تـنمية الرافديـن 81 (28) 2006
ص ص [119–125]

استخدام المصنف C4.5 في تمييز سمة الكائن
دراسة مقارنة

غيداء عبدالعزيز الطالب
مدرس-قسم علوم الحاسبات
كلية علوم الحاسبات والرياضيات-جامعة الموصل

رائد عبدالقادر الدباغ
مدرس-قسم نظم المعلومات
كلية الادارة والاقتصاد-جامعة الموصل

نعمة عبدالله الفخري
مدرس مساعد-قسم نظم المعلومات
كلية الادارة والاقتصاد-جامعة الموصل

المستخلص

ان تعدين البيانات فعالية الحصول على المعرفة لتحقيق هدف أساس وهو اكتشاف الحقائق الخفية (Hidden Facts) التي تتضمنها قواعد البيانات وذلك من خلال استخدام تقنيات متعددة تشتمل على الذكاء الاصطناعي، التحليلات الاحصائية، تقنيات و نمذجة البيانات ... الخ.

فأن عملية تعدين البيانات تولد نماذج وعلاقات واضحة في البيانات والتي تساعد على توقع النتائج في المستقبل. وقد ظهرت العديد من الخوارزميات التي في هذا المجال، وترتب عليها مقارنة بين هذه الخوارزميات لاختبار الخوارزمية المناسبة في الحصول على نتائج أفضل.

وقد هدف البحث الى استخدام المصنف C4.5 وربطها مع الشبكة العصبية نوع Back – Propagation (BP) وذلك لتكوين نموذج تصنيف يحمل خواص الطرفين، فضلا عن مقارنة النتائج المستحصلة مع نتائج التصنيف باستخدام الحزمة البرمجية الجاهزة Minitab. وتوصل البحث الى ان المعادلات الخاصة بالمصنف C4.5 كانت أكفأ في الاداء وخاصة بعد ربطها بالشبكة العصبية BP لازالة التناقض والتشويش الموجود في البيانات، كما عززت النتائج من افضلية استخدام لغات البرمجة مقارنة بنتائج التطبيق الجاهزة.

1. المقدمة

ان التطور الحاصل في موضوع تعدين البيانات (Data Mining) في المجالات والصناعات المختلفة أدى الى ظهور العديد من الخوارزميات، الأمر الذي جعل من الأهمية اختيار خوارزمية تعدين مناسبة للحصول على نتائج أفضل وذلك بسبب تنوع البيانات واختلافها، فما يعمل جيداً على بيانات معينة قد لا يعمل بنفس الجودة على بيانات أخرى. وتأخذ عملية تعدين البيانات الاعتبارات آلاتية (الفخري،2003، 1-2):

أولاً: تمثيل المعرفة باستخدام خوارزميات خاصة، وهي خوارزميات واسعة ومتنوعة، فمنها ما يعمل بأسلوب شجرة القرار (Decision Tree)، ومنها ما يستخدم قاعدة اذا ... فان (if – then – rule) ... الخ .

تاريخ التسلم 2005/7/4 ______________ **تاريخ قبول النشر 2005/9/11**

ثانياً: كيف تستطيع الخوارزمية الوصول الى أعلى مقياس اعتماداً على فضاء البحث المتوفر لديها.

وبناءاً على ذلك ظهرت عمليات مقارنة بين الخوارزميات و على مديات مختلفة من البيانات كمحاولة لوضع خصائص لهذه البيانات، ثم مطابقة تلك الخوارزميات مع خصائص تلك البيانات، وهذه المعلومات تساعد في تعدين البيانات لصنع قرارات ذكية في اختيار الخوارزمية الملائمة لملفات البيانات .

تأسيساً على ما تقدم، فقد هدف البحث الى استخدام المصنف C4.5 بوصفه أحد خوارزميات نوع شجرة القرار وربطها مع الشبكة العصبية Back – Propagation (BP) وذلك لتكوين نموذج تصنيف يحمل خواص الطرفين، فضلا عن مقارنة النتائج المستحصلة مع نتائج التصنيف للحزمة البرمجية الجاهزة Minitab(*) سعياً لتحقيق فرضية البحث ومفادها "ان استخدام البرمجيات الجاهزة في التصنيف Classification يعتمد على اسلوب محدد باستخدام احدى خوارزميات التصنيف دون مقارنة النتائج مع بقية الخوارزميات، الامر الذي يقلل من أهمية استخدام لغات البرمجة في كتابة البرامج الأكفأ و الأفضل في التصنيف".

