استخدام الخوارزميات الجينية المتوازية للكشف عن البيانات المتطرفة حسب مقياس BIC

شهلة عبدالوهاب عبدالقادر
قسم أنظمة الحاسوب ، المعهد التقني ، الموصل ، العراق

(تاريخ الاستلام: 8 / 6 /2011 ---- تاريخ القبول: 2 / 4 /2012)

الملخص

الخوارزمية الجينية هي إحدى تقنيات البحث الذكي التي تستخدم لإيجاد حلول تقريبية للمسائل التي تتطلب حلولاً أمثلية.

الهدف من هذا البحث استخدام الخوارزميات الجينية المتوازية بتقنية المدير/العامل (Manager/Worker) في حاسبات متوازية وبدالة معيارالمعلمات بايزي Bayesian Information Criterion(BIC') للكشف عن البيانات المتطرفة في نماذج الانحدار الخطي المتعدد .

تم استخدام بيانات الأرصاد الجوية من درجات الحرارة اليومية والإشعاعالشمسي وسرعة الر ياح واليوم من السنة لغرض تهيئتها لبناء نموذج تنبؤي لنسبة التبخر اليومي للأراضي المزروعة في مدينة الموصل للأعوام 1999-2001.أظهرت النتائج ان دالة معيار المعلمات بايزيا 'BIC كدالة لياقة في الخو ارزمية الجينية المتوازية ذات كفاءة عالية في الكشف عن البيانات المتطرفة وبأقل أخطاء ممكنة حيث وصلت نسبة تقليل الاخطاء مايقارب 98%. كما أظهرت الدراسة ان لعوامل الخوارزمية الجينية(الانتخاب، الطفرة، العبور ،الأجيال) تأثيرا واضحا في تقليل تلك الأخطاء، وان تقنية المدير /العامل في حاسبات متوازية التي تعمل بخوادم الحاسبات الموزعة ذات كفاءة في زيادة السرعة في هذا النوع من الخوارزميات الجينية الذكية . تم استخدام نموذج الانحدار الخطي المتعدد لتقييم النتائج واستخدم مقياس معامل التحديد المعدل Adjusted R Square في الحكم على النتائج ، حيث كانت قيمة معامل التحديد المعدل 0.732 قبل اكتشاف البيانات المتطرفة وازدادت الى 0.974 بعد اكتشاف البيانات المتطرفة

الكلمات المفتاحية: الخوارزمية الجينية المتوازية ، البيانات المتطرفة ، معيار المعلمات بايزي ،المدير/العامل.

1- المقدمة

الانحدار الخطي المتعدد هو واحد من الطرق الإحصائية الشائعة لتحليل البيانات وفهم الصفات الأساسية للبيانات. مع ذلك ، هناك صعوبات كثيرة تواجه تحليل البيانات منها اكتشافوا لغاء البيانات المتطرفة من البيانات الحقيقة . والبيانات المتطرفة هي بيانات خارجة عن نطاق البيانات الحقيقة لاي دراسة [1][4]. تحليل الانحدار الخطي المتعدد يواجه مشكلة البيانات المتطرفة التي تؤثر بشدة على نموذج الانحدار، وخاصة عند استخدام معيار تقديرالمربعات الأقل OLS(Ordinary Least Square). لذا عند تحليل البيانات وبناء النماذج يتم أولا معالجة البيانات للكشف عن البيانات المتطرفة ، وا زالتها ثم إعادة صياغة النماذج من البيانات الجديدة [2][7].لاحظ الباحثين ان البيانات المتطرفة تؤثر بصورة مباشرة على نتائج التحليل الإحصائي وتجعله اقل فائدة لذلك استخدمو العديد من الطرق الإحصائية التقليدية مثل hat , studentized residual, Cooks distance ,Dffits, diagonal elements في تعريف البيانات المتطرفة من خلال اكتشافها ومعالجتها [6][11][12][13]. لكن تلك الطرق الإحصائية لاتستطيع دائما ايجاد القيم المتطرفة . لذلك قام آخرون بحل مشكلة البيانات المتطرفة في نماذج الانحدار الخطي المتعدد واز التها باستخدام الخوارزمية الجينية [3][1][15] .

