**الجمهورية العربية السورية**

**جامعة دمشق**

**كلية الاقتصاد**

**الدراسات العليا**

**ماجستير إدارة أعمال**

**إعداد الطالبة : لـــــــيـــــــــرا وليد غنّام إشراف الدكتور : سليمان عوض**

***[[1]](#footnote-2) What is Data Mining?***

******

**التنقيب في البيانات**: هي عملية البحث داخل مستودع البيانات عن معرفة غير مستكشفة ( اكتشاف المعرفة).

**التنقيب في البيانات** : استطلاع وتحليل كميات كبيرة من البيانات لاكتشاف علاقات وقواعد ذات معنى بين المتغيرات

***لماذا التنقيب في قواعد البيانات؟***



***Why Data Mining?***

***نحن نغرق في البيانات ونتعطش للمعرفة***

إن التنقيب في قواعد البيانات يهدف إلى استخلاص المعلومات المخبأة فيها**،**

وهي تكنولوجيا حديثة فرضت نفسها بقوة في عصر المعلوماتية

وفي ظل التطور التكنولوجي الكبير وانتشار استخدام قواعد البيانات.

استخدامها يوفر للمؤسسات في جميع المجالات القدرة على استكشاف، والتركيز على، أهم المعلومات في قواعد البيانات.

تركز تقنيات التنقيب على بناء التنبؤات المستقبلية واستكشاف السلوك والاتجاهات، مما يسمح بتقدير القرارات الصحيحة واتخاذها في الوقت المناسب.

تجيب تقنيات التنقيب على العديد من الأسئلة، وفي وقت قياسي، بخاصة تلك النوعية من الأسئلة التي كان من الصعب الإجابة عليها، إن لم يكن مستحيلاً، باستخدام تقنيات الإحصاء الكلاسيكية، والتي كانت إن وجدت فإنها تستغرق وقتاً طويلاً والعديد من إجراءات التحليل

مثال : ليكن لدينا قاعدة بيانات تحتوي على 1000000 اسم لأصحاب البطاقات الائتمانية

ما يهم إدارة العلاقات مع الزبائن هو معرفة أي من الزبائن يحتمل أن يكونوا أكثرً ولاء وأيهم على الأغلب سيتجه إلى المنافسين ؟

التنقيب في البيانات يساعد في الإجابة على تلك التساؤلات من خلال انتزاع المعلومات المتعلقة بالزبائن



***التقنيات الحديثة للتنقيب في قواعد البيانات***

أساس التنقيب هو التصنيف والتنبؤ

وهذه التقنيات مستمدة من علم الإحصاء وعلوم تعلم الآلة وعلوم الحاسوب

* الجار الأقرب
* التجزئة العنقودية : تحليل المجموعات المتجانسة وهي من أساليب التحليل المتعددة المتغيرات تقوم على أساس تعظيم التجانس داخل كل مجموعة وتعظيم التباين بين المجموعات المختلفة
* شجر القرار
* الشبكات العصبية (( تعلم الآلة )) هنا تعطي أوزان ترجيحية
* استقراء القاعدة
* وتحليل الانحدار والسلاسل الزمنية
* اختبار الفروض أسلوب إحصائي يقوم على اختبارات معلمية ولا معلمية

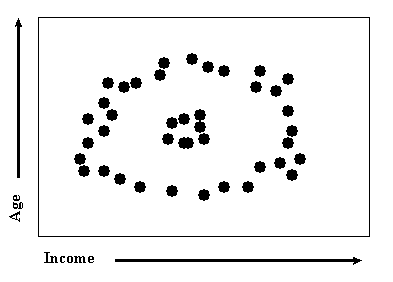
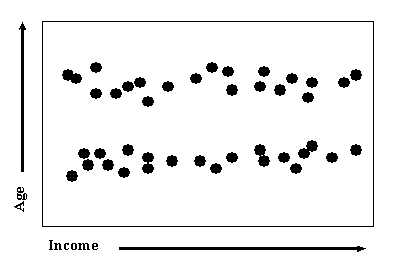
***- خوارزمية الجار الأقرب***

تعتبر من تقنيات التنقيب في البيانات التي تهدف للتنبؤ عن طريق مقارنة السجلات الشبيهة بالسجل المراد التنبؤ له وتقدير القيمة المجهولة لهذا السجل بناء على معلومات لتلك السجلات.

كثيراً ما تستخدم خوارزمية الجار الأقرب في مجال الأعمال، ومن الأمثلة الشائعة الاستخدام تلك التي تساعد المستخدمين في الشراء عن طريق اختيار السلع الأقرب لاحتياجاتهم مقارنة بسلع قد تم شراؤها بالفعل.