لقد اعتمد البحث في أسلوبه جانبين، الأول يمثل الجانب النظري و الذي يتم من خلاله وصف المصنف C4.5، فضلاً عن استخدام الشبكات العصبية الاصطناعية لغرض توليد البيانات بعد تغذية البرنامج بخصائص الحالات قيد الدراسة. في حين تناول الجانب العملي وصفا لنتائج تطبيق البرنامج على الحالات المولدة ومقارنتها مع نتائج التصنيف باستخدام الحزمة البرمجية الجاهزة Minitab .

وتم استخدام لغة البرمجة VB الاصدار 6.0 في كتابة البرامج كافة المستخدمة لتنفيذ خطوات الخوارزمية وربطها مع الشبكة العصبية BP .

2. الجانب النظري

1-2 تعدين البيانات Data Minig

يتميز عصرنا الحالي باستخدام تكنولوجيا البيانات المتطورة لحفظ و استرجاع البيانات وبكميات هائلة و التي توصف بمستودعات البيانات Data Warehousing. ان توفر هذه البيانات فتح الباب امام مجموعة من الموضوعات المتخصصة في ادارة تلك البيانات كان من ابرزها موضوع تعدين البيانات Data Minig والذي يعد من الاساليب المهمة للحصول على معلومات مفيدة من البيانات (Ibrahim, 1999,9).

وقد عرف العلماء مصطلح تعدين البيانات على انه "جزء من عملية اكتشاف المعرفة في قواعد البيانات و التي تتم باستخدام طرائق متعددة هدفها تكوين نماذج من البيانات" .

(*) الحزمة الجاهزة Minitab Release 13.0 والتي تعمل في بيئة النوافذ، و هي إحدى برامجيات لوائح العمل في مجال الإحصاء و الرياضيات .

وبصورة اخرى، فقد تم تعريف مصطلح تعدين البيانات بانه "عملية اكتشاف المعرفة وطريقة تحليلها من زوايا مختلفة و تلخيصها وتحويلها الى ما يسمى بـ (معلومات – معلومات) لتوضح امام صانعي القرار للعمل على اساسها في مجالات عدة مثل المحاسبة، الاتصالات، استخدام اوسع في المجالات الطبية، ... الخ و بشكل يعمل على زيادة الدخل او تقليل الكلف او كليهما معاً" .

وفي ضوء ما سبق من تعريفات، فانه يمكن القول بأن تعدين البيانات هو فعالية الحصول على المعرفة لتحقيق هدف اساس وهو اكتشاف الحقائق المخفية Hidden Facts التي تتضمنها قواعد البيانات وذلك من خلال استخدام تقنيات متعددة تشتمل على الذكاء الاصطناعي، التحليلات الاحصائية، تقنيات ونمذجة البيانات. ان عملية تعدين البيانات تولد نماذج وعلاقات واضحة في البيانات والتي تساعد على توقع النتائج في المستقبل.

اذن، لتعدين البيانات ادوات، وبرامجيات تعدين البيانات هي احدى هذه الادوات والتي تستخدم لتحليل البيانات من قبل المستخدم و تلخيص العلاقات التي تعرّفها، فضلا عن دورها في ايجاد روابط بين تقنية المعلومات و الانظمة التحليلية المستخدمة لتمثيل العلاقات بين نماذج البيانات المخزونة (www.Anderson.Ucla.edu).

وتقنياً، فأن تعدين البيانات هو عبارة عن إجراء أو معالجة لإيجاد الروابط بين مجاميع (الحقول/السجلات) في قواعد البيانات الكبيرة والتي تشترك فيها برامجيات أخرى فضلاً عن برامجيات تعدين البيانات كالبرامجيات الاحصائية، والشبكات العصبية، .... الخ. وبشكل عام فانه يمكن القول بأن أي من العلاقات الآتية مهمة في مجال تعدين البيانات:

1. الاصناف Classes: وتستخدم عادة لوضع البيانات المخزونة في مجاميع تم تحديدها مسبقا لبناء نموذج بالاعتماد على بعض المتغيرات المستقلة .
2. العناقيد Clusters : و تستخدم لوضع البيانات في مجاميع اعتماداً على العلاقات المنطقية. بعبارة اخرى، فان الخوارزميات المستخدمة للتصنيف في هذه الطريقة تسعى لتقسيم البيانات الى مجاميع (عناقيد) بحيث ان السجلات المتشابهة تقع في المجموعة نفسها وهذه المجاميع يجب ان تكون مختلفة عن بعضها قدر الامكان.
3. الروابط Associations : و هي تعرف العلاقات الخاصة بتعدين البيانات، اذ إن الخوارزميات المستخدمة فيها تنشئ قواعد لربط الحوادث التي تظهر سوية في البيانات .
4. النماذج المتسلسلة: يتم تعدين البيانات لتوقع سلوك واتجاهات النماذج المستحصلة (www.Anderson.Ucla.edu).

