Data Mining with application

The great scientific progress has led to widespread Information as information accumulates in large databases is important in trying to revise and compile this vast amount of data and, where its purpose to extract hidden information or classified data under their relations with each other in order to take advantage of them for technical purposes.

And work with data mining (DM) is appropriate in this area because of the importance of research in the (K-Means) algorithm for clustering data in fact applied with effect can be observed in variables by changing the sample size (n) and the number of clusters (K) and their impact on the process of clustering in the algorithm.

Key words :object ,data mining, clustering ,machine learning ,algorithm