الخوارزمية الجينية هي إحدى تقنيات البحث الذكي التي تستخدم لإيجاد حلول تقريبية للمسائل التي تتطلب حلولاً أمثلية ، والتي يتطلب إيجادها وقتاً طويلاً وفق تقنيات البرمجة التقليدية تعتمد الخوارزميات الجينية على تقنيات مستوحاة من التطور الطبيعي للأجيال في الكائنات الحية Evolutionary Biology،حيث تستخدم تقنيات الوراثة Inheritance والتزاوج Crossover والطفرات Mutations والاختيار الطبيعي Natural Selection ، وبالتالي فهي شبيهة بعملية تطورا لأجيال واختيار أفضل الأفراد من كل جيل التي تحدث بشكل طبيعي عند الكائنات الحية ،حيث من المعلوم أن مبدأ(البقاء للأفضل)هو السائد طبيعياً وهو المبدأ الذي تعتمد عليه الخوارزمية الجينية في حل المسائل التي تتطلب حلاً أمثلياً [1] Ishibuchi وآخرون [10]اقترحو ا خوارزمية جينية للكشف عن البيانات المتطرفة لمجاميع بيانية في تصنيف الأنماط . باستخدام معيارالمعلمات بايزي (BIC) [11].

الهدف من هذه الدراسة سرعة الكشف عن البيانات المتطرفة في نماذج الانحدار الخطي المتعدد باستخدام الخوارزميات الجينية المتوازية بدالة معيارالمعلمات بايزي BIC في حاسبات متوازية باستخدام بيانات الأرصاد الجوية لأيام السنة من درجات الحرارة والإشعاع الشمسي وسرعة الرياح ونسبة التبخر في مدينة الموصل /العراق . ماتبقى من البحث هو كما يلي : القسم الثاني يتضمن تفاصيل عن الطريقة المستخدمة في البحث. القسم الثالث يشرح المواد ومنهجية البحث المقترحة . القسم الرابع النتائج ومناقشتها . القسم الخامس الاستنتاجات .

2- الخوارزمية الجينية المتوازية للكشف عن البيانات المتطرفة

لتوضيح اكتشاف البيانات المتطرفة في نماذج الانحدار الخطي المتعدد باستخدام الخوارزميات الجينية المتوازية وبدالة معلمات بايزيا في هذا القسم تكون كالاتي:-

ISSN:1813 – 1662

2-1 البيانات المتطرفة في نماذج الانحدار المتعدد :

ان نموذج الانحدار الخطي المتعدد يهتم بتقدير العلاقة بين متغير كمي وهو المتغير التابع والعديد من المتغيرات الكمية هي المتغيرات المستقلة [7] وحسب المعادلة (1):

$$Y= B_0 +B_1X_1 + B_2X_2+....+ B_kX_k+U.....(1)$$

حيث ان Y المتغير التابع ، X المتغيرات المستقلة للنموذج ، B0 حد ثابت ، B المعالم المطلوب تقديرها في النموذج، U الاخطاء العشوائية .وبما ان المعادلة رقم (1) هي العلاقة الحقيقة المجهولة والمراد تقديرها باستخدام الإحصاءات المتوفرة عن المتغير التابع والمتغيرات المستقلة فانه يستوجب تحقق الفروض الأساسية التالية الخاصة ب U وحسب المعادلة (2) :

$$U \sim N(0, \sigma^2 Ln) \quad .....(2)$$

والتي تعني ان U يتوزع توزيعا طبيعيا لمتجه وسطه الحسابي μ صفري (0) ومصفوفة تباين هي σ^2 Ln ، وممكن استخدام طريقة معيار تقدير المربعات الاقل OLS في تقدير معلمات النموذج الخطي المتعدد لذلك يمكن كتابة المعادلة (1) بصيغة المعادلة (3) :

$$\hat{Y}= B`_0 +B`_1X_1 + B`_2X_2+.....+ B`_kX_k+U..... \quad (3)$$

ولما كان هدفنا الحصول على قيم (B`) التي تجعل مجموع مربعات الانحرافات اقل ما يمكن ، يمكن تصغير القيمة مبدا (المربعات الصغرى) الى اقل قيمة ممكنة ، حسب المعادلة(4) :

$$Min \ \sum e_i^2 = \sum (Y - \hat{Y})^2 \ ...(4)$$

ان الضرر الأساسي لطريقة OLS على الأداء وجود أخطاء (e) لاترضي بالفرضيات التقليدية . وواحدة من اهم الأخطاء الشائعة في التوزيع الطبيعي للبيانات هوا احتوائها على واحدة اواكثر من البيانات المتطرفة[7][8].

فرضت البيانات المتطرفة على شكل متغيرات زائفة Dummy variables . المتغيرات الزائفة هي جزء من المتجه (N*1) حيث ان N عدد بيانات نموذج الاختبار الكلية للمعادلة (1). البيانات الكلية تم تشفيرها باستخدام (1, 0) ، حيث تكون قيمتها (1) في حالة البيانات الزائفة ، وقيمتها (0) في حالة البيانات الحقيقة غير الزائفة، والمتغير الزائف يكافئ اكتشاف القيمة المتطرفة.والطريقة المقترحة من قبل Schwarz ،[14]و التي تعرف بمعيار المعلمات بايزيا (BIC) تستخدم لاكتشاف البيانات المتطرفة في نماذج الانحدار الخطي المتعدد والتي يمكن حسابها من المعادلة (5) :