2-2 خوارزميات التصنيف Classification Algorithms

ان طرق تعدين Data Mining Methods هي عبارة عن مجموعة الاجراءات و الخوارزميات المصممة لتحليل البيانات المخزونة Data Baise، و هي تتعامل مع

عدة عوامل أهمها، أولاً: الدقة بين النماذج المتكونة والبيانات المتوفرة، وثانياً: تمثيل المعرفة باستخدام خوارزميات خاصة (www.towcrows.com).

وبما ان ايجاد شكل موحد يمثل المعرفة لجميع البيانات أمر صعب المنال، فقد أوجد الباحثون أشكالا للتصنيف غير معتمدة على تمثيل المعرفة أي انها مصنفات للاغراض العامة وهي تقع في نوعين، الأول يطلق عليها المصنفات التي تعمل بمشرف و الثاني يعمل بدون مشرف (Ibrahim, 1999,11-23)

2-2-1 المصنف C4.5 C4.5 Classification Algorithm

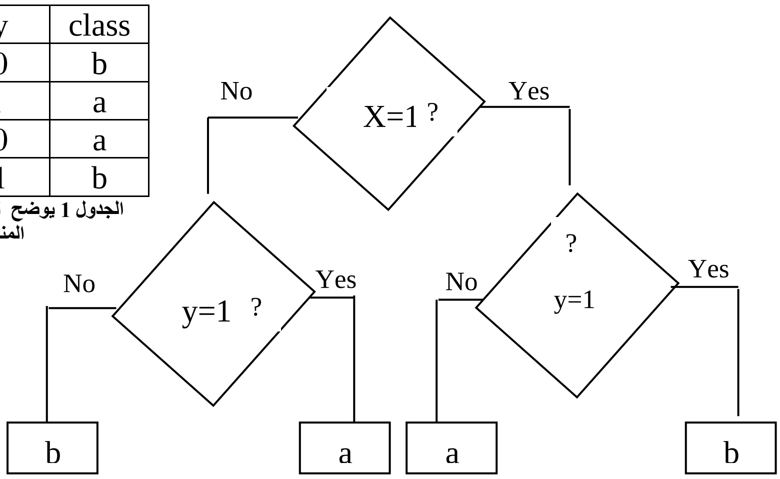
ان خوارزمية C4.5 هي احدى انواع الخوارزميات التي تعمل باسلوب شجرة القرار Decision Tree، حيث تتمثل فيها بتمثيل مجاميع القرارات و هذه القرارات تولد قوانين التصنيف الخاصة بمجاميع البيانات، وبشكل آخر فانها تعمل على تصنيف الحالات Instances الى فئات مختلفة، وهي من طرائق التصنيف شائعة الاستخدام.

تقع هذه الخوارزمية تحت مجموعة المصنفات التي تعمل بمشرف و التي تصنف حالات معينة الى عدد من الفئات (التصانيف) باسلوب فرق – تسد، اذ انها تجزئ المسألة المعقدة الى مسائل ابسط ثم يتم الاستدعاء الذاتي للدالة نفسها و لكل اجزاء المسألة، وبجمع حلول المسائل المجزأة يتم الحصول على حل المسألة المعقدة (Ali and Abraham , 2002, 2-3) .

ويمكن تشبيه عمل شجرة القرار في هذه الخوارزمية بعمل بوابة XOR المنطقية المتمثل بجدول الحقيقة الاتي : (keller,2000,8)

x	y	class
0	0	b
0	1	a
1	0	a
1	1	b

الجدول 1 يوضح نتائج بوابة XOR المنطقية



الشكل 1
عمل شجرة القرار للمصنف C4.5

اذ إن : العقد Nodes تمثل قرارات الخصائص Decisions On Attributes
الاوراق Leaves تمثل الاصناف Class**.**

2-2-2 مراحل عمل الخوارزمية C4.5

يمكن توضيح عمل الخوارزمية C4.5 بالمراحل الاتية:

1. تحديد التصنيف لكائن معين من خلال تنفيذ العلاقة الاتية:

$$I(p,n) = \frac{p}{p+n}\log\left[\frac{p}{p+n}\right] - \frac{n}{p+n}\log\left[\frac{n}{p+n}\right] \qquad ....(1)$$

اذ إن : p, n تمثلان أصناف مختلفة.
$P \neq n$

2. تعيين خاصية معينة واختيارها و لـ k من القيم وذلك باستخدام العلاقة الاتية:

$$E(A,p,n) = \sum_{i=1}^{k} \frac{p_i + n_i}{p+n} . I(p_i, n_i) \qquad ....(2)$$

اذ إن : A هي الخاصية التي تم اختيارها
p_i, n_i عدد الحالات لكل صنف من الشجيرات الناتجة عن شجرة القرار و تكون مرتبطة مع الجزء I.

