Evaluation of Clustering Validity

Rudhwan Yousif Sideek **Ghayda A.A. Al-Talib**
ghaydabdulaziz@uomosul.edu.iq
College of Computer Sciences and Mathematics
University of Mosul, Iraq

Received on: 01/11/2007 **Accepted on: 04/03/2008**

ABSTRACT

Clustering is a mostly unsupervised procedure and the majority of the clustering algorithms depend on certain assumptions in order to define the subgroups present in a data set. As a consequence, in most applications the resulting clustering scheme requires some sort of evaluation as regards its validity.

In this paper, we present a clustering validity procedure, which evaluates the results of clustering algorithms on data sets. We define a validity indexes, S_Dbw & SD, based on well-defined clustering criteria enabling the selection of the optimal input parameters values for a clustering algorithm that result in the best partitioning of a data set.

We evaluate the reliability of our indexes experimentally, considering clustering algorithm (K_Means) on real data sets.
Our approach is performed favorably in finding the correct number of clusters fitting a data set.

Keywords: Data Mining, K_Means, S_Dbw, SD

تقييم صحة العنقدة

غيداء عبد العزيز مجيد الطالب **رضوان يوسف صديق الجوادي**
كليةعلوم الحاسوب والرياضيات *الكلية التقنية*
جامعة الموصل *هيئة التعليم التقني/ الموصل*

تاريخ استلام البحث: 2007/11/1 **تاريخ قبول البحث: 2008/3/4**

الملخص

العنقدة هي إجراءات تكون على الأغلب دون مشرف ، واغلب خوارزميات العنقدة تعتمد على افتراضات معينة لغرض تعريف المجاميع الجزئية الموجودة في مجموعة البيانات.

نتيجة لذلك فان اغلب تطبيقات نماذج العنقدة الناتجة تتطلب شيئا من التقييم لإثبات صحة العنقدة .

في هذا البحث تم عرض إجراء لتقييم نتائج خوارزميات العنقدة في مجموعة البيانات، اذ تم تعريف مؤشري صحة هما S_Dbw و SD يستندان إلى معيار عنقدة كفوء يساهم في تحصيل أفضل قيمة في معاملات البيانات المدخلة لخوارزمية العنقدة والتي تنتج من أفضل تجزئة لمجموعة البيانات.

تم تقييم الوثوقية للموشرات المختارة عمليا ، استناداً إلى خوارزمية العنقدة (K_Means) المطبقة على مجموعة بيانات حقيقية .

يهدف البحث الى إيجاد العدد الصحيح للعناقيد التي تلائم مجموعة البيانات تحت الأختبار. واستخدمت لغة Visual basic 6 في تصميم البرامج وتنفيذها.

الكلمات المفتاحية: تنقيب البيانات، خوارزمية العنقدة (K_Means)، مؤشري الصحة S_Dbw و SD

المقدمة:

مما لا شك فيه أن انتشار قواعد المعلومات على اختلاف أنواعها (الببليوغرافية والنص الكامل) أدى إلى تضخم المعلومات فضلا عن تنوع المستخدم من الإنسان العادي إلى المتخصص، وكلاهما لم يبدل الهدف الأساسي من البحث والاسترجاع ألا وهو الوصول إلى المعلومة المطلوبة بدقة وسهولة مما يتطلب في أغلب الأحيان طريقة بحث تعتمد على اللغة المشتركة بين جميع المستخدمين ولذلك اقتصرت طريقة البحث على استخدام اللغات شبه الطبيعية مثل لغة SQL [5,4].

ان من سمات الانتشار الواسع لتكنولوجيا المعلومات تضخم حجم المعلومات تضخماً كبيراً بحيث أصبح عنصراً مهماً ومؤثراً في جوانب عديدة من المجتمع ، وإن معالجة هذه المعلومات والاستفادة منها ، مع الانتشار الواسع لشبكة الانترنيت التي أصبحت الوسيلة الأساسية للاتصال ولنشر المعلومات وتبادلها ، تتطلب توثيق المعلومات بطريقة آلية كفوءة مع الأخذ بنظر الاعتبار البحث في محتوى النصوص والبيانات مع ما يستلزم ذلك من أدوات معلوماتية لغوية فعالة .

لذلك فان التقنيات الحديثة مثل (Data Mining , Data Web , Data Media) المستخدمة في بنوك المعلومات وغيرها تتطلب تضافر جهود المعنيين من المعلوماتيين واللغويين والمتخصصين وحتى المستخدمين بهدف تصميم بنوك المعلومات ونشرها وإيجاد أدوات التوثيق والبحث وكيفية إيجاد واجهات سهلة الاستعمال.[3]

1. هدف البحث:

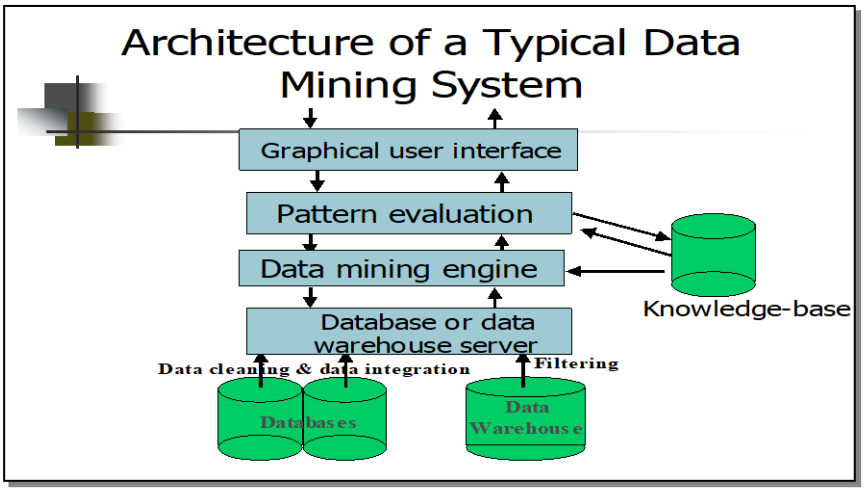
من الأنظمة المعلوماتية الجديدة أنظمة التنقيب في البيانات ضمن مخازن البيانات. ومن طرائق التنقيب في البيانات "العنقدة" التي تعمل على فصل مجموعة البيانات إلى عناقيد متشابهة ومترابطة داخليا ، مما يقودنا إلى التفكير في تقييم صحة النتائج المستخلصة من خوارزميات العنقدة وهذا هو هدف البحث.