$$BIC= Log(\sigma^2) + m \log(N) / N \ ...(5)$$

حيث ان $\sigma^2 = (e` e)/(N-K-1)$ ،التباين في النموذج المخمن $m=1+k+m_d$، تمثل مجموع المعلمات في النموذج المخمن، المكونة من عدد ثابت، k عدد المتغيرات المستقلة ، m_d عدد المتغيرات الزائفة المتطرفة (outlier dummy).عموما النموذج الجيد يمتلك أخطاء قليلة بأقل معلمات،وبأفضل قيمة صغرى للدالة (BIC) [2]. إن مشكلة استخدام BIC لاكتشاف البيانات المتطرفة تميل في بعض الأحيان إلى تضمين بيانات زائفة متطرفة غير ضرورية ولحل هذه المشكلة يمكن اخذ معيار إضافي في هذه الطريقة باستخدام kappa (κ>1) للبيانات الزائفة.و حسب المعادلة (6):

$$BIC`=Log(\sigma^2)+(1+ k) \log(N) / N+\kappa m_d \log(N) / N \ .....(6)$$

2-2 الخوارزميات الجينية لاكتشاف البيانات المتطرفة

الخوارزمية الجينية هي تقنيات بحث تستخدم لإيجاد الحلول الحقيقية والتقريبية في مشاكل البحث. حيث يتم فيها توصيف المشكلة لتمثيل الصبغيات Chromosomes الممثلة للحلول بواحدة من طرق الترميز. بعد ذلك تطبق مجموعة من العمليات الرياضية المستنبطة من العمليات البيولوجية كالعبور، التوارث ، الاختيار، التزاوج والطفرة للحصول في نهاية المطاف على مجموعة من الصبغيات التي تمثل الجيل النهائي وكل صبغي هو فرد من افراد الجيل، وأفضل صبغي هو الحل الأمثل الذي تبحث عنه المشكلة المطروحة. اذا ، تبدا عملية البحث انطلاقا من مجموعة حلول وليس حل واحد[9][1]. الحلول مثلت بصيغة ثنائية (1,0) لكل الاحتمالات الممكنة للمسألة، والجيل الأول يبدأ بتوليد جيل عشوائي لكل الصبغيات لتمثيل مساحة الحل ، اما الجيل الجديد يتكون من تقييم كل فرد في الجيل القديم حسب دالة اللياقة استخدمت هذه الطريقة في اكتشاف البيانات المتطرفة في نموذج الانحدار الخطي المتعدد ، والخوارزمية الجينية مكونة من الأتي:

- **تشفير المعلمات Parameter Encoding**: تشفير المعلمات (الصبغيات)على شكل متجه ثنائي (d)، ($d= ....,d_N$) ، حيث ان $d_i=0$ تشير الى ان البيان غير متطرف، (d_1 , d_2 ,) $i=1,.....,N$ تشيرالى ان البيان متطرف، لكل.

وهيكلية الصبغيات المتطرفة في الجيل تكون مكونة من الصبغيات P، حيث ان Pعدد البيانات المتطرفة (N,....,P=1 ,O_p) ، على سبيل المثال : عدد البيانات المتطرفة p=3: الاول ، الثاني :و N-1 كما في الشكل (1)الذي يوضح هيكلية الصبغيات في الخوارزمية الجينية :

	d_1	d_2	d_3	...	d_{N-1}	d_N
d=	1	1	0	...	1	0
	o_1	o_2		...	o_p	

الشكل رقم (1): هيكلية الصبغيات في الخوارزمية الجينية

- **حجم الجيل Population size** : يقصد به عدد الأفراد ضمن الجيل وهو من العوامل المهمة والتي يتوقف عليها أداء الخوارزمية .

ISSN:1813 – 1662

لذلك فإن زيادة حجم الجيل سوف يقلل من الخطأالمرتقب في النتيجة . أيضاً هناك مشكلة عند كون حجم الجيلصغيراً إذ تنخفض دقة النتائج بشكل ملحوظ، لأنه لن يشمل كامل المجال المدروس بشكل جيد.

• **دالة اللياقة Fitness Function**: سلسلة الصبغيات المكونة للجيلالقديم من البيانات المتطرفة وغير المتطرفة تتغيرالى الجيل الجديد اعتمادا على دالة اللياقة 'BIC في المعادلة (6) .

• **عدد تكرار الأجيال Generation**: كل تكرار للخوارزمية يعني الحصول على جيل جديد من الأبناءووفقاً لمبدأ داروين في الوراثة الذي ينص على أن البقاء للأنسب وأن جيل الأبناءعموماً يكون أكثرتكيفاً مع الظروف البيئية الجديدة بسبب فناء الأفراد الأقل تكيفاتأقلماً، فكلما زاد عدد الأجيال فإن الحلول الأقرب للحل الأمثل سوف تحافظ على نفسها على أقل تقدير في الأجيال اللاحقة.