3. الحصول على النسبة النهائية للمعلومات من المعادلة الاتية:

$$\text{Gain}(A,p,n) = I(p_i, n_i) - E(A,p,n) \qquad ....(3)$$

(Ali and Abraham, 2002, 3; Hidalgo and others, 2002, 327)

2 -1 الشبكات العصبية Neural Network

تميزت الشبكات العصبية ومنذ انتشارها في منتصف الثمانينات بامكاناتها العالية في اجراء المعالجة المتوازنة وعدم حاجتها الى علاقات معقدة في عملها، بل تحتاج الى بعض الامثلة لتعلمها فقط، ومن ثم سهولة الاجابة . وتعطي الاجابة الصحيحة للمدخلات التي ليست ضمن فقرة التدريب مع استبعاد ما يميل منها الى التناقض و التشويش.

واعتماداً على نوعية التطبيق فقد تم اختيار شبكة الانتشار الخلفي Back – Propagation (BP) في بناء النموذج، والتي تعد احدى الشبكات الواسعة الاستخدام والكفوءة في طرق التعلم و التي تستخدم لتدريب الشبكات متعددة الطبقات. وتعتمد هذه الشبكة في خوارزمياتها على القاعدة المعروفة باسم الانحدار التدريجي Stepwise Regression لمربع معدل الاوزان. ان تدرج الخطأ واوزان الشبكة يعطي الاتجاه الذي يتزايد فيه الخطأ بأسرع ما يمكن.

اما المعادلات المستخدمة في الإخراج و تصحيح الاوزان فتعطى بالاتي

$$\left.\begin{aligned} a &= f(w_1 * h) \\ h &= f(w_2 * e) \end{aligned}\right\} \quad ....(4)$$

اذ أن : a تمثل متجه الافراج، e تمثل متجه الادخال، h تمثل الطبقة المخفية (hidden Layer) و (W_1 , W_2) مصفوفتي الاوزان .

ومعادلة التفعيل Sigmoid function هي :

$$f(x) = 1/(1 + \exp(-c.x) \quad ....(5)$$

لقيم $c>0$

ويتم احتساب الاوزان بطريقة ما بحيث يكون الخطأ أقل ما يمكن و ان دالة الخطأ تتمثل بالعلاقة الاتية :

$$E = 1/2 * \sum_1 (z_1 - a_{i1})^2 \quad ....(6)$$

اذ إن : z_1 تمثل القيمة النهائية للدالة قيد التدريب.

a_1 تمثل القيمة الحقيقية (اخراج الشبكة)

ومن العلاقة 6 تمثل قيمة لـ E أحسن قيمة لدالة الخطأ للشبكة واذا كانت E=0 فان الشبكة تعمل بصورة أدق .

أما الأوزان فتتغير وفق المعادلتين الآتيتين:

$$\left.\begin{aligned} \Delta_{ij}^1 &= \alpha * \varepsilon_i * a_i (1 - a_i) * h_j \\ \Delta_{ij}^2 &= \alpha * \sum \varepsilon_m * a_m * (1 - a_m) w_{ij}^1 * h_j * (1 - h_i) * e_i \end{aligned}\right\} \quad ....(7)$$

وتعطى قيمة الخطأ بالعلاقة الاتية:

$$\varepsilon_i = z_i - a_1 = \text{Final Value} - \text{Network Value} = \text{error} \quad ....(8)$$

(kinnerbrock ,1995 ,40-41)

وعليه يمكن وصف خوارزمية التعليم للشبكة BP بالخطوات الاتية :

1. احتساب قيم الأوزان كافة و لأعداد عشوائية.
2. اختيار نموذج ادخال – اخراج عشوائي للدالة قيد التعلم، وحساب قيم h_i للطبقة المخفية.
3. لقيم الادخال e_i والقيم النهائية للشبكة z_i قيم تصحيح الاوزان وفقاً للمعادلة 7.
4. العودة الى الخطوة 2.

الجانب العملي

2-1 عينة الدراسة

قبل البدء في عرض النتائج المستحصلة من تطبيق المصنفات، لا بد من وصف البيانات التي استخدمت في البحث، فقد تم اختيار بذور النباتات (الحنطة، الشعير، الرز،...) لتنفيذ النموذج المقترح وذلك لما تحمله هذه البذور من خصائص وصفات مثل (نسبة الكربوهيدرات، الدهون ،البروتين، الألياف، ...الخ). وقد تم اعتماد نسبة وجود البروتين في البذور بوصفها صفة لدعم التصنيف فضلاً عن الصفات الأخرى، وتمييز أصناف بذور النباتات عن بعضها البعض والحصول على التصانيف الصحيحة.