***- التحليل بالتجزئة العنقودية***هي عملية تجميع السجلات المتشابهة في مجموعات، ويتم ذلك بهدف الاستكشاف عالي المستوى لما يجري داخل قاعدة البيانات.

|  |  |  |  |
| --- | --- | --- | --- |
| ***الزبائن السابقون***   * ***العمر*** * ***الرتب*** * ***المهنة*** * ***نوع الزبون*** |  | **قواعد القرار**  **الراتب أقل من 5**  **بيانات الزبون الجديد** | **Good / Bad** |

****

|  |
| --- |
|  |

***تجانس المجموعات التي يتم إنشاءها***

القاعدة العامة لضم أي عنصر في مجموعة هي أن يكون العنصر مائلاً للتشابه بعنصر منها أكثر من أن يكون شبيهاً لعنصر من مجموعة أخرى.

بشكل عام، تهدف التجزئة العنقودية إلى وضع العناصر المتجانسة في مجموعات منفصلة.

***- شجر القرار***

شجرة القرار هي نموذج استكشافي يظهر على شكل شجرة، كما يعبر اسمها، وبشكل دقيق يمثل كل فرع من فروعها سؤالاً تصنيفياً وتمثل أوراقها أجزاء من قاعدة البيانات تنتمي للتصنيفات التي تم بنائها.

تقسيم السلوك الشرائي في أحدى المناطق باستخدام خوارزمية شجر القرار

***استخدام شجرة القرار في التنبؤ***

القاعدة الأساسية في بناء شجرة القرار هي إيجاد أفضل سؤال عند كل فرع من فروع الشجرة بحيث يقسم هذا السؤال البيانات إلى قسمين، القسم الأول منها ينطبق عليهم السؤال والقسم الثاني لا ينطبق، وهكذا يتم من خلال سلسلة من الأسئلة بناء شجرة القرار بفروعها المتسلسلة

. بالرغم من أن شجرة القرار تستخدم في الاستكشاف وتحضير البيانات للعمليات الإحصائية إلا أنها أيضاً تستخدم وبشكل أكثر للتنبؤ. ومن المهم جداً عند بناء خوارزمية شجرة القرار أن يؤخذ بعين الاعتبار أن تكون قابلة للتطبيق بقدر الإمكان وبشكل مثالي على كل البيانات المتوفرة.

***الفرق بين شجرة القرار والتجزئة العنقودية***

تهدف تقنية شجرة القرار إلى تقسيم قاعدة البيانات بهدف معين سبق وأن تم تحديده.

وجود عنصر معين في إحدى المجموعات، وهي ممثلة هنا بالفروع، هو نتيجة لأنه حقق سلسلة الشروط الموضوعة وصولاً إلى هذا الفرع، وليس فقط لأنه يشبه بقية عناصره، بالرغم من أنه لم يتم تعريف التشابه في هذه الحالة.

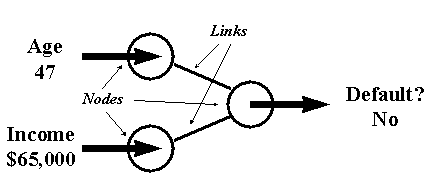
قد تكون شجرة القرار أكثر تعقيداً من التجزئة العنقودية ولكنها تؤدي إلى نتائج يمكن إظهارها بشكل مبسط وفائدة عالية المستوى.

***– الشبكات العصبية***تعتبر الشبكات العصبية هي وأشجار القرار من أهم تقنيات التنقيب في البيانات، نظراً للنتائج الدقيقة التي يتم التوصل إليها باستخدام هذه الخوارزميات ولإمكانية تطبيقهما في حل العديد من المشاكل وبكافة الأنواع، هذا بالرغم من صعوبتهما والتي أدت لعدم الانتشار بشكل واسع لهما.

خوارزمية الشبكة العصبية تشبه في تركيبتها تركيبة مخ الإنسان، فهي تعمل بنفس الطريقة كما يعمل المخ في نقل ومعالجة المعلومات والتوصل إلى الاستنتاجات واكتشاف الأنماط والتنبؤات ونستطيع من خلالها تطبيق بعض ما يطبقه المخ الطبيعي، رغم أن العلماء لا يزالون حتى اليوم يكتشفون المزيد ولم يلموا بكل تفاصيل عمل مخ الإنسان

***طريقة عمل خوارزمية الشبكة العصبية***

تتألف الشبكة العصبية من العقد (التي تناظر الخلايا العصبية) والروابط التي تصل بينها (التي تناظر الوصلات العصبية).

الشكل التالي يوضح تركيب شبكة عصبية بسيطة، وهي تأخذ متغيري العمر والدخل وتعطي نتيجة تنبؤية ما إذا كان الشخص سيقبل على ارتكاب فعل معين : ****

**مثال تطبيقي**

لكي يتم التنبؤ باستخدام الشبكة العصبية، يتم إدخال قيم المتغيرات المعلومة في العقد المخصصة للإدخال، ويصبح لكل عقدة قيمة المتغير الذي تم إدخاله، بعد ذلك يتم ضرب قيمة كل عقدة بقيمة الرابط المتصل بها لنحصل على النتيجة.