2. التنقيب في البيانات (Data Mining):

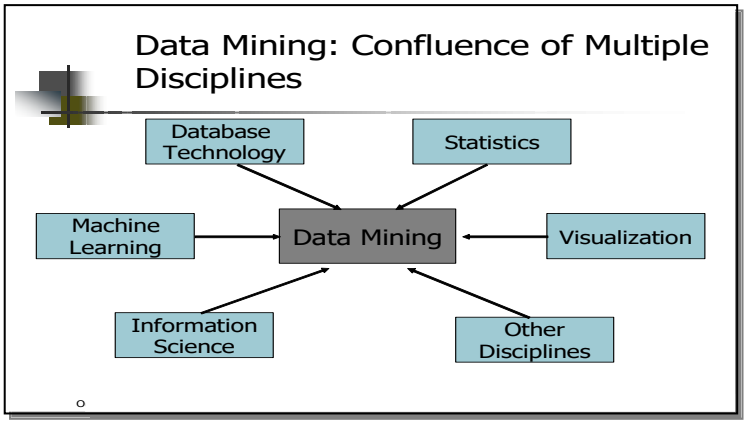
التنقيب في البيانات هو عملية الكشف والعثور على معلومات ذات فائدة من خلال استعمال مجموعة من الأدوات المعقدة. بعض من هذه الأدوات تشمل أدوات الإحصاء الاعتيادية والرسوم البيانية الحاسوبية .

فالتنقيب في البيانات منهجية تجمع بين نتائج الأبحاث في الذكاء الاصطناعي، الفهم الآلي، التعرف على الأشكال، قواعد المعلومات، الرياضيات الإحصائية، واجهات الاستعمال واللغة.

وبحسب مراكز الأبحاث يمكن الفصل بين نوعين من استخراج المعلومات فتسمية (Knowledge Discovery in Databases KDD) تستعمل عند الباحثين في الذكاء الاصطناعي والفهم الآلي وتسمية Data Mining تستعمل عند الباحثين في الرياضيات الإحصائية أو خبراء المعلومات . والشكل (1) يمثل معمارية التنقيب في البيانات.[9,10]



الشكل (1) معمارية التنقيب في البيانات



والشكل (2) يوضح عددا من المجالات التي تستخدم التنقيب في البيانات.

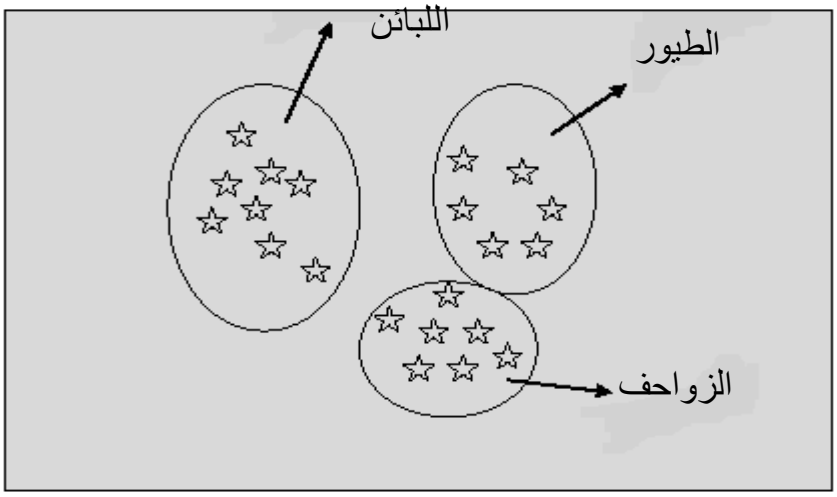
الشكل (2) علاقة التنقيب في البيانات مع مجالات أخرى

وتطبيقات التنقيب في البيانات بدأت تنمو بصورة كبيرة للأسباب الآتية:[1]

1) كمية البيانات الموجودة في مخزن البيانات وسوق البيانات تنمو بصورة كبيرة. ومن أجل ذلك فإن المستخدم يحتاج إلى أدوات متطورة من مثل التنقيب في البيانات من اجل استخلاص الفائدة والمعرفة من هذه البيانات.
2) الكثير من أدوات التنقيب عن البيانات بدأ يظهر مؤخرا، وكل أداة أفضل من الأخرى.
3) المنافسة الشديدة الموجودة في السوق تدفع الشركات إلى الاستفادة القصوى من البيانات التي بيدها، وعمليات التنقيب في البيانات توفر ذلك تماما.

