المجلة العراقية للعلوم الاحصائية (18) 2010
ص ص [297-320]

استخدام العنقدة في نمذجة سلم الرواتب الشهري

رنا وليد بهنام هندوش*

الملخص

تعد العنقدة واحدا من التطورات العلمية التي توصل إليها العلماء في مجال المعرفة والتقنيات الحديثة لاكتشاف المجاميع المتعددة . قدم مفهوم العنقدة لأول مرة من قبل العالم Ronald في عام 1955 . وان الفكرة الأساسية للعنقدة تتمثل بتجزئة البيانات إلى عناقيد. يهدف هذا البحث إلى استخدام مفهوم العنقدة لنمذجة البيانـات الواقعية لسلم الرواتب الشهري للملاك التدريسي لإحدى كليات جامعـة الموصـل ولعام 2009 , وتطبيق خوارزمية HCM على هذه البيانات . وقد تم استخدام الـ matlab أداة برمجية لكتابة برامجيات الخوارزمية . وأثبتت النتائج مدى كفاءة هذه الخوارزمية في عنقدة البيانات الحقيقية وكيفية تمثيلها بشكل عناقيد .

Using Clustering for Modeling Monthly Salary Grade

ABSTRUCT

Clustering is considered as one of the most scientifical developments which the scientists reached at in the field of recent knowledge and technologies to discover the cluster's group. The clustering concept was introduced firstly by Ronald in 1955. The clustering's fundamental notion is represented in dividing the data into clusters. This research aims to using clustering for actual data modeling for the monthly salary grade of the teaching staff for one of the Mosul University's College in

---

* مدرس مساعد/ قسم البرمجيات/ كلية علوم الحاسوب والرياضيات/جامعة الموصل

تاريخ التسلم :29 /6 /2009 ـــــــــــــــــــــــــــ تاريخ القبول : 10/12 /2009

2009, by using HCM algorithm to these data. Matlab software is used to write down the proposed algorithm programs. Results proved the efficiency of this algorithm in clustering the actual data and how to represent them as clusters .

1. مقدمة :

تعد العنقدة واحدا من نتاج التطور العلمي في مجـال تـصنيف البيانـات وتبويبها , فهي من التقنيات الشائعة والمفيدة في اكتشاف المجاميع المتعددة وتعريف الأنماط والتوزيعات المهمة في البيانات الخام. وتعود بدايات العنقدة إلـى النـصف الأول من القرن الماضي اذ وضع أسسها الأولـى العـالم البريطـاني Ronard fisher. إن الفكرة الأساسية للعنقدة تتلخص بتجزئة مجموعة البيانات إلى عناقيد , اذ إن نقاط البيانات للعنقود الواحد تتشابه أكثر مع بعضها البعض مقارنة مع تلـك النقاط في العناقيد الأخرى ((Basu et. al(2008). فقد كانت طرائق التصنيف سابقاً تعتمد على خبرة الخبراء وعلى الجهد الشخصي الكبير في هذا المجال لتؤدي إلـى نتائج دقيقة, ولكن التوسع العلمي الكبير في مجال الحاسوب جعل عملية الـسيطرة على المعلومات الكبيرة أمر سهل وبكلفة وجهد اقل مما جعـل العلمـاء يـستغلون الإمكانات الجبارة لهذا الجهاز العظيم لمصلحة اختصاصاتهم .

إن العنقدة يمكن تعريفها بأنها عبارة عن عملية تحليـل إحـصائي متعـدد المتغيرات ومبرمج ويتم من خلاله دراسة التشابه بين مجموعـة مـن المتغيـرات لنماذج مختلفة ومتعددة ثم مقارنة هذه النماذج مع بعضها البعض بالاعتمـاد علـى ماتحتويه من متغيرات وترتيب العلاقة مع بعضها بشكل عناقيـد, فيمكننـا إيجـاد التشابه بين المتغيرات المختلفة بعد إجراء المقارنة وترتيبها بشكل عناقيد بالاعتماد على مدى تواجدها في النماذج المختلفة((Lu et. al(2007). ويمكن أن يطلق علـى العنقدة مصطلحات عديدة حسب التخصص والمجال الذي تدرس فيه فمثلا تـسمى بالتجزئة في نظرية البيانات, التعلم غير الموجه في تمييز الأنماط, علم التـصنيف العددي في علوم الحياة والبيئة وكذلك يطلق عليها علم النماذج الشخصية في علـم الاجتماع.

2. تطبيقات العنقدة :

تستخدم العنقدة في تطبيقات عديدة وفي مجالات مختلفة مثل الطب, الهندسة, علم الالكترونيات, علوم الحياة, علم الإنسان, علم الآثار القديمة وأيضا في السوق والاقتصاد. وتتضمن تطبيقات العنقدة الحيوان والنبات, وتصنيف الأمراض, ومعالجة الصور, تمييز الأنماط واسترجاع المستندات وغيرها ((Fred et. al(2002)) ((Figueiredo et. al(2006)) ., ومن هذا كله يمكن تحديد أربعة اتجاهات رئيسية تستخدم فيها العنقدة وهي :

1. توليد الفرضيات: تستخدم العنقدة في هذه الحالة لتستنتج بعض الفرضيات المتعلقة بالبيانات.
2. اختبار الفرضيات: يتم استخدام العنقدة لأجل التأكد من صحة الفرضيات المحددة, ولكي نتحقق هل إن الفرضية صحيحة يتم تطبيق العنقدة على مجموعة نموذجية من البيانات.
3. التنبؤ المعتمد على المجاميع: إن العنقدة يتم تطبيقها على مجموعة من البيانات ونتائج العناقيد يتم تمييزها بواسطة تمييز الأنماط العائدة لتلك العناقيد (Anil et. al(2008)). لذا فان الأنماط غير المعروفة نستطيع تصنيفها إلى عناقيد خاصة بالاعتماد على تشابه العناقيد, وبهذا نستطيع استخلاص المعرفة المتعلقة بالبيانات .
4. تحليل البيانات: إن العنقدة يمكن أن تساعدنا في تقليل البيانات فعندما تكون البيانات المتوفرة كبيرة جدا سنحتاج إلى معالجة ويمكن باستخدام العنقدة تجزئة مجموعة البيانات إلى عدد من العناقيد المرغوب فيها ((Fred et. al(2002)).

3. الخصائص الأساسية للعنقدة :

هناك العديد من الخصائص المرغوب فيها في تكوين العناقيد بواسطة حل مسالة العنقدة واهم هذه الخصائص هي أن الزوج ضمن عنقود واحد يكون مشابها للأزواج ضمن ذلك العنقود أكثر من أن يكون مشابها للأزواج الخارجة عن ذلك العنقود, ويمكن إجمال أهم الخصائص الأساسية للعنقدة بالنقاط الاتية (Banerjee et. al(2007a)):

1. إن عملية العنقدة تعتبر من التعليم غير الموجه وذلك لأنه خلال عملية العنقـدة لاتوجد هناك معلومات مسبقة متعلقة بالصفات المميزة لكل عنقود فضلا عـن ذلك لاتوجد أمثلة تعرض نوع العلاقة المرغوب فيها التـي يجـب أن تكـون صحيحة بين البيانات .
2. تعتمد العنقدة على مفهوم المجاميع اذ أنها تحاول تقسيم مجموعة من البيانـات على مجاميع متشابهة فيما بينها ومختلفة عن باقي العناقيد .
3. عند الانتهاء من عملية العنقدة وتكوين مجموعة من العناقيد نلاحظ انـه مـن الممكن ألا يكون كل عنقود واضحاً وفي هذه الحالة سنحتاج إلى خبير ميـداني لتفسير كل عنقود اذ لايوجد أية معرفة مسبقة متعلقة بالعناقيد (Batistakis et. al(2001)).
4. إن تكوين العناقيد يعد اكتشافا اقرب علاقة بين النماذج والمتغيـرات ونتائجهـا عادة تكون حركية.
5. من الممكن أن يكون هناك العديد من الحلول لمسالة العنقدة الواحدة فلا يكـون من السهل تحديد عدد العناقيد المطلوبة, أي أن العدد الأمثل من العناقيد لايكون معروف عادة فليس هناك حل واحد صحيح لهذه المـسالة (Bekkerman et. al(2005)).
6. قبل تطبيق العنقدة على البيانات نحتاج إلى معالجة مسبقة للبيانات لنتمكن مـن تحليل المعلومات أكثر .