والجدول الاتي يوضح أنواع الحبوب مصنفة وفقاً لنسب البروتين الموجود في كل منها.

الجدول 2

نسب البروتين الموجودة في بذور بعض النباتات

نوعية الحبوب	النسبة المئوية للبروتين
الرز الخام (الشلب)	10 – 9
الحنطة	11.5 – 10.5
حبة الشوفان الكاملة	11.8 – 11.6
الشعير	12.0 – 11.8
ذرة صفراء حلوة	12.30 – 12.10
ذرة بيضاء	13.00 – 12.40
الشيلم	13.5 – فما فوق

المصدر: الفخري، واحمد صالح خلف، 1983، 29.

2 – 2 الخوارزميات المستخدمة في البحث

1. المصنف C4.5

وهي إحدى أنواع المصنفات التي تعمل بمشرف، وهي من نوع شجرة القرار، اذ تستخدم هذه الخوارزمية مبدأ تقسيم من الأعلى إلى الأسفل (Top – Down) وذلك وصولا الى الحل الامثل، و من ميزاتها انها تتعامل مع الارقام، الخصائص، القيم المفقودة والبيانات المشوشة. فضلا عن وصفها بأنها من أفضل خوارزميات التصنيف واكثرها دقة و سرعة في الوصول الى الحل النهائي. (Llora and *et al*,2001, 4).

2. الشبكات العصبية BP

تعتمد فكرة الشبكات العصبية الاصطناعية على ايجاد منظومة حسابية لها القدرة على التكييف و التعديل عن طريق التعلم و ذلك سعياً لايجاد دوال الربط بين المدخلات و المخرجات، أو استنتاج قرار مبني على آلاف الاحتمالات و العلاقات التي تشكل بدورها ملفاً تاريخيا تبنى من خلاله العلاقة او الدالة (العبيدي ،2000، 75).

وتعد شبكة الانتشار الخلفي Back Propagation من أكثر الشبكات العصبية شيوعا و استخداماً، اذ يجري تعديل اوزان الشبكة و تحسين ادائها من خلال دالة التعليم للوصول الى افضل نتيجة او نتيجة مقاربة. عليه اعتمد البحث هذا النوع من الشبكات BP في الحصول على النتائج وقد تم شرح المعادلات المستخدمة في الجانب النظري من البحث.

3-2 مراحل تنفيذ البرنامج المصمم

بعد وصف الطرائق المستخدمة في التصنيف فقد أعتمد البحث لغة البرمجة V.B في كتابة برامجه المتمثلة بربط نتائج الشبكة العصبية BP مع المصنف C4.5 وذلك للحصول على النسبة النهائية للمعلومات التي يتم التصنيف من خلالها الى فئات وذلك باعتماد البيانات المشار اليها في الجدول 2.

الجدول 3

جدول يوضح نتائج تنفيذ معادلات المصنف C4.5 من برنامج الـ V. Basic

ت	عدد العناقيد	نسبة المعلومات المستحصلة من المعادلة E(APn)	نسبة المعلومات النهائية Gain	التصنيف
1	10	23.9624	11.1	حبة الشوفان الكاملة
2	10	22.5530	10.4	حنطة
3	10	24.0284	11.1	حبة الشوفان الكاملة

ت	عدد العناقيد	نسبة المعلومات المستحصلة من المعادلة E(APn)	نسبة المعلومات النهائية Gain	التصنيف
4	10	24.1379	11.01	حنطة
5	10	22.1974	10.1	رز الخام
6	8	22.9378	9.7	رز الخام
7	8	23.3018	10.7	حنطة
8	8	25.9889	10.3	رز الخام
9	8	24.2070	11.1	حنطة
10	6	23.9629	11.1	حنطة
11	6	23.2454	10.7	حنطة
12	6	25.1428	9	رز خام
13	6	23.6219	10	رز خام
14	6	24.0149	11	حبة الشوفان
15	6	23.3053	10.1	رز خام
16	4	22.0781	10.01	رز خام
17	4	23.3018	10.1	رز خام
18	4	21.6978	9	رز خام
19	4	24.0165	11	حبة الشوفان
20	3	21.5680	9	رز خام
21	3	22.0781	10	رز خام
22	3	23.5868	10	رز خام
23	8	22.9378	9.75	رز خام
24	8	22.4245	9.5	رز خام
25	8	22.2467	9.45	رز خام
26	8	22.8557	9.77	رز خام
27	8	22.9384	9.35	رز خام