3. العنقدة:

هي عملية استخلاص او ايجاد مجموعة من الاشياء (الكائنات) المتشابهة فيما بينها بشكل مجاميع ، بحيث تكون عناصر كل مجموعة متشابهة مع بعضها بصورة أكبر ، مثال على ذلك مجموعة اللبائن، ومجموعة الزواحف ومجموعة الطيور في الشكل(3). [6,7]



الشكل (3) عناقيد البيانات

وهذه الأشياء في العنقود الواحد تكون متشابهة فيما بينها ومختلفة عن العنقود الآخر. وإن قوة تشابه الأشياء في العنقود الواحد تؤدي إلى أفضل عنقدة ودالة التشابه تعدّ مقياسا لجودة العنقدة. [8]

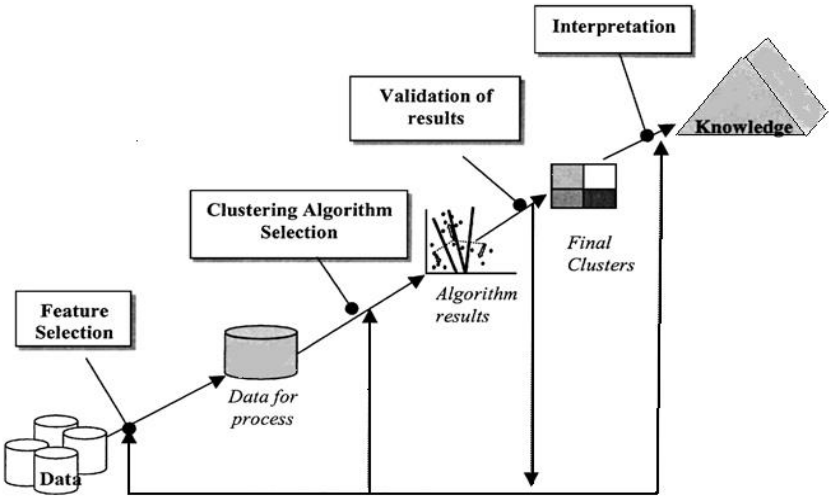
تعتمد معظم خوارزميات العنقدة على افتراضات مستمدة من تعريف المجاميع الجزئية والمتمثلة أو الموجودة في مجموعة البيانات تحت الاختبار واغلب هذه الخوارزميات تعمل بدون مشرف .[15]

ويعدّ تحليل العنقود من الأهداف المفيدة في تحديد التوزيعات البيانية المرغوب فيها وتحديد أنماط للبيانات الأساسية، والخطوات الأساسية لتطوير عملية العنقدة ممثلة في الشكل–4 والتي يمكن أن تلخص بالخطوات الآتية [12]:

3.1 اختيار الصفات:

الهدف هو اختيار الصفات الصحيحة التي من خلالها سوف تصاغ العنقدة التي تمكننا من فك رموز المعلومات الكبيرة المتوقعة والمتعلقة بالمسألة المطلوب حلها . وان مصطلح فك

الرموز هو عملية تحويل الصفات الحرفية إلى قيم عددية . لذلك من الضروري المعالجة المسبقة للبيانات قبل الاستفادة منها .



الشكل(4) خطوات عملية العنقدة

3.2 اختيار خوارزمية العنقدة [12,13]:

في هذه الخطوة سوف يتم اختيار الخوارزمية المناسبة اعتماداً على نوعية البيانات الناتجة من مخطط العنقدة الجيدة في الخطوة السابقة . المقياس التقريبي للعنقدة ومقياس العنقدة الرئيس سوف يحددان صفات خوارزمية العنقدة فضلا عن ذلك يحددان كفاءة مخطط العنقدة الذي يلائم مجموعة البيانات .

المقياس التقاريبي:

هذا المقياس يحدد التشابه بين نقطتين بيانيتين في العنقود نفسه ، والنقاط البيانية هي عبارة عن متجهات أي صفات البيانات لسجلين او قيدين ، وفي اغلب الحالات يجب التأكد من ان كل الصفات المختارة والمساهمة متشابهة او متساوية لحسابات مقياس التقارب وليس هنالك صفات تسيطر او تهيمن على الصفات الاخرى .

مقياس العنقدة الرئيس:

في هذه الخطوة سوف يتم التعرف على مقياس العنقدة الذي يعبر عن دالة الكلفة الخاصة أو بعض القواعد الأخرى لذلك يجب أن نأخذ بالحسبان نوعية العناقيد المتوقعة الحدوث في مجموعة البيانات وهنا يجب تعريف مقياس عنقدة جيد يقودنا إلى تقسيم يلائم مجموعة البيانات .

4. مواصفات خوارزميات العنقدة [12,13]:

يمكن تصنيف خوارزميات العنقدة على وفق ما يأتي :

- نوعية البيانات المدخلة للخوارزمية.
- معيار او مقياس العنقدة والذي يعرف مدى التشابه بين عناصر او كائنات البيانات .
- النظرية والمفاهيم الاساسية التي تعتمد عليها تقنية تحليل العنقدة (مثال على ذلك النظرية المضببة والاحصاء).

المشكلة الأساسية للعناقيد هي تقسيم مجموعة البيانات المعطاة إلى عدد عناقيد صحيحة ومثال على ذلك في العنقود الواحد يجب أن تكون متجهات البيانات متشابهة جداً بعضها مع البعض الآخر.وقد اصبحت مجموعة من خوارزميات العنقدة في السنين الأخيرة مجهزة ومتوافرة نظريا ، وبناءً على الطريقة المعطاة في تعريف العناقيد ، فان الخوارزميات يمكن تصنيفها إلى الأنواع الآتية : [17,16]

- العنقدة التجزيئية.
- العنقدة الهرمية .
- العنقدة المستندة إلى الكثافة.
- العنقدة المستندة إلى الشبكة المتعامدة.
- العنقدة المضببة.