4. أصناف خوارزميات العنقدة :

يمكن تصنيف خوارزميات العنقدة وفقاً لنوعية البيانات المدخلة للخوارزميات ومعيار العنقدة الذي يعرف التشابه بين نقاط البيانات فضلا عن المفاهيم الأساسـية والنظرية المعتمدة عليها تقنيـات العنقـدة ((Ajmera et. al(2003).كمـا يمكننـا تصنيف هذه الخوارزميات حسب نوع المتغيرات الموجودة في البيانات وكما ياتي:

1. العنقدة الحادة اوالواضحة: يفترض فيها أن نقطة البيانات إما ان تنتمـي إلـى العنقود أو لاتنتمي واغلب خوارزميات العنقدة التي تنتج عناقيد هـشة يمكـن تصنيفها إلى العنقدة الحادة .

2. العنقدة المضببة: تفترض انه يمكن تصنيف النماذج إلى أكثر من عنقود وتستخدم فيها تقنيات مضببة لبيانات العنقود وفي هذا النوع نحصل على مخططات عنقدة متوافقة مع حياتنا اليومية وتتعامل مع بيانات حقيقية وغير مؤكدة, ومن أهم خوارزميات هذا النوع (Basu FCM) , ((Anil et. al(2008)) et. al(2008)).
3. العنقدة الإدراكية: وتستخدم لعنقدة البيانات الصنفية , اذ تتعنقد النماذج وفقاً للمفاهيم التي تحملها وليس تبعاً للقيم .
4. العنقدة الإحصائية: هذا النوع من العنقدة يعتمد على مفاهيم التحليل الإحصائي, اذ تستخدم مقاييس التشابه لتجزئة النماذج والتي تكون محددة ببيانات عددية (Lu et.al(2007)).
5. عنقدة شبكة كوهين: تعتمد العنقدة في هذا النوع على مفاهيم الشبكة العصبية. ويكون لهذه الشبكة عقد إدخال وإخراج , اذ إن عقد الإدخال تكون مرتبطة مع كل عقدة في الإخراج وكل ارتباط يكون مجهزا بوزن يحدد موقع مايماثله في عنقدة الإخراج , وان العقد الخارجية ستشكل العناقيد ((Witold et. al(2005)).

إن العديد من خوارزميات العنقدة تم تطويرها من خلال علوم الحواسيب والمواضيع المتعلقة بها مثل تعلم الماكنة , تمييز الأنماط , تنقيب البيانات وعلم الإحصاء. ويمكن تصنيف خوارزميات العنقدة بصورة شاملة وعامة إلى الأصناف الاتية((Batistakis et. al(2001)):

أولاً / العنقدة الهرمية :

إن خوارزميات العنقدة الهرمية تكون مجموعة من العناقيد المتداخلة وتختلف هذه الخوارزميات في كيفية تكوين هذه المجاميع من العناقيد ولتوضيح هذه التقنية والمجاميع المختلفة من العناقيد يمكن استخدام شجرة العناقيد التي تسمى Dendogram والتي تعرض العناقيد, فكل مستوى من الشجرة الهرمية يمثل

تجزئة مجموعة من البيانات إلى مجموعة من العناقيد ففي المستوى الأعلى من المخطط والذي يمثل الجذر يكون كل عنصر من العناصر منتميا إلى عنقود واحد فقط بينما في المستوى الأدنى من مخطط الشجرة والذي يكون بمثابة الأوراق يكون كل عنصر له عنقود وحيد لوحده في حين كل مستوى من مخطط الشجرة يحوي على عقد داخلية والتي تمثل مجموعة من العناقيد المنفصلة. ويمكن تقسيم خوارزميات العنقدة الهرمية تبعاً لطريقة إنتاج العناقيد فمنها خوارزميات التكتيل وتُنتَج العناقيد هنا بأسلوب من الأسفل إلى الأعلى عن طريق تكوين سلسلة من مخططات العنقدة وتبدأ بعنصر لكل عنقود ثم ندمج العناقيد إلى أن تصبح كل العناصر في عنقود واحد وتختلف هذه الخوارزميات في كيفية دمج العناقيد في كل مستويا ((Ajmera et. al(2003)). والطريقة الأخرى تتمثل بـ ـ خوارزميات التقسيم والتي تنتج العناقيد بأسلوب من الأعلى إلى الأسفل بتكوين سلسلة من مخططات العنقدة وتزداد عدد العناقيد في كل خطوة, اذ تبدأ بوضع كل العناصر بعنقود واحد ويتم فصل العنقود الواحد إلى اثنين عدة مرات إلى أن يصبح كل عنصر في عنقود لوحده. وان العنقدة الهرمية تحتاج إلى تحديد المسافة بين العناقيد ففي كل مستوى من الشجرة يكون مشتركا بمقاييس المسافة الذي يستخدم لدمج العناقيد ((Basu et. al(2008)) .

ثانياً / خوارزميات العنقدة الجزيئية :

يتم في هذه الخوارزميات مباشرة تفكيك مجموعة البيانات إلى مجموعة من العناقيد المنفصلة. وهذا يعني تكوين مجموعة واحدة فقط من العناقيد في خطوة واحدة وليس في عدة خطوات لذا يجب على المستخدم إدخال عدد العناقيد المطلوبة. ومن أشهر الخوارزميات التي تقع ضمن هذا الصنف هي خوارزمية HCM (والتي تعد الخوارزمية التي تستخدم في تطبيق هذا البحث) فضلا عن العنقدة مع الخوارزميات الجينية والعنقدة مع الشبكات العصبية وغيرها ((Anil et. al(2008)) .

ثالثاً / خوارزميات عنقدة قواعد البيانات الكبيرة :

إن تقنيات العنقدة يجب أن تكون قادرة على التكيـف مـع تغيـرات قاعـدة البيانات وهذا ما ساعد على ظهور خوارزميات عنقدة حديثة أكثر مثل خوارزميـة GK, وتهدف هذه الخوارزميات إلى قواعد بيانات اكبر يمكنهـا أن تتكيـف مـع تقيدات الذاكرة .

رابعاً / خوارزميات العنقدة مع الصفات الصنفية :

إن خوارزميات العنقدة التقليدية لاتعمل دائما مع البيانـات الـصنفية لـذلك ظهرت خوارزميات حديثة تنظر في البيانات الصنفية.

خامساً / خوارزميات العنقدة المعتمدة على الكثافة :

إن الفكرة الأساسية لهذا النوع هو تجميع النماذج المتجاورة لمجموعـة مـن البيانات إلى عناقيد بالاعتماد على شروط الكثافة فالكائنـات ذات الكثافـة العاليـة سوف تشكل عناقيد في فضاء البيانات أما الكائنات ذات الكثافة الواطئـة فـسوف تكون منفصولة عن العناقيد في تلك البيانات ((Bekkerman et. al(2005).