يتبع

ماقبله

28	8	22.8822	9.59	رز خام
29	8	24.4865	11.20	حنطة
30	8	24.0371	10.97	حنطة
31	8	24.3657	11.14	حنطة
32	8	24.4103	11.16	حنطة
33	8	23.9241	10.90	حنطة
34	8	22.3122	10.02	رز خام
35	8	23.9650	11.11	شعير
36	8	26.0664	10.31	رز خام

37	8	23.0837	11.00	حنطة
38	4	24.486	11.2	حنطة
39	4	24.0371	10.9	حنطة
40	4	24.1379	11.08	حنطة
41	3	24.4402	11.18	حنطة
42	3	24.4266	11.18	حنطة
43	3	23.3567	10.80	حنطة
44	3	22.3726	10.24	رز خام
45	3	22.3726	10.25	رز خام
46	2	21.5124	9.58	رز خام
47	2	21.5582	9.61	رز خام
48	2	29.0202	12.72	ذرة بيضاء

الجدول 4

النتائج من برنامج الـ Minitab بعد تنفيذ المعادلات

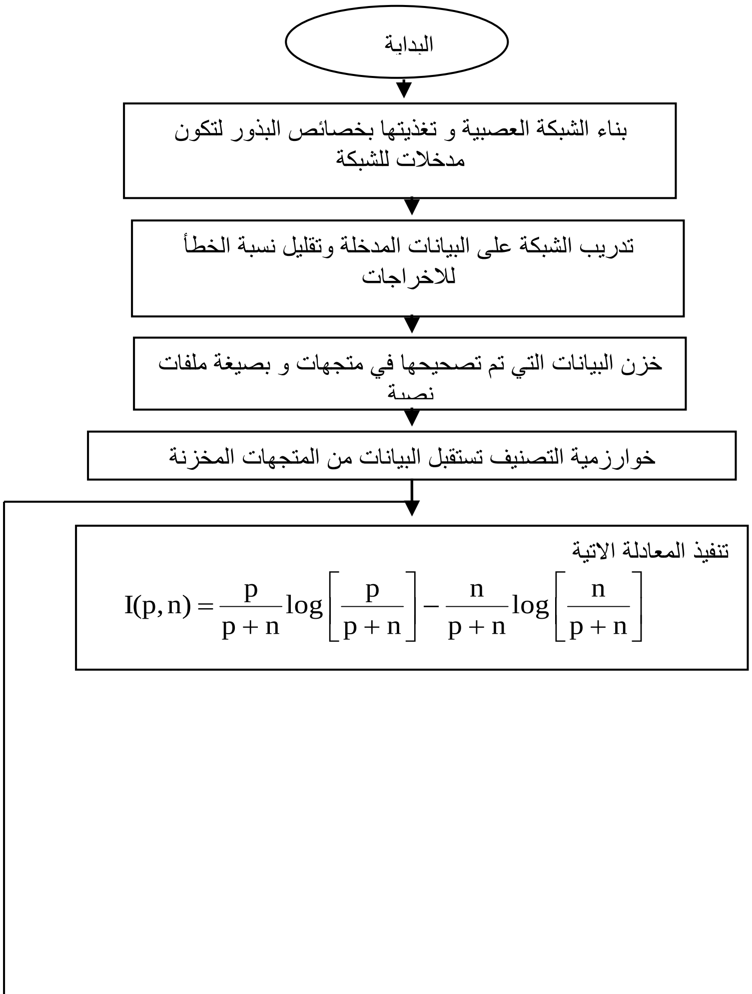
ت	عدد العناقيد	نسبة المعلومات النهائية	التصنيف
1	10	$9 \approx 8.9$	رز خام
2	10	9.2	رز خام
3	10	13.6	ذرة بيضاء
4	10	9.3	رز خام
5	10	10.8	حنطة
6	8	10.1	حنطة
7	8	10.5	حنطة
8	8	14.5	شيلم
9	8	9.5	رز خام
10	6	9.2	رز خام
11	6	10.5	حنطة
12	6	13.7	شيلم
13	6	9.2	رز خام
14	6	9.3	رز خام
15	6	11.6	شعير
16	4	8.8	ادنى من المستوى
17	4	10.2	رز خام
18	4	10.5	حنطة
19	4	10.6	حنطة
20	3	9	رز خام