5. المقاييس الأساسية لصحة العنقود[12,14]:

تستخدم مقاييس كثيرة لعرض قياس كفاءة خوارزميات تحليل البيانات في مجال التنقيب عن البيانات وعنقدتها .ومن هذه المقاييس هي المقاييس الخارجية ، المقاييس الداخلية والمقاييس النسبية وسوف نتطرق في هذا البحث إلى التعريف بهذه المقاييس بشكل عام وصولاً إلى استخدامات المقاييس النسبية التي هي موضوع بحثنا .

5.1 المقياس الخارجي:

الفكرة الأساسية في هذا المبحث هي اختيار فيما إذا كانت النقاط في مجموعة البيانات مهيكلة عشوائيا أم لا ؟ وهذا يقود إلى حساب نتائج خوارزمية العنقدة بالاعتماد على هيكل محدد مسبقا والذي يُفرض على مجموعة البيانات ويعكس حدسنا حول هيكل العنقدة لمجموعة البيانات. بالنتيجة فان تقنيات موندي كارلو(Mondi Carlo) الإحصائية تستخدم حلولاً للمسائل ذات الحسابات العالية.

5.2 المقياس الداخلي:

تستخدم هذه الطريقة مقياساً لصحة العنقدة ، هدفنا هو حساب نتائج العنقدة للخوارزمية باستخدام الكميات والصفات الموروثة من مجموعة البيانات . إذ بإمكاننا حساب نتائج خوارزمية العنقدة في صيغ من الكميات الممثلة بالمتجهات لمجموعة البيانات نفسها ، و هنالك صيغتان سوف تستخدم المقياس الداخلي لصحة العنقود وهما :

- مخطط العنقدة الهرمية .
- مخطط العنقدة المفردة .

5.3 المقياس النسبي:

إن أساس طرائق التقييم السابقة هي الفحوص الإحصائية لذلك فان نقاط الضعف في التقنيات التي تستند إلى المقاييس الخارجية والمقاييس الداخلية هي المطالب الحسابية العالية ، فضلا عن ذلك فان المؤشرات المتعلقة بهذه الطرائق تهدف الى قياس درجة أية مجموعة من البيانات التي تؤكد المخطط المحدد مسبقا .

أما المقياس الثالث (المقياس النسبي) فانه يهدف إلى اختيار أفضل مخطط لمجموعة معرفة من مخططات العنقدة والذي تستطيع خوارزمية العنقدة تعريفه وفقاً لافتراضات ومعاملات محددة مسبقاً . والفكرة الأساسية هنا هي حساب هيكل العنقدة بالمقارنة مع مخططات عنقدة أخرى ناتجة من الخوارزمية نفسها لكن لقيم معاملات مختلفة.

6. عينة البحث:

تم اختيار خوارزميتي (SDوS_Dbw) لتقييم صحة نتائج العنقدة المستحصلة من خوارزمية (K_means) التي تعمل بدون مشرف ، وذلك لإيجاد أفضل مخطط عنقدة يلائم مجموعة البيانات المختارة ، وحجم البيانات هو مصفوفة تحتوي على خمس سمات لسبعة محاصيل حقلية ولمنتج سبع سنوات، إذ تم اختيار بذور المحاصيل (الحنطة ، الشعير ،

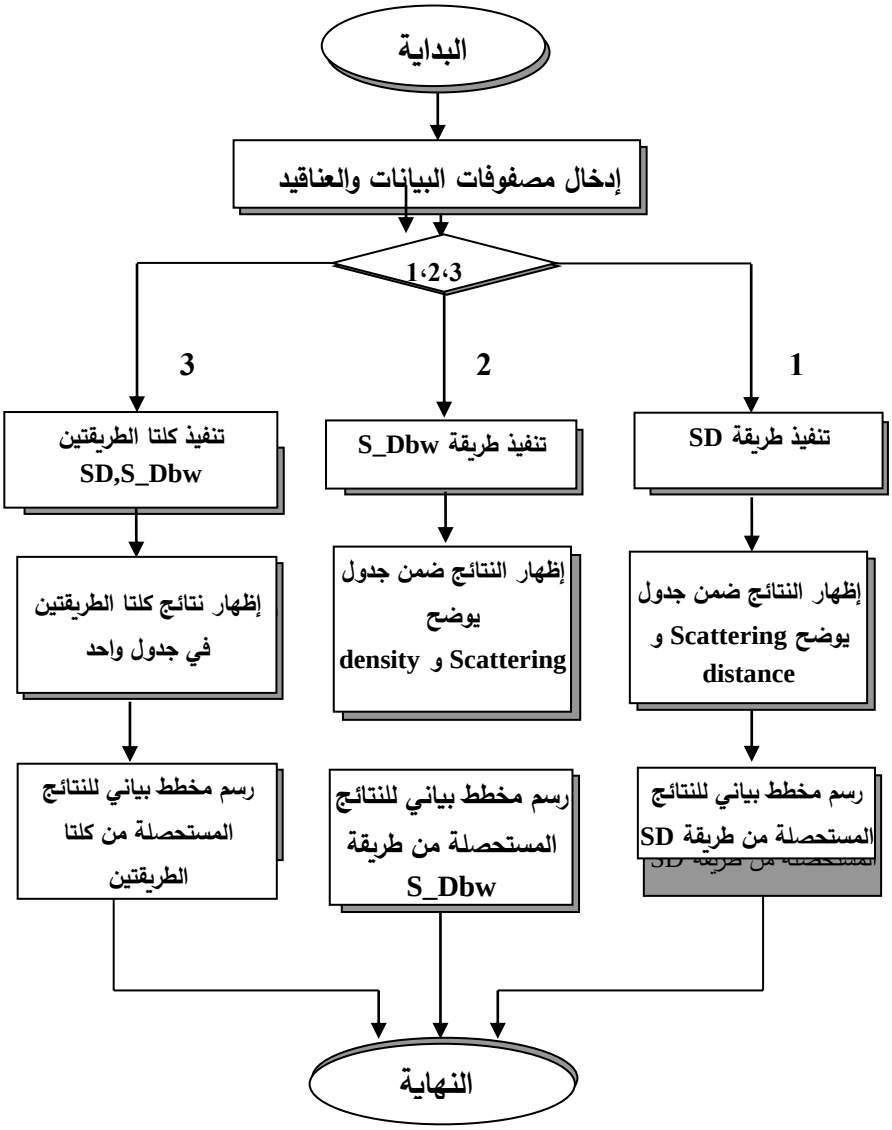
الذرة .. الخ) لتنفيذ النموذج المقترح وهي البيانات المستخدمة في الأطروحة [2] ، واعتمدت على أصناف الحبوب بوصفها عناقيد لمجموعة البيانات المختارة. الجدول (1) يوضح مراكز العناقيد، والشكل (5) يوضح خطوات عمل البرنامج.