سادساً / خوارزميات العنقدة المعتمدة على قاعدة الشبكة المتعامدة :

إن الميزة الأساسية لهذا النوع هو تحديد الفضاء بعدد محدد من الخلايا ومـن ثم إجراء العمليات في هذا الفضاء, ويستخدم هذا النوع من الخوارزميات في تنقيب البيانات المكانية .

سابعاً / خوارزميات العنقدة المضببة :

إن كل أنواع الخوارزميات التي ذكرناها سابقا تنتج عناقيد هشة, اذ إن نقطة البيانات إما ان تنتمي أو لا وبهذا تكون العناقيد غير متداخلة, أما في هـذا النـوع فتستخدم مفاهيم المنطق المضبب التي تصف فيها النتائج على شكل عناقيد مـضببة وهذا يعني أن نقطة البيانات ستنتمي إلى عدة عناقيـد فـي آنٍ واحـد وبـدرجات

عضوية مختلفة. واهم خوارزمية تعود الى هذا الصنف هـي خوارزميـة FCM . (Banerjee et. al(2007a)) , (Batistakis et. al(2001))

5. الجانب التطبيقي :

تتضمن هذه الفقرة الجانب العملي للبحث والمتمثل بالخوارزمية المـستخدمة وخطواتها ومخططها الانسيابي وكيفية برمجتها وبناء الدوال الخاصة بها ومراحـل تنفيذ البرنامج وعنقدة البيانات الواقعية لسلم الرواتب الشهري لإحدى كليات جامعة الموصل وكما موضح بالتفصيل في الملحق A, اذ تم بناء هذا النظام اعتماداً علـى لغةmatlab7.0 , وقد استخدمت الخوارزمية المقترحة لعنقدة البيانات.

1.5 الخوارزمية المستخدمة في نمذجة سلم الرواتب الشهري :

هناك العديد من الخوارزميات الشائعة في العنقدة ولكن سوف يقتصر عملنـا في هذا البحث على خوارزمية HCM, وان الهدف الرئيسي في هذه الخوارزميـة هو تحديد كل نقطة في فضاء البيانات إلى عنقود بالاعتماد علـى تجزئـة فـضاء البيانات X إلى مجموعة من العناقيد C_i, فلو فرضنا أن

$$X = \{x_1, x_2, ..., x_n\}$$

وان cc هي عدد العناقيد وتحقق الشرط الاتي:

$$\bigcup_{i=1}^{cc} C_i = X \qquad , \quad 2 \le cc \le n \qquad (1)$$

فعند $cc=n$ سيتم وضع كل نقطة بيانات لوحدها في عنقود, *وفي* حالـة $cc=1$ سيتم وضع كل نقاط البيانات في العنقود نفسه, ومن الواضح انه لايوجد أي تداخل بين أي زوج من هذه العناقيد أي أن :

$$C_i \cap C_J = \varphi \qquad \forall\, i \neq j$$
$$0 \subset C_i \subset X \qquad \forall\, i$$

والشرط الأخير يعني أن العنقود لا يمكن أن يكون فارغاً وهو في الوقت نفسه مجموعة جزئية من X . وفي هذه الخوارزمية سيتم إدخال عدد العناقيد المطلوبة منذ البداية لأنها تحتاج إلى تحديد مسبق لعدد العناقيد وتهدف هذه الخوارزمية الى ايجاد المركز لكل عنقود وتقليل دالة عدم التشابه (دالة الهدفobj) لقياس المسافة ويتم اختيار المسافة الاقليدية دالة للمسافة, اذ أن هذه الدالة تعتمد على المسافة $d(x_n - c_i)$ بين نقطة البيانات x_n ومركز العنقود c_i ويمكن التعبير عن دالة الهدف بالمعادلة الاتية ((Tsuda et. al(2006)):

$$obj = \sum_{i=1}^{cc} (obj)_i = \sum_{i=1}^{cc} \sum_{x_n \in c_i} d(x_n - c_i) \qquad (2)$$

اذ إن $(obj)_i$ هي دالة الهدف داخل العنقود i و $d(x_k - c_i)$ تمثل المسافة الاقليدية والتي تحسب بأنها الجذر ألتربيعي لمجموع تربيعات الفروق الحسابية للإحداثيات المقابلة لنقطتين وكما في المعادلة الاتية:

$$d(x_n - c_i) = \|x_n - c_i\| = \sqrt{\sum_{j=1}^{n} (x_{nj} - c_{ij})^2} \qquad (3)$$

ويمكن تعريف العناقيد المجزأة بالمصفوفة الثنائية $U_{cc \times n}$ والتي يطلق عليها مصفوفة العضوية, اذ أن كل مدخل في المصفوفة U يشار إليه بـ u_{in} والمعبر عنها كما ياتي:

$$u_{ik} = \begin{cases} 1 & if \ \|x_n - c_i\|^2 \le \|x_n - c_j\|^2 \quad \forall \ j \ne i \\ 0 & otherwise \end{cases} \qquad (4)$$

وإن الرمز $\|\ \|$ يرمز للمسافة الاقليدية والتي تحسب كما في المعادلة *(3)* .

<u>ملاحظة:</u> تتميز مصفوفة العضوية بالخاصيتين الآتيتين :

$$\sum_{i=1}^{c} u_{ik} = 1 \qquad \forall \ k = 1,2,.....,n$$

$$\sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik} = n \qquad (6)$$

إن المركز الأمثل c_i والذي يمثل المركز المتوسط Centerior يكون هو المعدل لجميع نقاط البيانات في العنقود i وكما ياتي:

$$c_i = \frac{1}{|C_i|} \sum_{n, x_k \in c_i} x_n \qquad (7)$$

اذ إن نقطة البيانات x_n تنتمي الى العنقود i إذا كان المركز c_i هو المركز الأقرب من بين كل المراكز, وإن $|C_i|$ تمثل عدد نقاط البيانات في العنقود C_i .

إن المدخلات لهذه الخوارزمية هو تحديد عدد العناقيد المطلوبة وإدخال البيانات على شكل متجهات وكذلك إدخال قيمة الخطأ ((Anil et. al(2008). أما المخرجات فهي العناقيد, مصفوفة المسافة, مصفوفة العضوية, مصفوفة المراكز, ***sum-dis*** , دالة الهدف .

2.5 خطوات خوارزمية HCM :

Step1: يتم انتقاء مراكز العناقيد الأولية C_i بصورة عشوائية من بين المشاهدات x_n وبالاعتماد على عدد العناقيد cc المختار.

Step2: حساب مصفوفة المسافة الاقليدية D و مصفوفة العضوية U المعتمدة على عناصر المصفوفة D .

Step3: حساب دالة الهدف obj ويتم التوقف في حالة الحصول على فرق عن التكرار السابق بمقدار اقل من قيمة الخطأ المحددة e .

Step4: حساب مراكز العناقيد الجديدة c_i ثم نعود إلى Step2 .

إن خوارزمية HCM تعتبر خوارزمية تكرارية ولا تضمن أن تصل الى الحل الأمثل والشكل(1) يوضح المخطط الانسيابي لخوارزمية HCM في العنقدة((Dhillon et.al(2004).