• **الاختيار وفق مبدا العجلة المتدحرجة Roulette Wheel Selection**:لمحاكاة هذه الطريقة نعتبر وجود عجلة تم تقسيمها إلى 100قطاعيوزع أفراد الجيل على هذه القطاعاتوفقاً لمتوسط احتمالية الانتخاب لكل فرد من أفراد الجيل الحالي والذي يعطى رياضيا بالمعادلة(7) :

$$P_{select}(i) = \frac{F_i}{\sum_{i=1}^{N} F_i} \quad .....7$$

حيث :

Pselect(i) : احتمالية انتخاب الفرد i

Fi : درجة ملاءمة الفرد i

N : عدد أفراد الجيل.

تتم طريقة الانتخاب بإدارة العجلة بشكل عشوائيوانتظار وقوف العجلة عند مؤشرما،

ويؤخذ الفرد الذي أشار إليه المؤشر.فكلما زاد عدد قطاعاته زادت درجة لياقته ومن ثم زاد احتمال انتخاب هذا الفرد وهذا يعني دخوله في الجيل الخلفاو تأثيره الإيجابي في هذا الجيل على أقل تقدير.[1]

• **العبور Crossover**:عملية العبورمن العمليات المهمة في الخوارزميات الجينية فالمعتقد السائد هو أن التزاوج بين أفراد يتمتعون بمواصفات جيدة سوف ينتج عنهغالباً أفراد يتمتعون بمواصفات جيدة على أقل تقدير. ومنثَمَّ يتم انتخاب الأفراد ذوي مقياس اللياقة الجيد بعد ذلك تنفذ عملية مزج عشوائي Shuffling للتخلص قدر الإمكان من إجراء العبور بين الفرد ونفسه والناتج عن احتمال تكرار الفرد أكثر من مرة بشكلمتتالٍ نتيجة عملية الانتخاب التي تسبق العبور.أيضاً تتم عملية العبور باحتمال العبور لكل فرد ضمن الجيل ،بعد عملية المزج العشوائي لأفراد الجيل تتم عملية العبوربين الفرد والذي يليه باستخدام إحدىطرائق العبور كالعبورالمنسق او نقطة البسيط المستخدمة في هذه الدراسة) (Simple n-point Crossover وحسب الشكل (2) الذي يوضح كيفية تطبيق العبورمن نوع نقطة البسيط –2 على صبغتيين متتاليتين من جيل الآباء وفق القيم العشوائية لقيمة العبور= 2، وقيمة العبور =6 .

قيمة نقطة العبور		1	2	3	4	5	6	7	8
جيل الآباء	Chromosome 1	1	1	0	1	1	1	1	0
	Chromosome 2	0	0	0	0	1	0	0	1
		↓	↓↑	↓↑	↓↑	↓↑	↓↑	↓	↓
جيل الأبناء	Chromosome 1	1	0	0	0	1	0	1	0
	Chromosome 2	0	1	0	1	1	1	0	1

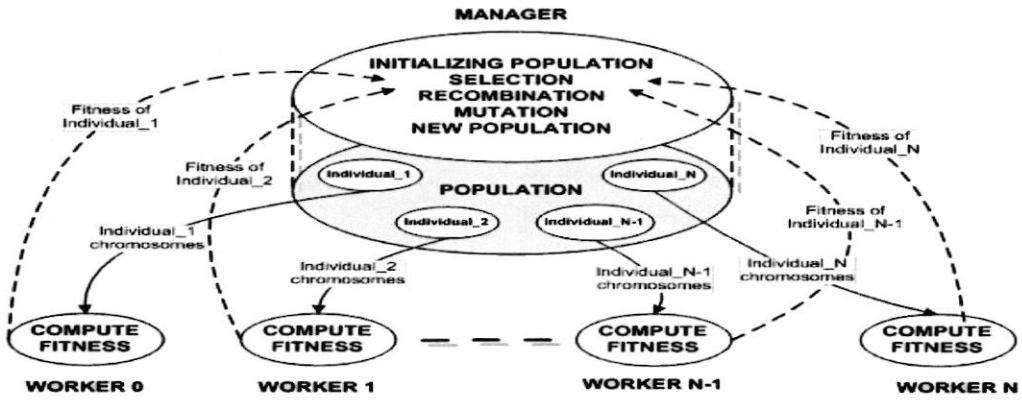
الشكل رقم (2): عبور نقطة البسيط –2

• **الطفرة Mutation** : هي المرحلة الأخيرة من سلسلة العمليات التكرارية التي تسهم بشكل جيد في الوصول إلى الحل الأمثل بسرعة وهي مستمدة من كون الخوارزميات الجينيةقريباً من تقنيات البحث العشوائي عن الحل الأمثل ومنثَمَّ فإن حدوث تغير مفاجئ(عشوائي) وغير متوقع في الجيل ولو كان باحتمال حدوث طفيف سوف يكون له الأثرالإيجابي في الاقتراب من الحل الأمثل، فعند استخدام طريقة الترميز الثنائي تكون الطفرة عبارةعن عكس إحدى المورثات ضمن الصبغي أي من الصفر إلى الواحد أو العكس، واحتمال حدوث هذا التغير يعرف باحتمال الطفرة (Pm probability of Mutation) حيث يتم اختياره بقيمة متدنيةنسبياً. [15]