الجدول (1) يمثل عينة من مراكز العناقيد للأصناف المختارة

Feature attributed

الحبوب	البروتين %	الدهون%	كربوهيدرات ذائبة %	الياف خام %	مواد معدنية %
الحنطة الانكليزية	10.5	2.6	78.6	2.5	1.8
ذرة صفراء حلوة	12.1	9.1	74.5	2.2	2.0
ذرة بيضاء	12.4	3.6	79.7	2.7	1.7
الشعير	11.8	1.8	78.1	5.3	3.1
الرز الخام (الشلب)	9.1	2.2	71.2	10.2	7.2
حبة الشوفان الكاملة	11.6	5.2	69.8	10.4	2.9
الشليم	13.8	1.4	79.7	2.6	2.2

Vector table instance record



الشكل (5) مخطط عمل البرنامج

6.1 خوارزمية SD validity index [11]:

ومؤشر الصحة (SD) معرف بالاعتماد على مفهومين اساسين هما معدل التشتت الداخلي للعناقيد والفصل الكلي بين العناقيد والتعاريف الاساسية لهذا المؤشر موضحة بالاتي :

أ. **معدل تشتت العناقيد Average scattering for clusters**

ومعرف بالمعادلة الآتية:

$$Scatt(n_c) = \frac{1}{n_c}\sum_{i=1}^{n_c} \| \sigma(v_i) \| / \| \sigma(X) \| \quad \text{... (1)}$$

حيث $\sigma(v_i)$ تمثل تباين العنقود (Vi) (variance of cluster) ويقاس بالمعادلة الآتية

$$\sigma_{v_i}^{p} = \sum_{k=1}^{n_i} (x_k^p - v_i^p)^2 / n_i \quad \text{....(2)}$$

حيث ان P هي عبارة عن الصفات.

اما σ(x) فتمثل التباين داخل مجموعة البيانات ويعرف بالمعادلة الآتية:

$$\sigma_{v_i}^{p} = \frac{1}{n}\sum_{k=1}^{n_i} (x_k^p - \bar{x}^p)^2 \quad \text{....(3)}$$

حيث ان $\bar{x}^p$ هي معدل الصفات (p) لمجموعة البيانات X ويعرف بالمعادلة الآتية:

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n_i} x_k, \forall x_k \in X \quad \text{....(4)}$$

حيث ان nc = عدد العناقيد

اما n = عدد البيانات في مجموعة البيانات

ب. الفصل الكلي بين العناقيد Total separation between cluster

الفصل الكلي بين العناقيد يعرف بالمعادلة الآتية :

$$Dis(nc) = \frac{D_{\max}}{D_{\min}} \sum_{k=1}^{nc} \left(\sum_{z=1}^{nc} \| v_k - v_z \| \right)^{-1} \quad \text{....(5)}$$

حيث ان Dmax = max(||Vk-Vz ||) وهي تمثل ابعد مسافة بين مراكز العناقيد عندما j,i تنتمي الى مجموعة العناقيد (1..nc).

أما Dmin = min(||Vk-Vz ||) التي تمثل اقل مسافة بين مراكز العناقيد عندما z,k تنتمي إلى مجموعة العناقيد (1..nc). و Vk,Vz فهي مراكز للعناقيد Ck,Cz

الآن يمكن تعريف مؤشر صحة العنقدة (SD) بالاعتماد على المعادلات من التعريفين السابقين:

SD(nc)= a . Scat(nc)+ Dis (nc) ....(6)

حيث ان a تمثل عامل الوزن الذي يساوي Dis(Cmax) ، وان Cmax يمثل اكبر عدد من العناقيد المدخلة (مجموع المسافات بين العناقيد).

المتغير الأول Scat(nc) والمعرف بالمعادلة الأخيرة يشير إلى معدل الدمج داخل العناقيد وهو يشير إلى (intra-cluster distance) والقيمة الصغيرة لهذا المتغير تشير إلى الترابط الجيد للعناقيد ، أما زيادة القيمة لهذا المتغير Scat(nc) فسوف تشير إلى زيادة التشتت ضمن العناقيد.