3.5 تصميم خوارزمية HCM وبرمجتها :

تمت برمجة الخوارزمية HCM عن طريق بناء دوال برمجية باستخدام لغة Matlab وفيما ياتي شرح موجز لكل دالة من هذه الدوال المبرمجة:

دالة R-I-C :

إن هذه الدالة مختصر للكلمات (Randomly Initialize Centroids). اذ إن المدخلات لهذه الدالة هي عدد العناقيد, مصفوفة البيانات. أما المخرجات فهي مصفوفة مراكز العناقيد الأولية C_i المختارة بصورة عشوائية .

دالة Class-R-Q :

تم إنشاء هذه الدالة لتستلم دالة مصفوفة البيانات كإدخال لها ثم تسال المستخدم بتحديد النمط المطلوب استخدامه, فعند الإجابة بـ "Q" سوف يتم عمل transpose للبيانات لكي تتم عنقدة البيانات على أساس المشاهدات كما في محور بحثنا, في حين عند إدخال الحرف "R" سوف تبقى البيانات كما هي لكي تحصل العنقدة لهذه البيانات على أساس المتغيرات.

دالة Cal - D :

تم بناء هذه الدالة لغرض حساب المسافة الاقليدية, والمدخلات لهذه الدالة هي مصفوفتا البيانات والمراكز, أما المخرجات فهي مصفوفة المسافة D. وتقوم هذه الدالة بإيجاد مربع فروقات المسافة بين نقاط البيانات x_n لمصفوفتي البيانات والمراكز ثم تكوين مصفوفة المسافة وذلك بإيجاد الجذر التربيعي لحاصل جمع مربع الفروقات .

دالة Cal - U :

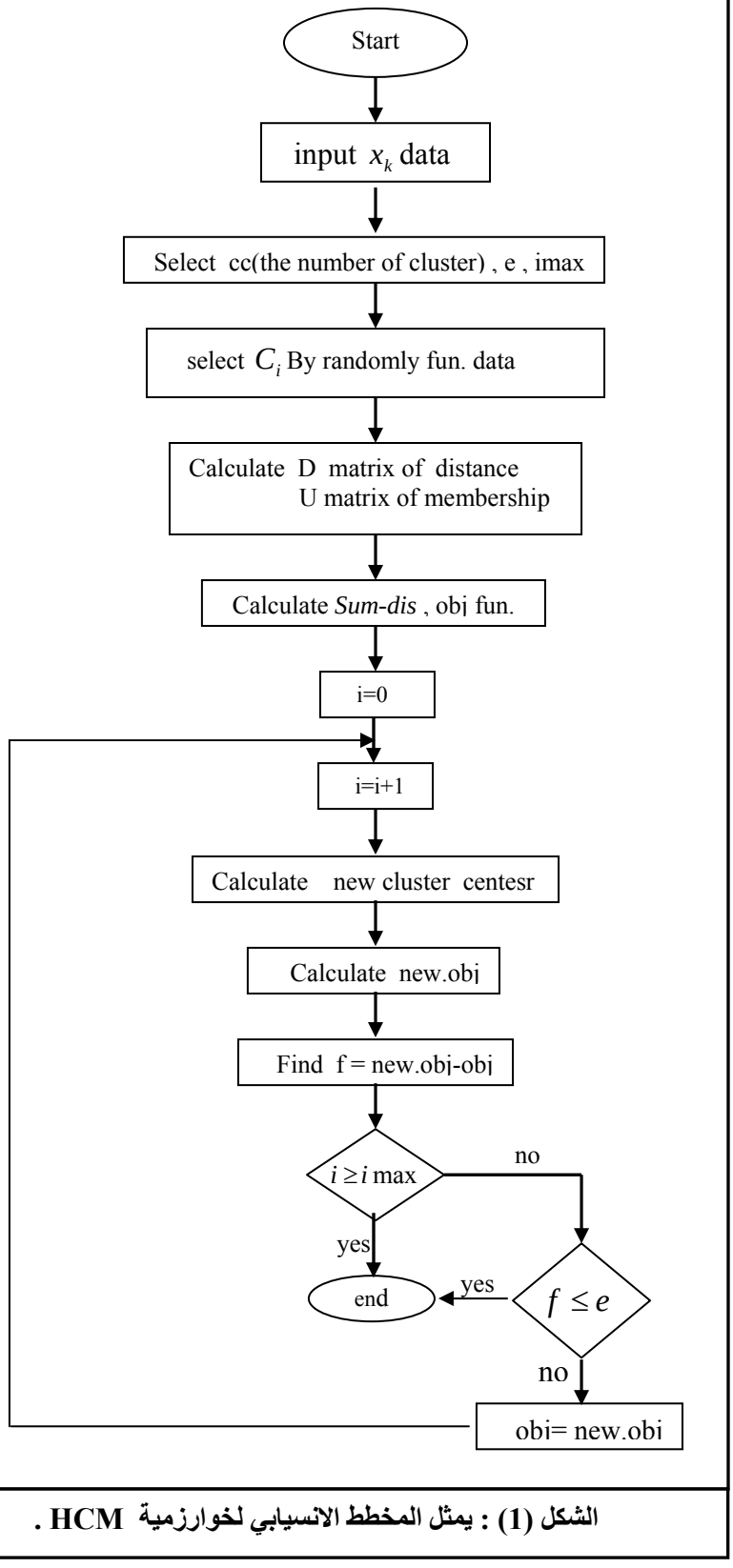
تم بناء هذه الدالة لغرض حساب مصفوفة العضوية كمخرجات لهذه الدالة, أما مدخلاتها فهي مصفوفة المسافة D. ويتم في هذه الدالة حساب درجات انتماء نقاط البيانات إلى العناقيد بالاعتماد على مصفوفة المسافة.

دالة Sum-dis :

تم بناء هذه الدالة لحساب مجموع المسافة الاقليدية بين كل عنصر و مركزه في العنقود الواحد, وان المدخلات لهذه الدالة هي مصفوفة D .

دالة Obj-Fun :

إن مدخلات هذه الدالة هي مصفوفة ***Sum-dis*** المحسوبة مسبقاً. وقد تم بناء هذه الدالة لحساب مجموع عناصر مصفوفة ***Sum-dis*** وتكون دالـة الهـدف obj كمخرجات لها .



الشكل (1) : يمثل المخطط الانسيابي لخوارزمية HCM .

دالة Cal - ci :

تم بناء هذه الدالة لحساب مراكز العناقيد الجديدة في خوارزمية HCM , وإن المدخلات لهذه الدالة هي عدد العناقيد ومصفوفتا البيانات والعضوية . وتقوم هـذه الدالة بحساب مراكز العناقيد الجديدة c_i عن طريق إيجاد المعدل لنقـاط البيانـات التي تنتمي إلى عنقود معين, وتكون مخرجاتها مصفوفة مراكز العناقيد الجديدة .

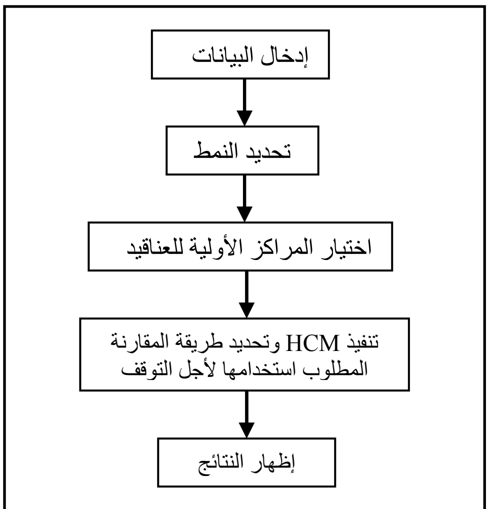
4.5 مراحل تنفيذ البرنامج :

تقسم مراحل تنفيذ البرنامج إلى ثلاث مراحل أساسية :

المرحلة الأولى: تحديد النمط المراد استخدامه وقد اخترنا البيانـات علـى أسـاس المشاهدات Q.