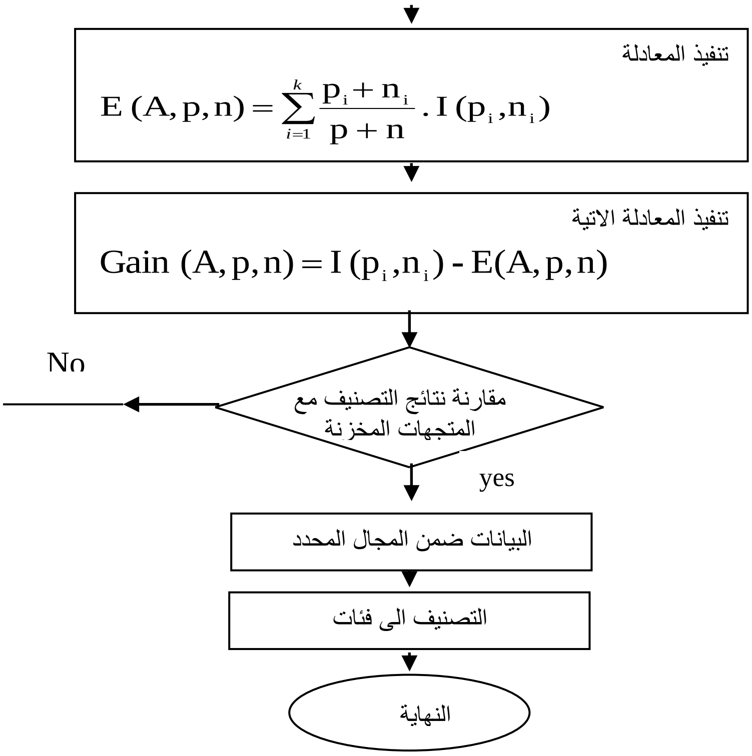
يتبع ⇦

ماقبله ⇦

21	3	7.1	ادنى من المستوى
22	3	6.6	ادنى من المستوى
23	8	9.7	رز خام
24	8	2.7	ادنى من المستوى
25	8	12.6	ذرة بيضاء
26	8	7.5	ادنى من المستوى
27	8	-7.7	ادنى من المستوى
28	8	39.4	اعلى من الحد الطبيعي
29	8	39.4	اعلى من الحد الطبيعي
30	8	42.5	اعلى من الحد الطبيعي
31	8	43.2	اعلى من الحد الطبيعي
32	8	36.1	اعلى من الحد الطبيعي
33	8	32.4	اعلى من الحد الطبيعي
34	8	42.7	اعلى من الحد الطبيعي
35	8	44.6	اعلى من الحد الطبيعي
36	8	38.8	اعلى من الحد الطبيعي
37	8	42.7	اعلى من الحد الطبيعي
38	8	44.6	اعلى من الحد الطبيعي
39	8	46.1	اعلى من الحد الطبيعي
40	4	41.2	اعلى من الحد الطبيعي
41	4	37.4	اعلى من الحد الطبيعي

ولفهم عمل البرنامج المصمم يمكن تتبع المخطط الانسيابي في الشكل 2، والذي يبين الخطوات كافة التي تمر بها الخوارزمية وصولاً الى النتائج.





الشكل 2

مخطط انسيابي يبين مراحل البرمجة المستخدمة في البحث

2-4

، 3، الاول

يمثل نتائج تنفيذ البرنامج المصمم والمكتوب بلغة V.B والمتضمن استخدام المصنف C4.5، في حين يظهر الجدول الثاني جدول 4 نتائج تنفيذ الحزمة البرمجية الجاهزة Minitab Release 13.0 .

من ملاحظة نتائج التصنيف في الجدول 3 تبين بأن النتائج المستحصلة من تطبيق البرنامج المعد من قبل الباحثين تقع جميعها ضمن فئات التصنيف الصحيحة، بعبارة اخرى فان نسبة المعلومات النهائية Gain والتي تمثل بالنتيجة نسبة البروتين الموجود في البذور قيد التصنيف، ولجميع الحالات التي تم تنفيذ البرنامج عليها تقع ضمن فئات الجدول 2 ولاتوجد أي حال شاذة في نتائج تنفيذ الخوارزمية C4.5 . وهذه النتيجة تشير الى توصل الخوارزمية C4.5 الى حالات التصنيف الصحيحة او المقاربة في جميع الحالات .

في حين تشير نتائج الجدول 4 الى وجود بعض من الحالات غير الصحيحة (حالات شاذة) اذ كانت النسب النهائية للبروتين دون المستوى المطلوب او انها اعلى بكثير من المستوى. هذا يعني ان استخدام الحزمة الجاهزة Minitab قد لايؤدي الى نتائج صحيحة في التصنيف، ناهيك عن استخدام بعض المعالجات (معالجة رياضية) على النتائج المستحصلة من تطبيق دالة التصنيف للحصول على النتائج النهائية وهذا

ما ينسجم مع فرضية البحث ويفرز من صحة النتائج المستحصلة من تطبيق البرنامج المصمم بلغة البرمجة V.B.