المتغير الثاني Dis(nc) يشير إلى الفصل الكلي بين العناقيد (nc) وهو يشير إلى (inter-cluster distance) . وبما أن هذين المتغيرين يقاسان بمعدلات مختلفة لذلك سنحتاج إلى استعمال المعامل (a) الذي يمكننا من دمج كلا المتغيرين في طريقة متوازنة ، والذي يتأثر تأثيراً كبيراً بعدد العناقيد (nc) .

ومن هذا المؤشر يتم استنتاج انه كلما كانت قيمة التشتت Scat(nc) صغيرة وكانت قيمة المسافة Dis(nc) صغيرة أدى ذلك إلى العنقدة المثلى . [6]

6.2 خوارزمية S_Dbw [14]:

هذا المؤشر سوف يستقبل الصفات الموروثة لتقييم صحة نتائج العنقدة واختيار أفضل تقسيمات ملائمة لمجموعة البيانات . إن هذا المؤشر يشبه المؤشر السابق (SD) من حيث التعريف إذ انه يعتمد على مفاهيم الدمج والفصل بين العناقيد ولكنه يأخذ بنظر الاعتبار الكثافة أيضا. هنالك مقياسان أساسيان في هذه الخوارزمية وهما:

أ. مقياس التشتت لكل عنقود باستخدام مقياس التباين داخل العنقود (intra-cluster variance).

ب. مقياس الكثافة بين العناقيد باستخدام مقياس الكثافة بين العناقيد (inter-cluster density).

وفيما يأتي شرح لكل منهما:

أ. مقياس التشتت ضمن العناقيد intra–cluster variance هو مقياس التشتت ضمن العناقيد والذي تم شرحه في المقياس (SD) (2 الفقرة أ).

ب. مقياس الكثافة بين العناقيد inter-cluster density هذا المقياس هو لحساب معدل الكثافة في المنطقة المحصورة بين العناقيد مع كثافة العناقيد نفسها. والغاية هو أن تكون الكثافة بين العناقيد قليلة بالمقارنة مع كثافة العناقيد نفسها. وهذا سوف يعرف بالمعادلة الآتية :

$$Dens\_bw(c) = \frac{1}{c.(c-1)} \sum_{i=1}^{c} \left(\sum_{\substack{j=1 \\ i \neq j}}^{c} \frac{denity(u_y)}{\max\{density(v_i), density(v_j)\}} \right) \quad \ldots(7)$$

حيث أن:

Vi,Vj : هما مراكز للعناقيد Ci,Cj على التوالي .

Uij : هي النقطة الوسطى لقطعة المستقيم بين مراكز العناقيد Vi,Vj .

C : هي عدد العناقيد .

أما الكثافة density(U) فمعرفة بالمعادلة الآتية

$$density(u) = \sum_{l=1}^{n_y} f(x_1, u), \qquad \dots (8)$$

حيث أن

nij : هي عدد الأزواج التي تنتمي إلى العناقيد Ci,Cj ، والتي تمثل عدد النقاط المجاورة لـ (U).

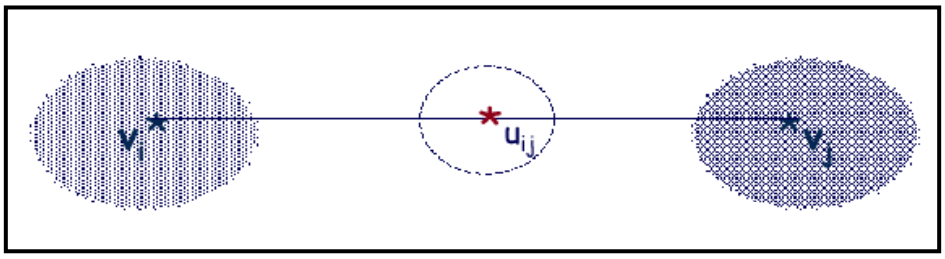
Xl : هو قيمة تنتمي الى Ci,Cj.

Ci,Cj: هي مجوعة جزئية من مجموعة البيانات S .

ثم أن الدالة f(x,u) معرفة كالآتي :

$$f(x,u) = \begin{cases} 0, & if\ d(x,u) > stdev \\ 1, & otherwies \end{cases}$$

في هذه الطريقة فان التجاور لنقاط البيانات U معرفة كجسم كروي والذي مركزه Uij ونصف قطره هو الانحراف المعياري القياسي(stdev) للعناقيد. النقطة تكون مجاورة لـ (U) في حالة كون المسافة بينهما اقل من (stdev) . هنا سوف نفترض أن قيم البيانات المجاورة لـ (U) هي قيم محسوبة من متجهات الكائنات التابعة للعناقيد Ci,Cj على التوالي لتكوين كائن ينتمي إلى النقطة (U) . وهكذا لبقية العناقيد . والشكل (6) يوضح ذلك



الشكل (6) يوضح مراكز العناقيد Vi,Vj والخط والواصل بينها ونقطة الوسط Uij

أما مصطلح stdev فهو معدل الانحراف القياسي للعناقيد

$$stdev = \frac{1}{c}\sqrt{\sum_{i=1}^{n_c} \| \sigma(v_i) \|} \quad \ldots(9)$$

فضلا عن ذلك فان النموذج (||X||) يعرف كالآتي :

$$||X|| = (XTX)1/2 \quad \ldots(10)$$

حيث أن X متجه.

و XT معكوس المتجه X.

على افتراض أن D={Vi: i=1,.. c} هي مجموعة مراكز أجزاء في مجموعة البيانات S والتي عددها (c) من العناقيد الكروية الشكل ، حيث أن Vi هي مركز للعنقود Ci والناتج من تطبيق خوارزمية العنقدة على مجموعة البيانات S.