المرحلة الثانية: بعد تحديد الخوارزميـة المطلوبـة لعنقـدة البيانـات والمتمثلـة بخوارزمية HCM لعنقدة سلم الرواتب الشهري وبالاعتماد على النمط Q تم تحديد طريقة اختيار المراكز الأولية للعناقيد بصورة عشوائية.

المرحلة الثالثة: تنفيذ الخوارزمية وتحديد الشرط المطلوب لأجل التوقف والمتمثل بـ حساب الفرق بين دالة الهدف الجديدة (new.obj) المحسوبة بالدورة الأخيـرة ودالة الهدف القديمة (obj) المحسوبة في الدورة السابقة ومقارنة هذا الفرق f مـع قيمة الخطأ المفروضة e , والشكل (2) يوضـح أجـزاء البرنـامج عنـد تنفيـذ خوارزمية HCM .



الشكل(2): يمثل أجزاء برنامج خوارزميةHCM

5.5 تحليل وعنقدة البيانات الواقعية لسلم الرواتب :

لقد تم تطبيق خوارزمية HCM في هذا البحث على بيانـات حقيقيـة لـسلم الرواتب الشهري للملاك التدريسي لإحدى كليات جامعة الموصـل ولعـام 2009 وكما موضحة بالتفصيل بالملحق A, اذ اعتبر أعضاء الملاك التدريسي المشاهدات المختلفة ورمزنا لها بـ $X = \{x_1, x_2, ..., x_n\}$, $n = 130$, اذ أن n تمثل عدد المشاهدات (عدد التدريسيين), وان عدد المتغيرات هو 5 وتتمثل بـ :

(1) اللقب العلمي والذي يتضمن(مدرس مساعد , مدرس , أستاذ مساعد , أسـتاذ) وفرضنا لكل لقب رمز معين (رقم مفـروض) وكمـا يـاتي علـى التـوالي (100,75,50,25), ومن ثم تم تقييسها باذ أصبحت القيم محصورة بين 1,0 .

(2) الشهادة وتتمثل بـ (ماجستير, دكتوراه) وأيضاً فرضنا لهـا قيمـا معينـة, اذ فرضنا "0" للماجستير و "1" للدكتوراه .

(3) عدد سنوات الخدمة وتتراوح بين (3-49) سنة, حاولنا ايضاً تقييس هذه القـيم باذ أصبحت محصورة بين 1,0 .

(4) الدرجة الوظيفية والمرحلة التابعة لكل درجة, اذ أن كل درجة مـن الـدرجات الوظيفيـة تكـون واحـدة مـن الـدرجات الاتيـة (الأولى,الثانية,الثالثة,الرابعة,الخامسة,السادسة) بينما تكون المرحلة التابعة لتلك الدرجة إحدى المراحل الاتية (الأولى, الثانيـة, الثالثـة, الرابعـة, الخامـسة, السادسة, السابعة) وفرضنا أن الترابط بين الدرجة والمرحلة متمثـل بـرقمين مرتبة الآحاد منه تدل على المرحلة ومرتبة العشرات منه تمثل الدرجـة لكـل مشاهدة (تدريسي) من المشاهدات وحسب البيانات الحقيقية, فمثلا الدرجة الثالثة المرحلة الخامسة ستمثل بالرقم 35 بينما الدرجة الخامسة المرحلة الثالثة ستمثل بالرقم 53 وهكذا..الخ. وهذه الأرقام حاولنا تقييسها وتنسيبها وبهذا حصلنا على قيم مرتبة تصاعدياً باذ أصبحت اكبر مرحلة ولأعلى درجة وظيفية اقرب لـ "1" منه لـ "0" .

(5) الراتب الاسمي والذي يعطي قيمة الراتب الاسمي لكل مشاهدة من المـشاهدات وقيمته تقع ضمن الفترة [374,1068] , وأيضا تـم تقييـسها بـاذ أصـبحت محصورة بين 1,0 .

ولتطبيق خوارزمية العنقدة HCM على البيانات الحقيقيـة لـسلم الرواتـب الشهري وتنفيذها على أساس المتغيرات التي حددناها(اللقب العلمي, الشهادة, عـدد سنين الخدمة, الدرجة الوظيفية والمرحلة, الراتب الاسمي) وباختيار المراكز الأولية للعناقيد C_i عشوائياً بالاعتماد على عدد العناقيد المختار سـنقوم بإيجـاد المـسافة الاقليدية بين كل C_i من مراكز العناقيد الأولية وبين أية مشاهدة مـن المـشاهدات $n=1:130$, x_n , اذ سنطرح كل عنقود مختار C_i من كـل المـشاهدات x_n ثـم نضرب كل متجه نحصل عليه بـ *transpose* لذلك المتجه وبأخذ الجذر ألتربيعي نحصل على المسافة d_{ij} بين كل عنقود من العناقيـد وبـين كـل مـشاهدة مـن المشاهدات الـ (130 مشاهدة) وبهذا تتكـون لـدينا مـصفوفة المـسافة $D_{130\times5}$, وبالاعتماد على عناصر مصفوفة المسافة سنحاول إيجاد مصفوفة العضوية U , اذ أننا نفحص كل صف من صفوف D الـ (130 صـف) ونقـوم بتحديـد اصـغر قيمة d_{ij} في ذلك الصف ونضع بالموقع المقابل لها (لنفس الصف والعمود) من U قيمة "1" , أما باقي قيم الصف فتأخذ القيمة "0" وبهذا تكـون أبعـاد مـصفوفة العضوية نفس أبعاد مصفوفة المسافة, بعد ذلك وبالاعتماد على مصفوفة D سـنقوم بإيجاد ***sum-dis*** والتي تمثل مجموع أبعاد كل عنصر في العنقود عن مركزه بحيث ان قيمته تكون بشكل متجه عمودي عدد عناصره تعتمد على عدد العناقيد cc, وبعد أن وجدنا عناصر المصفوفة ***sum-dis*** سنقوم بحساب دالـة الهـدف obj بإيجـاد المجموع لعناصر مصفوفة ***sum-dis*** التي وجدناها سـابقاً. ثـم نـضع i عـدادا للتكرارات ونقوم بزيادته بمقدار واحد في كل مرة, بعـد ذلـك سنحـسب مراكـز مصفوفة العناقيد الجديدة c_i من خلال المعادلة (7) والتي تمثل المعدل, وسـتتكون لدينا مصفوفة أبعادها تعتمد على عدد العناقيـد (الـصفوف) وعـدد المتغيـرات (الأعمدة), وبعد هذا سنحسب دالة الهدف الجديدة new.obj , ثم نقـوم بإيجـاد f الفرق بين دالة الهدف القديمة ودالة الهدف الجديدة , وبعد أن حصلنا على قيمـة f

سنقوم بفحص هل أن اكبر تكرار imax اقل أو مساوي لقيمة i , إذا كان نعم سينتهي البرنامج وإذا لم يتحقق هذا الشرط سوف نفحص هل أن الفرق بين دالة الهدف القديمة والجديدة والذي رمزنا له بـ f اقل أو مساوي لمقدار الخطأ والذي فرضناه $e=0.0001$, فإذا تحقق الشرط الأخير سنتوقف وينتهي البرنامج وإذا لم يتحقق سنضع قيمة دالة الهدف الجديدة new-obj محل القيمة القديمة لدالة الهدف obj ثم نزيد التكرار بمقدار1 ونستمر إلى أن نحصل بالنهاية على العناقيد المطلوبة, مصفوفة المراكز, دالة الهدف النهائية .