2-3 الخوارزمية الجينية المتوازية

الخوارزمية الجينية تتطلب وقت طويل لتنفيذها لكن في تقنيات الحاسوب الحديث تم التغلب على طول الوقت بتنفيذ هذا النوع من الخوارزميات في حاسبات متوازية الهدف منه الحصول على الحل الأمثل بسرعة عالية. الخوارزمية الجينية المتوازية طبقت على أساس المدير/ العامل Manager/Worker بعدة حاسبات، حاسبة رئيسية تعتبر المدير، و N من الحاسبات الطرفية كعمال، الموضحة بالشكل (3). المدير يعمل على تأسيس الاتصال مع العمال من خلال إرسال إشارة الاتصال، ثم يقوم بتوليد صبغيات الجيل الابتدائي عشوائيا وتوزيعهاعلى العمال بالتساوي ، تهيئة متطلبات الخوارزمية الجينية وتوزيعها على العمال Worker-N.أيضا يقوم المدير باستلام النتائج من كل عامل لإعادة تمثيلها عشوائيا ليتسنى تكرار تلك العمليات،

إرسال إشارة لإنهاء الاتصال مع كل العمال بعد الحصول على الحل الأمثل النهائي. أما العمال N ، يقوم كل عامل من البداية باستلام الإشارة من المدير لتأسيس الاتصال والانتظار لحين البدء بإرسال متطلبات الخوارزمية الجينية ،وجزء من الصبغيات الكلية لتنفيذ الخوارزمية بدالة اللياقة 'BIC ، ومن ثم إرسال ناتج الخوارزمية الجزئية الى المدير،العمليات تكرر لكل عامل مع كل إرسال من المدير إلى أن يستلم إشارة بإنهاء الاتصال بعد الوصول إلى الحل الأمثل النهائي [5][3]. والنظام المقترح للخوارزمية الجينية المتوازية (PGA)يوضح في الشكل (4).



الشكل (3) الخوارزمية الجينية المتوازية

3- مواد البحث وطرائقه :

وفقا لما تقدم في المقطع الثاني ، تم استخدام بيانات الأرصاد الجوية من درجات الحرارة اليومية ،الإشعاع الشمسي ،سرعة الرياح ،اليوم من السنةونسبة التبخر اليومي للأراضي المزروعة في مدينة الموصل للأعوام 1999-2001للكشف عن البيانات المتطرفة في نموذج الانحدار الخطي المتعدد. تم استخدام بيانات الجدول رقم (1) في تمثيل نموذج الخوارزمية الجينية المتوازية .

كما استخدمت **MATLAB Distributed Computing Server (MDCS)** المعروفة بخوادم الحاسبات الموزعة والموجودة في **MatlabR2010b** وتأسيسها على مجموعة حاسبات. واستخدام تقنية المدير/العامل Manager/Worker ، حيث ربطت 5 حاسبات ، (المدير الذي يعمل على إدارة العمل كحاسبة Server خادم ،العمال Worker-N الذين يعملون كحاسبات طرفية وهنا عدد الحاسبات Worker-N=4.وتم تاسيس الاتصال بين المدير والعمال مع تنفيذ الخوارزمية الجينية فيها حسب الخطوات التالية:

1- Install the **MDCE** Service on all computers
2- In **Dos** for the Folder Matlab execute instructions:
   Matlabroot2010b\distcomp\bin**Mdce install**
   Matlabroot2010b\distcomp\bin**Mdce start**
   Matlabroot2010\distcomp\bin**admincenter.bat**
3- In admincenter: Create Job Manager : Jm1,Worker (1 – 4)
4- In Matlab: matlab pool **open** local Jm1 فتح الاتصال

[Initialize GA] :

options = gaoptimset(@ga)----- استحداث قيد بكل متطلبات الخوارزمية الجينية

Option.UseParallel='Always'
Option.PopulationSize=1000 :
Option.Generations= 50
Option.CrossoverFraction=0.0.01
Initial Population : option.population= Rand (population)

Create M-file for (Fitness):

[BIC]=function Fitness(N , K, σ , md)
BIC=Log(σ^2)+(1+ k) log (N) / N+κ* md* log(N) / N
End