2-4 الاستنتاجات

تم التوصل من خلال نتائج تنفيذ البرنامج المصمم والنتائج المستحصلة من تطبيق الحزمة البرمجية الجاهزة Minitab الى جملة من الاستنتاجات ندرجها بالاتي:

1. ان المعادلات الخاصة بالمصنف C4.5 كانت اكفأ في الاداء وهذا ما أظهرته نتائج تنفيذ البرنامج بلغة V.B .
2. ان ربط نتائج الشبكة العصبية BP بخوارزمية التصنيف ادى الى ازالة التناقض والتشويش الموجودة في البيانات، اذ تعمل الشبكة العصبية ومن خلال مصفوفات الاوزان الى ازالة هذه الحالات من البيانات .
3. من ملاحظة المعادلات المستخدمة في المصنف C4.5، فان الحدود التي تحتوي على دالة اللوغاريتم قد تؤدي الى عدم الوصول الى حل نهائي وذلك في الحالات التي يكون فيها دليل الدالة Log سالبة، وقد تمت معالجة هذه الحالات الجاهزة نتيجة لمثل هذه الحالات .
4. لاتعد النتائج المستحصلة من تطبيق الحزمة الجاهزة Minitab نتائجاً نهائية، اذ تتطلب عملية التصنيف القيام بالعديد من العمليات الحسابية للحصول على النتيجة النهائية .
5. ان كانت الحالات التي تم تنفيذ البرنامج المصمم عليها حصلت على تصنيف، وذلك من خلال الحصول على نسبة المعلومات النهائية ضمن حدود الفئات الواردة في الجدول 2 في حين ظهرت العديد من الحالات غير المصنفة لدى استخدام الحزمة البرمجية الجاهزة Minitab.

المراجع

اولاً- المراجع باللغة العربية

1. محمود خليل ابراهيم العبيدي، "الشبكات العصبية الاصطناعية " مجلة ابحاث الحاسوب، المجلد الرابع، العدد الاول، 2000 .
2. عبدالله قاسم الفخري، و احمد صالح خلف، "بذور المحاصيل انتاجها ونوعيتها"، دار الكتب للطباعة والنشر، 1983.
3. نعمة عبدالله الفخري، "استخلاص نموذج بياني من قاعدة بيانات باستخدام خوارزميتي BK , K-means "، رسالة ماجستير، كلية علوم الحاسبات والرياضيات / جامعة الموصل، 2003 .
4. محاضرات مأخوذة من شبكة المعلومات الدولية (الانترنيت) :

1- Data Mining Glossary Courses .
www.towcrows.com/glossary.html/2004.
2- Data Mining : What is data mining .
www.Anderson.Ucla.edu/faculty/jason.f.Fraud/teacher/technologies/2004.

ثانياً- المراجع باللغة الاجنبية

1. Ali, S., and Abraham A. (2002), "An Empirical comparison of kernel selection for support vestor machine", Gippsland school of computing and information technology, Monash university, Victoria-Australia.
2. G`omez Hidalgo J. M., Mana Lop`ez M., and Puertas Sanz E. (2002), " Evaluating cost-sensitive un solicited Bulk Email categorization ", Journees internationales d'Analyse statistique des Donnees Textuelle (JADT) - http:// citeseer.nj.nec.com
3. Ibrahim R. S. (1999), "Data Mining of machine learning performance data" (M.Sc. thesis), RMIT university-Australia. http://goanna.cs.rmit.edu.au/~vc/papers/ibrahim-mbc.
4. Kinnebrock W.,(1995),"Neural Networks: Fundamentals, Applications, Examples", Galgotia publications Pvt. LTD, NewDelhi .
5. Keller F. (2000), "Introduction to machine learning", Journal of Machine Learning Research (JMLR). http://www.aai.org/AI Topics/html/mahine.html.
6. Llora X., and Garrell J. M. (2001), "Knowledge-Independent data mining with fine-grained parallel evolutionary algorithms", Enginyeria I Arauitectura La Salle, Uinveristat Ramon Llull. http://gal4.ge.uivc.edu/~xllora/curriculum/cu

ABSTRACT

The Use of The Classification Algorithms C4.5 in Distingushing

Data mining is an action to obtain the knowledge to achieve the main purpose that is detecting hidden facts which contain database by using various techniques that include artificial intelligent, statistic analysis, techniques and modeling.

Data mining method brings models and obvious relations in data which help to expect future results. Algorithm appears in this field which indicates comparisons between algorithm to choose the suitable one to have the best results.

This paper aims at using classification algorithms. C4.5 and connect it with neural nets (Back propagation BP) to form a classification model hold a two methods characteristics. In addition to comparing the total results with the classification results by using Minitab.

This paper results in classified equation C4.5 which is sufficient in performance especially after connecting it with the neural net BP to erase the contradictions in data.

The results also confirm the importance to use programming language in achieving the best in comparison with the results of ready made applications.