وهنا تم اعتبار أنه إذا كانت النتيجة 1 يكون من المتوقع إقدام الشخص على شراء السلعة المدروسة ، وإذا كانت 0 فيكون من المتوقع عدم الاستجابة.

تم التعبير عن العمر بقيمة تقع بين 0.0 و 1.0 وهي هنا 0.47، وهذا يمثل عمر 47 سنة، والدخل بالقيمة 0.65، وهذا يمثل دخلاً بقيمة 65000$.

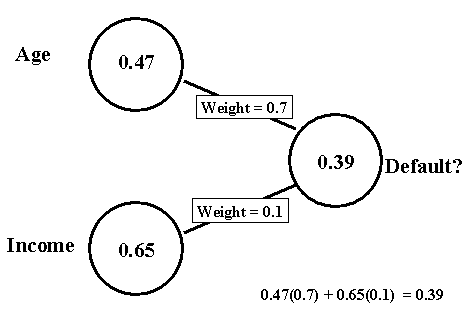
بالنسبة للوصلات والتي تعبر عن الأوزان فقد تم تقديرها بالقيم 0.7، 0.1 على الترتيب وذلك بناء على معرفتنا بسجلات سابقة من قاعدة البيانات.

بعد ضرب قيم العقد في قيم الوصلات وجمعها نحصل على قيمة المتغير الذي نريد التوقع له فيكون هنا 0.39 وهو رقم أقرب للصفر منه للواحد الصحيح وبذلك تكون النتيجة هي عدم احتمال شراء الشخص للسلعة

شكل توضيحي للمثال

الشكل التالي يبسط ما تم عمله في المثال السابق



****

***العقد المخفية في الشبكات العصبية***

قد تحتوي خوارزمية الشبكة العصبية على نوعية أخرى من العقد والتي تسمى العقد المخفية.

مهمة هذه العقد استشارية ولا يؤخذ بقيمها إلا بعد أن يتم اعتماد استشارتها في حالة صحتها وبعد التجربة الفعلية.

ومثلما يحدث في المنظمات فالمدير يستمع إلى العديد من الاستشارات ممن حوله من المستشارين قبل اتخاذ قرار معين، ولكنه بعد اتخاذ القرار واستكشاف نتائجه ومدى صحته، يصبح بإمكانه تمييز المستشارين الجيدين والذين كانت آرائهم أقرب للقرار الذي كان من المفترض أن يكون أنسب، وبالتالي سوف يعتمد آرائهم في المستقبل ويأخذ بها أكثر من آراء غيرهم، وهكذا، فالعقد المخفية تلعب نفس هذا الدور، كلما تم تطبيق الخوارزمية يتم تطوير وتحديث العقد الأصلية بأن تأخذ بالاعتبار قيم العقد المخفية المناسبة والتي تدعم الحصول على نتائج أكثر دقة، وبالمقابل يتم إهمال قيم العقد المخفية التي لم تحقق ذلك.

***– استقراء القاعدة***أن تقنية استقراء القاعدة من التقنيات الأساسية في التنقيب في **البيانات** وأكثرها شيوعا في مجال استكشاف المعرفة، وهي أقرب ما تكون إلى ما يسمى بعملية التنقيب بحد ذاتها، والذهب في هذه الحالة هو "القاعدة”.

تبين هذه القاعدة ما يجري داخل قاعدة البيانات وتظهر لنا ما لم نكن نعرفه من قبل، وربما أيضاً ما لن نستطيع أن نعرفه إلا من خلالها.

***كيف نستكشف القاعدة***

في تحليل قاعدة بيانات سجلات الزبائن في المنظمة يمكن أن نستكشف القاعدة الشيقة التالية:

"إذا قام المستهلك بشراء السلعة من النوع A فإنه يقوم بشراء السلعة من النوع B معها باحتمال 80%، وهذه الثنائية تحدث بإجمالي 3% من كافة المبيعات المسجلة" .

ولكي تكون القاعدة مكتملة وذات فائدة، فإنه يلزم لها تقييم، وهو عبارة عن نوعين إضافيين من المعلومات التي يجب أن تلازمها، وهذه المعلومات الإضافية هي :

الصحة : كم هي نسبة صحة القاعدة (وقوع النتيجة في حال وقوع السبب).