[Repeat Until Termination Condition is satisfied **Do]** :

spmd (Instruction of Single programming multiple data)
D = codistributed(population,option);
worker 1: This lab stores D(:,1:250).
worker 2: This lab stores D(:,251:500).
worker 3: This lab stores D(:,501:750).
worker 4: This lab stores D(:,751:1000).
FitnessResult= codistributed (Fitness);
[option] = ga(Fitness ,options)
End

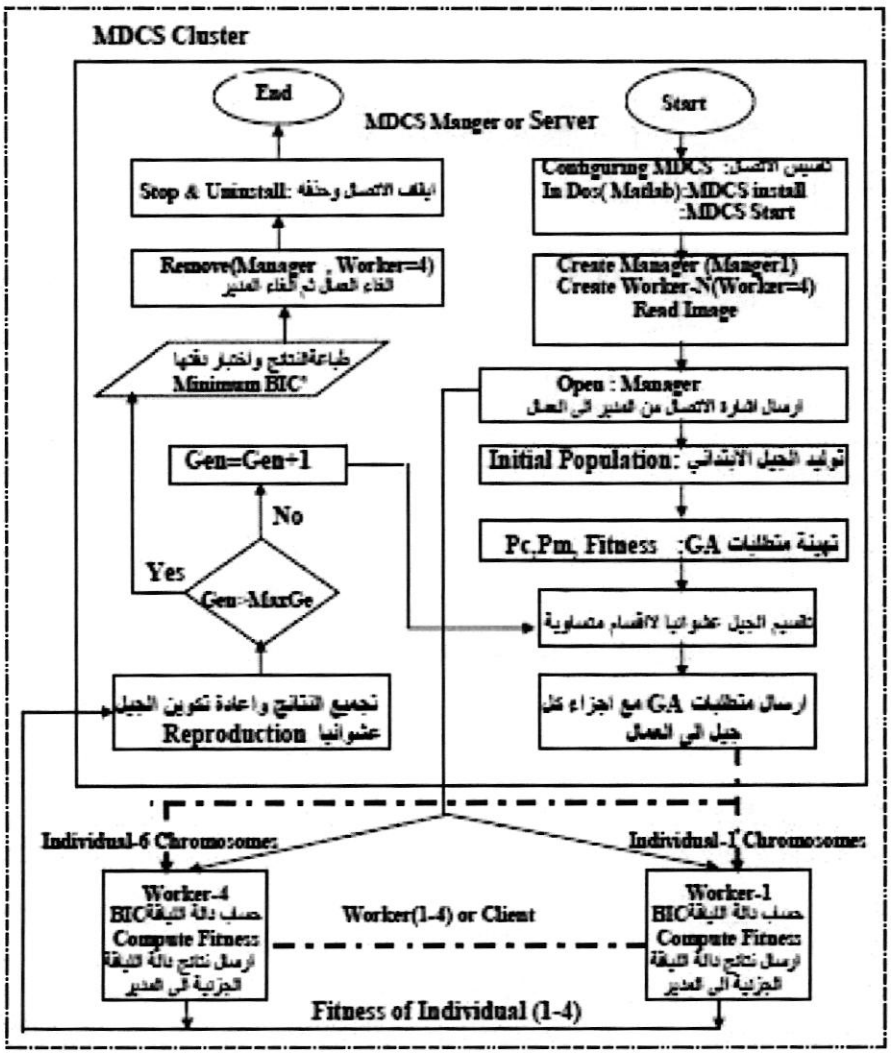
[Replace]Use new generation population and look for minimum of BIC'

[Test] If the final condition is satisfied on BIC' **Stop**, and return to best solution in current population.

[Loop] Go to **[Repeat Until]**

الجدول رقم (1) بيانات نموذج الخوارزمية الجينية المتوازية

حجم البيانات N= 1000	
50	عد الاجيال
1000	حجم افراد الجيل الكلي المولد في المدير Manager
250	حجم الجيل المستلم لكل عامل Worker
BIC'	دالة اللياقة
1	العبور
0,01	الطفرة



الشكل (4) مخطط النظام المقترح PGA للكشف عن البيانات المتطرفة حسب مقياس BIC'

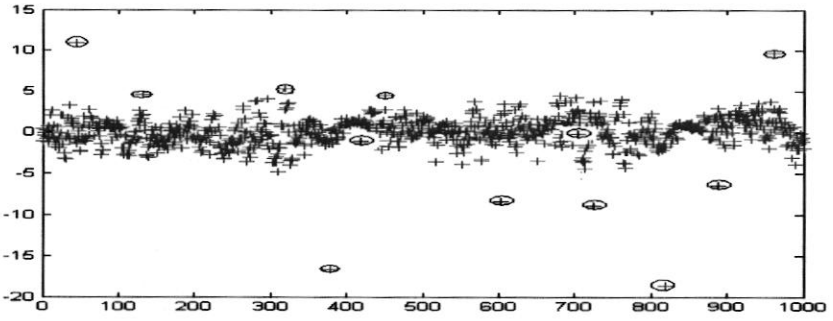
ISSN:1813 – 1662

4- مناقشة النتائج:

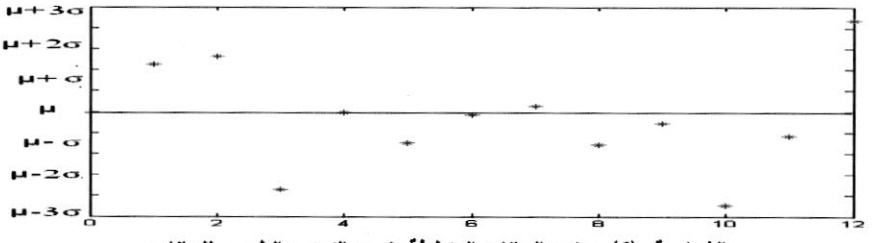
تشير النتائج التي تم الحصول عليها ان تطبيق الخوارزمية الجينية المتوازية لها تأثير في الكشف عن البيانات المتطرفة في نماذج الانحدار الخطي المتعدد. تبين ذلك من خلال استخدام بيانات الأرصاد الجوية من درجات الحرارة اليومية، الإشعاع الشمسي، وسرعة الرياح لاي يوم من السنة كمتغيرات مستقلة(X)، ونسبة التبخر اليومي للأراضي المزروعة كمتغير تابع (Y)لثلاث سنوات بما يعادل 1000عينة ، وقيمة K=4 (عدد المتغيرات المستقلة).

فعند تحليل البيانات احصائيا باستخدام نموذج الانحدار الخطي المتعدد تبين ان قيمة معامل التحديد المعدل تصل الى 0.732 وهي قيمة مقبولة احصائيا والتي جاءت من العلاقة الخطية بين المتغيرات المستقلة والمتغير التابع. ازدادت قيمة معامل التحديد المعدل بعد تطبيق الخوارزمية الجينية المتوازيةو استخراج القيم المتطرفة من البيانات الأصلية حيث كانت مساوية لـ 0.974 .

يلاحظ من زيادة قيمة معامل التحديد ان البيانات المستخدمة في تحليل الانحدار المتعدد حصلت على تصحيح بنسبة عالية جدا من خلال الكشف عن البيانات المتطرفة باستخدام الخوارزمية الجينة المتوازية. فبعد إجراء التجارب على البيانات المستخدمةو بواقع 50 عملية تكرار كإرسال من المدير الى العمال وقيمة Kappa (κ)=4 .تم الكشف عن و جود بيانات متطرفة بقيمة معيارية 'BIC مساوية (-6.6037) وبواقع (12) عينة في المواقع (937,801,714,646,567,547,395,378,346,272,126,75) من حجم البيانات الكلية يوضحها الشكل (5)و الموضحة بدائرة ضمن البيانات الكليةو التي كانت ضمن التوزيع الطبيعي للبيانات والمحصورة بين (μ + 3σ) و(μ - 3σ) يوضحها الشكل (6). تم إلغاء البيانات المتطرفة من البيانات الأصلية ليصبح حجم العينة الجديد 988 عينة وعند إعادة التجارب وبنفس التكرار عليها لم يكتشف أي حالة أخرى من البيانات المتطرفة.



الشكل رقم (5) يوضح البيانات المتطرفة في البيانات الكلية

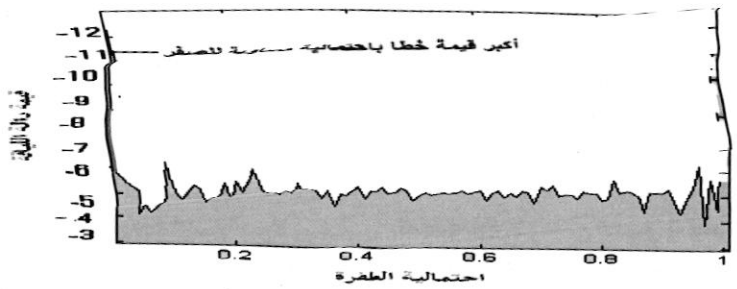


الشكل رقم (6) يوضح البيانات المتطرفة ضمن التوزيع الطبيعي للبيانات

أما من ناحية قيمة Kappa (κ) تبين أن لها تأثير كبير على تصحيح البيانات من خلال التباين في اكتشاف البيانات المتطرفة فعندما كانت قيمتها = 1 أو 2 كانت هناك كشف عن حالات قليلة من البيانات المتطرفة تراوحت بين (2-5) ، لكن مع زيادة قيمة Kappa (κ)=3 كان هناك زيادة في اكتشاف البيانات المتطرفة وصلت إلى (7) حالات . لكن عندما كانت قيمة Kappa (κ)=4 كانت هناك زيادة ملحوظة في الكشف عن البيانات المتطرفة وصلت إلى (12) حالة وبأقل قيمة لدالة اللياقة 'BIC والتي يوضحها الشكل (6). من جهة أخرى، احتمال الطفرة كان له تأثير ايجابي في سرعة الوصول إلى الحل وبخطوة صغيرة كانت احتمالية الطفرة مساوية