6.5 *تفسير النتائج :*

بعد بنائنا لخطوات الخوارزمية المحددة HCM وتحديدنا النمط الذي سنعمل عليه وإدخالنا البيانات الخاصة بتطبيق سلم الرواتب الشهري Xdat وكما موضحة بالملحق A, سنقوم بتنفيذ الخوارزمية وسيتطلب منا إدخال عدد العناقيد cc المراد عنقدة البيانات على أساسها والتي يتم اختيارها بكل تنفيذ بقيمة معينة بحيث ان اقل عدد مختار للعناقيد cc=2 وفي حالة اخترنا cc=1 سنلاحظ أن كل البيانات ستقع في عنقود واحد ويكون لها مركز واحد وهذه الحالة لن تفيدنا في تفسير أسلوب العنقدة بشكل واضح لان كل البيانات ستتمركز حول نقطة واحدة فقط, وإن اكبر عدد مختار للعناقيد يجب ألا يتجاوز عدد المشاهدات (130), وفي حالة cc=130 سنلاحظ أن كل عنصر(مشاهدة) ستقع في عنقود مستقل وهذا أيضا لايعطي التفسير المنطقي المطلوب. حاولنا تنفيذ الخوارزمية باختيارنا في كل مرة عدد العناقيد مختلف عن سابقه وكمایاتي {2,3,4,5,6,7,8,9,10} وبعد هذا حاولنا تكرار العملية السابقة خمس مرات ولاحظنا في كل مرة نحصل على عناقيد, ومصفوفتي U,D ومصفوفتي المراكز و ***sum-dis*** وكذلك سنحصل على دالة الهدف النهائية والتي تختلف قيمها في كل مرة عن المرة السابقة لها وهذا السبب يعود إلى انه في كل مرة تقوم دالة rand الخاصة باختيار المراكز الأولية للعناقيد C_i بصورة عشوائية باختيار العناقيد بشكل معين يختلف عن سابقتها. وقمنا باختيار أفضل نتائج حصلنا عليها من التكرارات الخمسة لتنفيذ الخوارزمية ولكل العناقيد المختارة. وإن النتائج المثلى التي حصلنا عليها يمكن أن ندرجها في الجدول (1) اذ أن كل من ,

$c_i Ai$ ($i=1,...,10$, i تمثل عدد العناقيد) موضحة بالتفـصيل فـي الملحـق B , ونلاحظ من الجدول ادناه مايلتي :

الجدول(1) : يبين عدد العناقيد والعناقيد الناتجة ودالة الهدف التابعة لها والفرق بين دالة الهدف القديمة والجديدة

عدد العناقيد المختارة cc	العناقيد الناتجة	مصفوفة المراكز	دالة الهدف النهائية obj	الفرق d
2	*A2*	c_2	15.1153	4.2444
3	*A3*	c_3	10.8709	2.2890
4	*A4*	c_4	8.5819	1.8508
5	*A5*	c_5	6.7311	1.5494
6	*A6*	c_6	5.1817	1.2635
7	*A7*	c_7	3.9182	0.8807
8	*A8*	c_8	3.0375	0.5635
9	*A9*	c_9	2.4740	0.0463
10	*A10*	c_{10}	2.4277	

- العمودان الأول والثاني من هذا الجدول يمـثلان عـدد العناقيـد المختـارة والعناقيد الناتجة على التوالي, ففي حالة اختيار $cc=2$ أي عندما يكون عـدد العناقيد المختار اثنين نحصل على عنقودين أي أن البيانـات (المـشاهدات) تتعنقد إلى عنقودين ورمزنا لهذه المجموعة المكونة من العنقودين بـالرمز *A2*وعند اختيار عدد العناقيد $cc=3$ سوف تتعنقد البيانات إلى ثلاثة عناقيـد وهكذا إلى أن نختار عدد العناقيد $cc=10$ فنحصل على عشرة عناقيد أي أن البيانات تتعنقد إلى عشرة عناقيد, علماً أن هذه العناقيد موضحة بالتفصيل في الملحق B .
- العمود الثالث من الجدول يمثل مصفوفة المراكز c_i المقابلة لكل مجموعة من العناقيد الناتجة, وان هذه المصفوفة هي مصفوفة مكونـة مـن عـدد مـن الصفوف تبعاً لعدد العناقيد المختار وخمسة أعمدة تبعاً لعدد المتغيرات (اللقب العلمي, الشهادة, عدد سنين الخدمة, الدرجة الوظيفيـة والمرحلـة, الراتـب

الاسمي) وكل مصفوفة من هذه المصفوفات تتكـون مـن مجموعـة مـن العناصر, فمثلاً عندما يكون عدد العناقيد $cc=2$ ستكون مصفوفة المراكز c_2 مكونة من صفين وخمسة أعمدة , وأن العنصر الأول من هـذه المـصفوفة $c_2(1,1)$ والذي يقع في الصف الأول والعمود الأول من هذه المصفوفة يعـد المركز الأمثل (المعدل) لجميع المشاهدات في العنقود الأول وللمتغيـر الأول المتمثل باللقب العلمي, والعنصر الثاني $c_2(1,2)$ الذي يقع في الـصف الأول والعمود الثاني يمثل المركز الأمثل (المعدل) لجميع المشاهدات في العنقـود الأول وللمتغير الثاني المتمثل بـ الشهادة وهكذا إلى أن نصل الى العنـصر الخامس من الصف الأول $c_2(1,5)$ الـذي يمثـل المركـز الأمثـل لجميـع المشاهدات في العنقود الأول وللمتغير الخامس المتمثل بالراتب الاسمي, أمـا العنصر الأول من الصف الثاني $c_2(2,1)$ من هذه المصفوفة يمثـل المركـز الأمثل (المعدل) لجميع المشاهدات في العنقود الثاني وللمتغير الأول (اللقـب العلمي) وهكذا بقية العناصر لهذه المصفوفة إلى أن نصل إلى أخر عنـصر بالمصفوفة $c_2(2,5)$ والذي يمثل المعدل لجميع المشاهدات في العنقود الثاني وللمتغير الخامس المتمثل بـ الراتب الاسمي, وبالأسلوب نفسه يكون تحليـل باقي مصفوفات المراكز. ومن الملاحظات المهمة التي توصلنا إليهـا مـن التجربة المتكررة لهذا التطبيق انه في هذه المصفوفة لو كـان فـي العمـود الواحد أي لمتغير معين قيمة المراكز متقاربة سوف يـدل علـى إن ذلـك المتغير ليس له أهمية ولا يؤثر كثيرا في النواتج.