لذلك فان مؤشر الصحة SD_bw سيحسب بالمعادلة الآتية

$$S\_Dbw(c) = Scat(c) + Des\_bw(c) \quad \ldots(11)$$

يشير هذا التعريف إلى كلا المقياسين الأساسيين للعنقدة الجيدة وهما " الترابط والفصل" . ويجمعها كما ينبغي بدقة ليمكننا من حسابات موثوقة لنتائج العنقدة ، كذلك تباين الكثافة بين العناقيد أُخذت بالحسبان للوصول إلى أكثر النتائج وثوقيةً .

إن أفضل عدد للعناقيد (c) هو الذي يقلل قيمة المؤشر S_Dbw إلى اقل ما يمكن ، إذ ان اقل قيمة لهذا الموشر مع عدد العناقيد تشير إلى العدد الأمثل من العناقيد الملائمة لمجموعة البيانات .

7. مناقشة النتائج:

7.1 نتائج خوارزمية SD:

بعد تنفيذ خوارزمية SD على مجموعة البيانات ومصفوفة العناقيد التابعة لها كانت نتائج خوارزمية S_Dbw كما في الجدول(3)

الجدول (3) نتائج خوارزمية SD

No.of clusters	Scattering	Distance Dis(c)	SD value
2	0.0216383	0.00985	1.723
3	0.0015355	0.00365	0.218
4	0.000283	0.00190	0.123
5	0.000080	0.00046	0.106
6	0.000029	0.000109	0.103
7	0.000012	0.000063	0.101

نلاحظ من الجدول المذكور آنفا إن قيم التشتت تقل عند زيادة عدد العناقيد .
أما المسافة النسبية Dis(c) فإنها تقل كلما زاد عدد العناقيد وهذا يعني أنه عند زيادة عدد العناقيد إلى هذا العدد فان المسافة بين مراكز العناقيد والعناقيد نفسها تكون اقل ما يمكن وهذا يعد أفضل عنقدة لمجموعة البيانات تحت الاختبار .
والقيمة النهائية لهذا المؤشر (SD) تقل مع زيادة عدد العناقيد . لذا فقد ظهر بان أفضل عدد عنقدة هو عندما يكون عدد العناقيد =7 .

7.2 نتائج خوارزمية S_Dbw:

بعد تنفيذ خوارزمية S_Dbw على مجموعة البيانات ومصفوفة العناقيد التابعة لها وكان عدد العناقيد مختلفاً في كل تنفيذه ومراكز عناقيد مختلفة أيضا فان نتائج خوارزمية S_Dbw موضحة في الجدول (4).

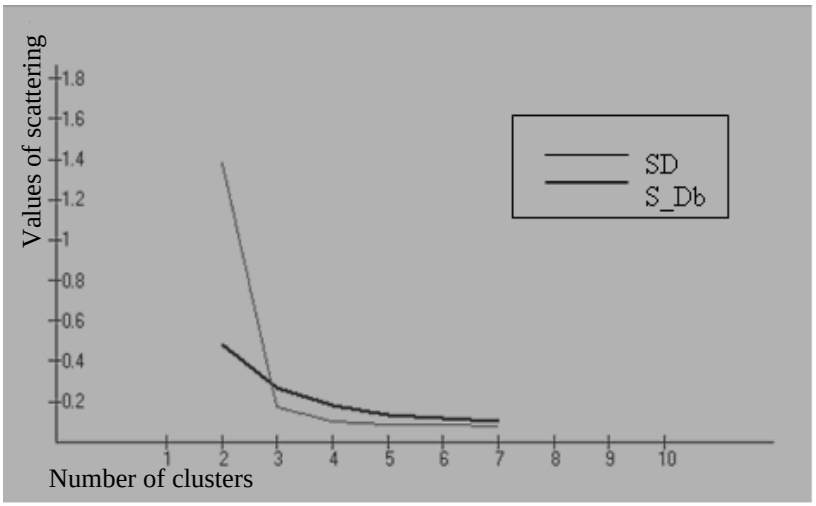
الجدول (4) نتائج خوارزمية S_Dbw

No.of clusters	Scattering	Density	S_Dbw value
2	0.0216383	15	0.599
3	0.0015355	15	0.334
4	0.000283	15	0.226
5	0.000080	16	0.168
6	0.000029	16	0.142
7	0.000012	16	0.123

نلاحظ من خلال الجدول المذكور آنفاً إن قيم التشتت تقل عند زيادة عدد العناقيد ، إذ أن عدد العناقيد عندما يكون ما بين (5−7) يكون أكثر ملاءمة لتقسيم مجموعة البيانات بهذا العدد واقل قيمة للتشتت هو عندما يكون عدد العناقيد (7) .

وكذلك فان كثافة العناقيد Density تزداد عندما يكون عدد العناقيد بين (5−7)، وهذا يؤدي إلى انخفاض قيمة الكثافة النسبية المحسوبة. والقيمة النهائية لهذا المؤشر (S_Dbw) تقل مع زيادة عدد العناقيد . لذا ظهر بان أفضل عدد عنقدة هو عندما يكون عدد العناقيد =7 في مجموعة البيانات تحت الاختبار.
الشكل (7) يوضح منحني نتائج الخوارزميتين S_Dbw و SD إذ يلاحظ فيه الاختلاف بين الخوارزميتين في قيمة التشتت، إذ تصل إلى اقل قيمة عندما يكون عدد العناقيد =7. ويلاحظ كذلك

في المخطط أن خوارزمية SD لها انحدار حاد في قيمة التشتت عند زيادة عدد العناقيد، وتلتقي مع خوارزمية S_Dbw التي يظهر انحدارها تدريجيا في قيمة التشتت.