التغطية : كم نسبة السجلات المحققة للقاعدة إلى كافة السجلات في قاعدة البيانات

***مثال تطبيقي لطريقة تقييم القاعدة***

إذا اشترى المستهلك 3 قطع من النوع الأول فإنه سيشتري كامل المجموعة التي يشكل النوع الأول عنصر فيها

في أحد أمثلة تحليل قاعدة بيانات الشراء ، ليكن لدينا القاعدة التالية**:**

**والأعداد التالية كما يلي:**

ت= 100 : العدد الكلي للسجلات في قاعدة البيانات

ك  = 30  : عدد سجلات المستهلكين الذين اشتروا كامل المجموعة

ل = 40  : عدد سجلات المستهلكين الذين اشتروا 3 قطع من النوع الأول

ب = 20  : عدد سجلات المستهلكين الذين اشتروا 3 قطع من النوع الأول و اشتروا كامل المجموعة

فتكون نسبة صحة القاعدة هي حاصل قسمة عدد سجلات المستهلكين الذين اشتروا 3 قطع من النوع الأول و اشتروا كامل المجموعة معاً مقسوماً على عدد سجلات المستهلكين الذين اشتروا 3 قطع من النوع الأول وتكون في هذه الحالة مساوية لـ 40/20=50%.

أما التغطية فتكون حاصل قسمة عدد سجلات الذين اشتروا 3 قطع من النوع الأول مقسوماً على العدد الكلي للسجلات في قاعدة البيانات. وهي هنا تساوي 40/100 = 40 %.

القاعدة المفيدة وغير المفيدة

إن أكبر مشكلة يمكن مواجهتها في خوارزميات استقراء القاعدة هي كيف يمكن تمييز القاعدة المفيدة من القاعدة غير المفيدة.

أمثلة للقاعدة غير المفيدة :

أن لا تكون لها فائدة عملية

أن تكون صحتها نادرة.

أن تكون تغطيتها نادرة ولا يمكن تطبيقها.

أن توضح معلومات بديهية لا حاجة لها.

وقد أدت عمليات البحث إلى نتائج يمكن من خلالها تحديد مدى فائدة القاعدة من خلال العديد من الاعتبارات، وكان أهمها هو طبيعة القضية موضوع البحث.

من جهة أخرى، تعتبر القاعدة البسيطة والواضحة مفيدة أيضاً بعكس ما إذا كانت معقدة وصعبة الفهم، وكذلك يمكن اعتبار القاعدة مفيدة في حالة ما إذا كانت توضح علاقة فريدة.

***اختيار تقنية التنقيب المناسبة***

أخيراً، لا يوجد نظرية محددة يتم بناء عليها اختيار تقنية من تقنيات التنقيب، ويتم الاختيار عادة بناءاً على الخبرة في هذا المجال والتجربة الفعلية للتقنيات ومدى فاعليتها، ومن جهة أخرى قد تكون المفاضلة أيضاً بين التقنيات التقليدية والتقنيات الحديثة بقدر ما يكون هناك توفراً للأدوات المناسبة، ومع ازدياد الخبرة نستطيع أن نقيّم الخيارات ونحدد منها المناسب ونطبقه

***تخطيط عمليات التنقيب في قواعد البيانات***

إن تخطيط عمليات التنقيب في قواعد البيانات من الأمور المهمة للحصول على أفضل النتائج، فالتخطيط الجيد يؤدي للنتائج الجيدة.

ويمكن تلخيص الخطوات الأولية للتنقيب في قواعد البيانات في ما يلي:

1 – تحديد المشكلة المراد بحثها وإيجاد الحلول لها



2 – بناء قاعدة بيانات التنقيب

3 – استكشاف البيانات

4 – تحضير البيانات للتنقيب

5 – بناء نموذج التنقيب المناسب

6 – تطبيق النموذج

7 – استخراج النتائج

المراجع

Two Crows Corporation, *Introduction to Data Mining and Knowledge Discovery, Third*

*Edition* (Potomac, MD: Two Crows Corporation, 1999); Pieter Adriaans and Dolf Zantinge,

*Data Mining* (New York: Addison Wesley, 1996).

For a more technically-oriented definition of data mining, see

* + [http://searchcrm.techtarget.com/gDefinition/0,294236,sid11\_gci211901,00.html].
  + **www.pentaho.com**
  + **www.neurosolutions.com**
  + [**www.math.ntu.edu.tw**](http://www.math.ntu.edu.tw)
  + **www.chasmgroup.com**
  + **http://www.cs.sfu.ca/~han/dmbook**
  + **http://db.cs.sfu.ca/ or http://www.cs.sfu.ca/~han**
  + [**http://www.stat.wisc.edu/~limt/treeprogs.html**](http://www.stat.wisc.edu/~limt/treeprogs.html)
  + **سعد غالب ياسين ، مساندة القرارات دار المناهج**
  + **نمذجة البيانات باستخدام تقنيات التنقيب في أنظمة قواعد المعطيات ، محمد عبيد ، رسالة ماجستير**

1. [↑](#footnote-ref-2)