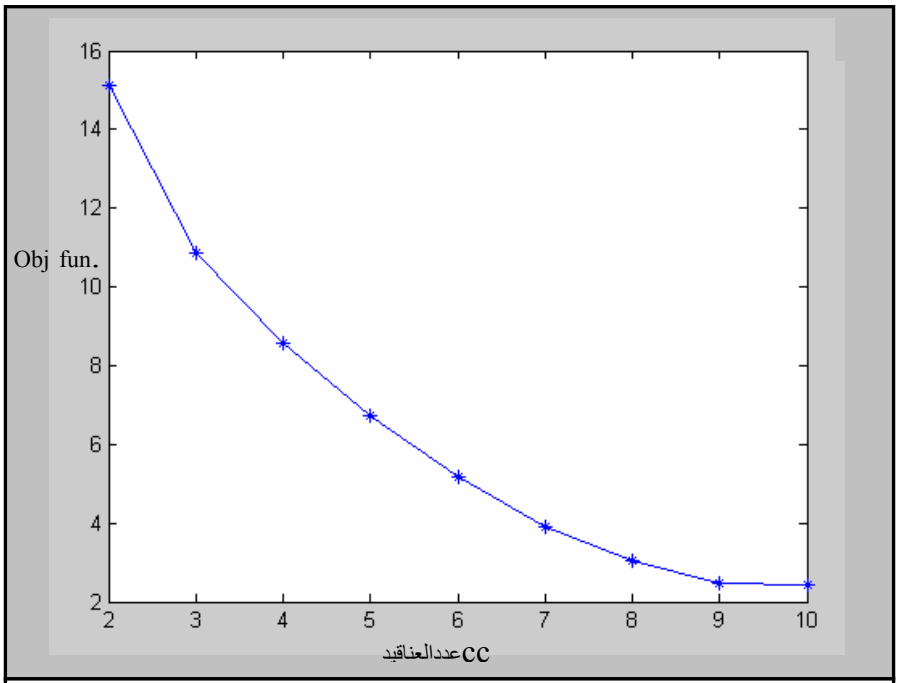
- وفي الحقل الرابع من الجدول المذكور انفا وضعنا قيمة دالة الهدف النهائيـة والتي تحقق لنا الشرط الذي يكون فيه مقدار الفرق f بين قيمة أية دالة مـن دوال الهدف وقيمة دالة الهدف السابقة لها مساوي للصفر, اذ أن قيمة دالـة الهدف تعتمد على قيم عناصر مصفوفة sum-dis , وان قيم دالة الهدف تعد أهم فرق تتضح لنا فيه كيفية سير عملية العنقدة في هذا التطبيق, فكلمـا زاد عدد العناقيد سوف تقل قيمة دالة الهدف النهائية, ولو أمعنا النظر فـي هـذا الحقل من الجدول (1) سنلاحظ مايأتي :

- ◊ عند اختيارنا عدد العناقيد *cc=2* يكون مقدار مجموع المسافات الاقليدية بين كل عنصر عن مركزه في العنقود الأول مع مجموع المسافات الاقليدية بين كل عنصر عن مركزه في العنقود الثاني مساويا لـ (15.1153).
- ◊ وعندما نختار عدد العناقيد *cc=3* يكون المقدار لمجموع المسافات الاقليدية بين كل العناصر ومراكزها وللعناقيد الثلاثة مساويا لـ (10.8709).
- ◊ وأخيراً عندما نصل إلى عدد العناقيد *cc=10* يكون المقدار لمجموع المسافات الاقليدية بين كل العناصر ومراكزها وللعناقيد العشرة مساويا لـ (2.4277) .

- أما الحقل الأخير من الجدول (1) فيحوي الفرق بين أية دالة من دوال الهدف النهائية ودالة الهدف التي تليها, وكلما زاد عدد العناقيد قل الفرق بين أية قيمة والقيمة التي تليها لدالة الهدف obj إلى أن تصل إلى درجة لايكون لهذا الفرق أهمية اوتاثير كبير في عنقدة البيانات, فعند اختيارنا عدد العناقيد *cc=2* و *cc=3* فان الفرق بين المجموع عندما يكون عدد العناقيد 2 والمجموع عندما يكون عدد العناقيد 3 هو(4.2444) والفرق بين المجموع عندما يكون *cc=3* و *cc=4* يكون مساوي لـ(2.2890) وهكذا نستمر وكما مبينة القيم بالجدول(1) إلى أن نصل *cc=9* و *cc=10* فيكون الفرق بين المجموع عندما يكون عدد العناقيد 9 والمجموع عندما يكون عدد العناقيد 10 هو(0.0463).

ومن الاستنتاجات المهمة التي توصلنا إليها هو انه كلما كانت قيمة دالة الهدف اقل يكون أفضل أي أن العناصر الموجودة في كل عنقود تكون غير بعيدة عن مركزها ولكل العناقيد, ويمكننا تمثيل قيم دالة الهدف النهائية من خلال الشكل (3). ولتقدير أفضل عدد للعناقيد اعتمدنا على إيجاد مربعات الخطأ لقاعدة البيانات ومن ثم حاولنا تحويل مربعات الأخطاء إلى قيم معيارية لمربعات الأخطاء وقارنا بين القيم المعيارية هذه وبشكل متتال ابتداءً من أول قيمة معيارية لمربعات الأخطاء

وهكذا صعوداً إلى باقي القيم . ومن خلال العديد من الممارسات العملية لوحظ بأن القيم المعيارية لمربعات الأخطاء تتناقص ببطء ثم عند الوصول إلى العدد المناسب من العناقيد يحدث هبوط قوي في القيمة المعيارية لمربعات الأخطاء, ثم بعد ذلـك العدد فان الأعداد الاتية سوف تستقر تقريباً عند قيم معيارية قريبة من الصفر وهذا ماحدث بالفعل ومن خلال البرمجة فحصلنا على إن أفضل عدد للعناقيد هو $cc=6$ والذي تكون فيه قيمة دالة الهدف مساوية لـ(5.1817) بينما تكون قيمة دالة الهدف السابقة لها عندما $cc=5$ هي (6.7311) وقيمة دالة الهدف التي تليها عنـدما $cc=7$ هي (3.9182). ولو أمعنا النظر بالشكل (3) سوف نلاحظ أن عدد العناقيد المختار قبل العدد $cc=6$ يحدث هبوط قوي في انحدار المنحني وبعد ذلك العدد فان الأعداد الاتية تستقر نسبياً وكما موضح بالشكل(3) :



الشكل (3): يمثل عدد العناقيد cc المختارة وقيم دالة الهدف obj المقابلة لها .

ومن ناحية أخرى لو تفحصنا عناصر كل عنقود من العناقيد الناتجـة فـي الملحق B سنلاحظ مايأتي :

- في حالة اختيار عدد العناقيد $cc=2$ إن البيانات سوف تتعنقد إلى عنقودين على أساس الشهادة أي أن العنقود الأول يضم كل المشاهدات التي شهادتهم ماجستير والعنقود الثاني يضم كل المشاهدات التي تكون شهادتهم دكتوراه.
- وفي حالة اختيار عدد العناقيد $cc=3$ ستتعنقد البيانات على أساس سنين الخدمة فالعنقود الأول , الثاني , الثالث سيحوي كل منهم المشاهدات التي تكون سنين خدمتهم تقع ضمن الفترات وعلى التوالي $[3,10]$, $[11,20]$, $[21,49]$.
- وعندما يكون عدد العناقيد $cc=4$ سوف تتعنقد البيانات على أساس اللقب العلمي فالعنقود الأول يضم كل المشاهدات التي لقبها العلمي مدرس مساعد والعنقود الثاني يضم كل المشاهدات التي لقبها العلمي مدرس والثالث والرابع كل واحد منهما يضم كل المشاهدات التي لقبها العلمي على التوالي أستاذ مساعد وأستاذ.
- وعنما وضعنا $cc=5$ لاحظنا أن المشاهدات التي يكون لقبها العلمي مدرس مساعد وشهادتها ماجستير ستتجمع بعنقود, والمشاهدات التي يكون لقبها العلمي مدرس وشهادتها ماجستير ستتجمع بالعنقود الثاني ,أما العنقود الثالث فيضم المشاهدات التي لقبها العلمي مدرسا وشهادتها دكتوراه, والعنقود الرابع سيحوي المشاهدات التي لقبها العلمي أستاذا مساعدا وشهادتها ماجستير, أما العنقود الأخير سيضم المشاهدات التي يكون لقبها العلمي أستاذا و أستاذا مساعدا وشهادتهم دكتوراه.
- أما في حالة اختيار $cc=6$ فان البيانات تتعنقد على أساس الدرجة فمثلا العنقود الأول يضم كل المشاهدات التي درجتها السادسة والعنقود الثاني يضم كل المشاهدات التي درجتها الخامسة وهكذا...الخ.
- وفي حالة اختيار $cc=7$ تتعنقد على أساس المراحل وسنين الخدمة.
- ولو وضعنا $cc=8$ سنلاحظ أن العناقيد المتكونة قد تعنقدت على أساس الراتب الاسمي اذ يضم العنقود الأول المشاهدات التي يكون راتبها الاسمي ضمن الـ $[374-399]$ ألف, والعنقود الثاني يضم المشاهدات التي يكون راتبها الاسمي ضمن الـ $[400-499]$ ألف,أما العنقود الثالث فيضم المشاهدات التي يكون