الشكل (7) منحني كفاءة الخوارزميتين (SD وS_Dbw)

8. الاستنتاجات:

إن التقنيات الحديثة ومنها تقنيات التنقيب في البيانات وخوارزميات التصنيف التي لها دور كبير في تصنيف مجموعة البيانات إلى أصناف محددة لاكتشاف النماذج المخفية في البيانات والتي تعمل بدون مشرف ، قد تؤدي إلى الاستغناء عن مجاميع من البيانات المأخوذة من مخزن البيانات والاحتفاظ بالأصناف فقط والاستفادة منها، وذلك لكي تكون هذه المعلومات مفيدة ، وكون هذه المعلومات مفيدة يعتمد على أساليب تحليل هذه المعلومات لاكتشاف الأنماط والعلاقات الداخلية التي تحتوي عليها . وخوارزميات العنقدة هي الوسيلة الآلية للوصول إلى جوهر العلاقة بين مجموعة البيانات بعضها ببعض.

وبما أن هذه الخوارزميات تعمل من دون مشرف فانه لا بد من اكتشاف دقة هذا التصنيف وقوته من خلال خوارزميات صحة العنقدة (validity clustering algorithms). ومن هذه الخوارزميات خوارزمية (S_Dbw) و خوارزمية (SD) إذ استعملتا في هذا البحث ، فالخوارزمية الأولى (S_Dbw) سهلة ويسيرة ونتائجها جيدة ولاسيما أنها تستعمل مبدأ الكثافة ضمن العناقيد ويكون انحدار القيم فيها تدريجيا موازنة بالخوارزمية الثانية (SD) إذ كان الانحدار فيها حاداً .

المصادر

[1] العسيري ، د. ابراهيم محمد عبد الله (2001)، ملخص لأطروحة دكتوراه بعنوان " بحث وتصميم وانجاز نظام جديد للمعلومات التربوية بالمملكة العربية السعودية متضمناً دعم القرارات وتقنية مخزن البيانات " ، مجلة كلية المعلمين / الإصدار الثاني المجلد الأول

[2] الفخري، نعمة عبد الله (2003) " استخلاص نموذج بياني من قاعدة بيانات باستخدام خوارزميتي K-Means و IBK" رسا لة ماجستير ، كلية علوم الحاسبات والرياضيات / قسم علوم الحاسبات جامعة الموصل .

[3] جريمس ،براد ، (2003)، "التنقيب في البيانات -دراسة حالة :موقع اكسبوكس" ، مجلة pc 1اكتوبر 2003.

[4] رمال ،د.محمود استاذ مساعد الجامعة اللبنانية (2002) ، "دور التقنيات الحديثة لقواعد المعلومات في بناء مجتمع المعلومات العربي" ندوة المعلومات الخامسة التي أقيمت في النادي العربي للمعلومات في الفترة من 2-2002/7/4 ، النادي العربي للمعلومات .

[5] قطيشات ، د. منيب ، (1999)، "قواعد البيانات " ، الاكاديمية العربية للعلوم المالية والمصرفية ،ص331. الاردن

[6] Abhay S. Harpale , (2003) , “Clustering” , Indian Institute of technology , Bombay, powai .

[7] Christos Amanatidis, Maria Halkidi, Michalis Vazirgiannis, (2001), “Uminer: A Data mining system handling uncertainty and quality” , Dept of Informatics, Athens Univ. of Economics and Business

[8] Hsiao-Fan Wang (2003) , “Beyond Optimal Clustering” , NEWSLETTER , Operational Research Society of New Zealand, Inc. Auckland, New Zealand.

[9] I. K. Ravichandra Rao ,(2003), “ Data Mining and Clustering Techniques” , Professor and Head ,Documentation Research and Training Center , Indian Statistical Institute , Bangalore .

[10] Jiawei Han , Micheline Kamber ,"data mining : concept and techniques" , (2000) , intelligent data base system research lab., school of computer science , simon fraser university , Canada.

[11] M. Halkidi, Y. Batistakis, M. Vazirgiannis , (2000), "Quality scheme assessment in the clustering process" , Athens Univ. of Economics & Business.

[12] M. Halkidi, Y. Batistakis, M. Vazirgiannis , (2001),"Clustering algorithms and validity measures" , Department of Informatics , Athens University of Economics & Business

[13] Maria Halkidi ,Yannis Batistakis , Michalis Vazirgiannis , (2001), "On clustering validation techniques" , Department Of Informatics, Athens University Of Economics & Business, Patision 76, 10434, Athens, Greece (Hellas)

[14] Maria Halkidi Maria Halkidi Michalis (2001), "Clustering Validaty Assessment : Finding The Optimal Partitioning Of Data Set" , Dept. Of Informatics Dept. Of Informatics, Athens Univ. Of Economics & Business.

[15] Maria Halkidi, Michalis Vazirgiannis,(2001), "A data set oriented approach for clustering algorithm selection" ,Department of Informatics, Athens University of Economics & Business , Patision 76, 10434, Athens, Greece (Hellas)

[16] Maria Halkidi, Yannis Batistakis, Michalis Vazirgiannis (2001), "Cluster Validity Methods : Part I" , Department of Informatics, Athens University of Economics & Business

[17] Ying Zhao and George Karypis (2002), "Criterion Functions for Document Clustering" ,Experiments and Analysis , University of Minnesota, Department of Computer Science / Army HPC Research