راتبها الاسمي ضمن الـ *[500-599]* ألف, والعنقود الرابع والخامس والسادس والسابع والثامن يضم كل واحد منهم المشاهدات التي راتبهـا الاسـمي علـى التوالي*[900-999],[1000-1100],[600-699],[700-799],[800-899]*. وبهـذا يكون لكل عدد عناقيد معين تفسير خاص به يتناسب مع بيانات المـسالة ومـا يحيط بها من تعلقات وصفات تميزها .

6. الاستنتاجات والتوصيات :

لقد حاولنا في هذا البحث تقديم مفاهيم أساسية لإحدى التقنيات الحديثـة فـي تكنولوجيا المعلومات بمفهوم العنقدة وإحدى الخوارزميات المستخدمة لهذا المفهـوم الذي يلاقي اهتمامات كبيرة في الوقت الحاضر والذي تم تطبيقـه علـى البيانـات الواقعية لسلم الرواتب الشهري للملاك التدريسي لإحدى كليات جامعـة الموصـل ويمكن إجمال أهم الاستنتاجات والتوصيات التي توصلنا إليها بالنقاط الاتية:

- إن خوارزمية العنقدة لها دور كبير في تصنيف مجموعة البيانات إلـى أصـناف محددة لاكتشاف المعلومات القيمة في البيانات اذ أن التعامل مع هذه المعلومـات يكون أفضل بكثير من التعامل مع البيانات قبل إجراء عملية العنقدة عليها لأنـه في بعض الحالات تكون كمية البيانات المتوفرة كبيرة جـدا ومـن ثـم تـصبح معالجتها ضرورية جدا لذا فان العنقدة تساهم في تقلـيص البيانـات, وإن عـدد العناقيد الناتجة اصغر بكثير من عدد نقاط البيانات الأصلية, وبهذا تحتـاج إلـى مساحة خزن قليلة في الذاكرة وإمكانية بحث كفوءة أكثر.
- إن طبيعة البيانات المراد تطبيق العنقدة عليهـا لهـا دور كبيـر فـي اختيـار الخوارزمية الملائمة للحصول على النتيجة المطلوبة والكفـوءة , ففـي بعـض التطبيقات تكون خوارزمية HCM هي الأفضل ولكن في تطبيقات أخـرى قـد لاتكون ملائمة للوصول إلى نتيجة مرضية .
- عند إعادة تنفيذ الخوارزمية وباستخدام نفس البيانات ونفس العدد مـن العناقيـد يظهر اختلاف نتائج الخوارزمية في حالة اختيار المراكز الأولية للعناقيد بصورة عشوائية .

- إن نظام الحاسوب المصمم هو نظام يتمتع بقدرة كبيرة من العمومية وهو ليس مصمما فقط للمعالجة في حالة عنقدة البيانات على أساس المشاهدات عند اختيارQ-mode لإيجاد التشابه بين المشاهدات , بل أيضاً ممكن إيجاد التشابه بين المتغيرات عند اختيارR-mode في حالة عنقدة البيانات على أساس المتغيرات.
- إن البرنامج عام ويعمل على n من المشاهدات و m من المتغيرات , لذا فيمكننا زيادة عدد البيانات على مرور الفترات الزمنية, ففي حالة تعيين تدريسي جديد ضمن الملاك التدريسي للكلية المدخلة بياناتها وإدخال البيانات الخاصة بالتدريسي الجديد وتنفيذ البرنامج سوف نلاحظ أن البرنامج يسير بصورة طبيعية وبنفس الكفاءة ونحصل على كل من obj,C,U,D تبعاً للبيانات الجديدة , وكذلك في حالة تغيير البيانات وإدخال بيانات واقعية أخرى لكلية ثانية سنلاحظ أن البرنامج لهذه الخوارزمية سيعمل ايضاً بكفاءة ونحصل على النتائج والعناقيد المطلوبة بسهولة ووضوح .

المصادر

Batistakis,M. and Vazirgiannis,M.(2001)."**Clustering Algorithim and Validi-ty Measures**"; proceedings of the 13th international conference on scientific and Statistical Database Management.IEEE Computer Society, Virginia, USA..

Anil,K.(2008)," **Data Clustering: 50 Years Beyond K-Means** " ; delivered at the 19th International Conference on pattern Recognition , Department of Computer Science & Engineering Michigan University 48824 USA .

Ajmera,J. and Wooters,C. (2003). " **A robust speaker clustering algorithem**", pp. 411-216. IEEE Automatic Speech Recognition .

Witold,P.,(2005)." **Knowledge – based clustering** "; John Wiley & Sons Inc. ISBN: 0-471-46966-1.

Banerjee,A., Merugu,S., Dhillon,I. and Ghosh,J.,(2004)." **Clustering with bregman divergences** "; Journal of machine learning research, pp 234-245.

Banerjee,A.,Basu,S.and Merugu,S.,(2007a)."**Multi-way clustering on relation graphs** "; Proceedings of the 7th siam International Conference on data mining.

Basu,S.Davidson,I.and Wagstaff,K.,(2008)," **Constrained clustering: Advan-ces in algorithms, theory and applications**"; vol. 3 .Chapman & Hall/ CRC.

Bekkerman,R.,Ran,E. and Andrew,M.,(2005)." **Multi-way distributional clus-tering via pairwise interactions** "; pp 41-48., Proceedings of the 22nd International Conference on machine learning. NY, USA .

Dhillon,I.,Guan,Y. and Kulis,B.,(2004)." **Kemenl k-means: spectral cluster-ing and normalized cuts** "; pp 551-556., Proceedings of the tenth acm sigkdd International Conference on knowledge discovery and data mining . NY, USA .

Figueiredo,M.,Chang,D.and Murino,V.,(2006)." **Clustering under prior kno-wledge with applications to image segmentation** " ; pp 401-408.

Fred,A.and Jain,A.,(2002)." **Data clustering using evidence accumulateion** "; Proceedings of the International Conference on Pattern Recognition.

Lu,Z.and Leen,T.,(2007)." **Penalized probabilistic clustering** "; Neural comput., pp 1528-1567.

Pelleg,D. and Moore,A.,(1999)." **Accelerating exact k-means algorithms with geometric reasoning** "; pp 277-281., Proceedings of the 5th International Conference on knowledge discovery in databases. AAAI.

Tsuda,K.and Kudo,T.,(2006)." **clustering graphs by weighted substructure mining** "; pp 953-960., Proceedings of the 23rd International Conference on machine learning. NY, USA .