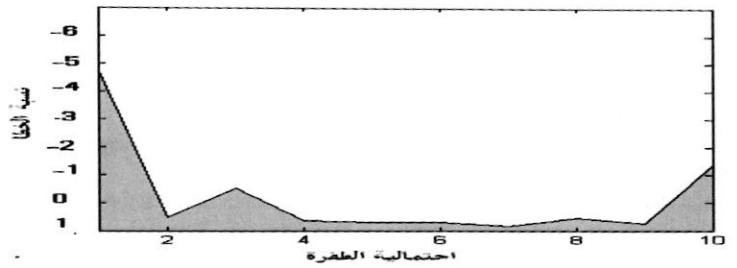
[Pm = [0 : 0.01 : 1 التي يوضحها الشكل رقم (7). حيث يمثل المحور الشاقولي قيمة دالة اللياقة 'BIC ضمن مجال الحل لها في حين يمثل المحور الأفقي احتمال حدوث الطفرة، كانت النتيجة تتأرجح حول قيمة (-6.6037)التي تم الحصول عليها أي أن أعظم خطا لا يتجاوز 0.06 وهذه نتيجة جيدة .ونتيجة التكرار المستمر للبرنامج تم الحصول على نتائج متشابهة بشكل كبير وخاصة فيما يتعلق بكون الخطأ ضئيل عند كون احتمال الطفرة ضمن نفس المجالو أن أكبر خطأ هو عندما كانت الطفرة باحتمالية مساوية للصفر

ISSN:1813 – 1662



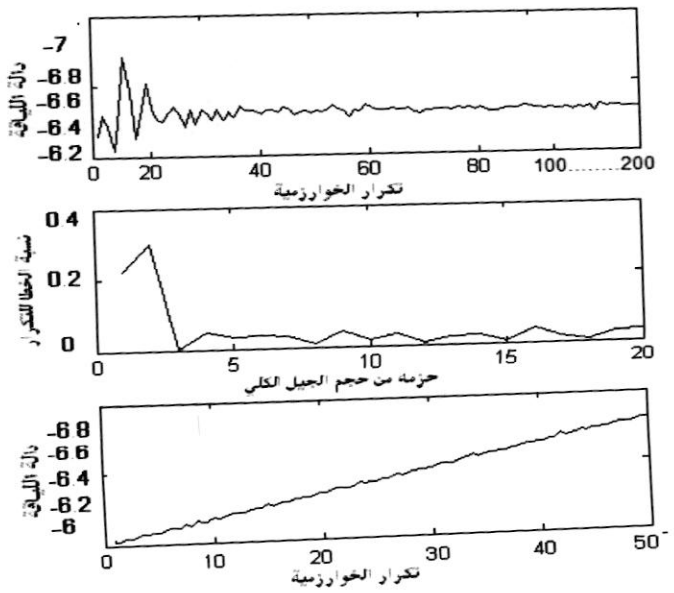
الشكل رقم (7) تأثير احتمال الطفرة في نتيجة الحل

ولغرض بيان تأثير احتمالية الطفرة تم تقسيم قيم احتمالية الطفرة إلى عشرة مجالات متساوية ومتدرجةوا ,يجاد المتوسط الحسابي للنتائج التي تم الحصول عليها، مع إيجاد الخطأ النسبي لكل مجال، لوحظ ما يؤكد أن الاستغناء عن الطفرة يسيء للنتائج التي تم الحصول عليها بزيادة الخطأ وهذا ما يوضحه الشكل رقم (8).



الشكل (8) الخطأ النسبي مع تغير مجال احتمال الطفرة

اما عدد الأجيال في الخوارزمية الجينية المتوازية كان له تأثير في سرعة الوصول الى الحل الامثل ، حيث لوحظ أن الحلول الأقرب للحل الأمثل سوف تحافظ على نفسها بأقل تقدير في الأجيال اللاحقة فكانت النتيجة تتأرجح حول قيمة(6.6037-)بعدد تكراريتراوح بين(100-200) جيل أي ان الخطا لايتجاوز 0.04 وهذا مايوضحه الجزء الاول من الشكل (9) . كما لوحظ ان زيادة عدد مرات التكرار يقلل من الخطا خاصة بعد زيادته عن حد معين. وان عدد مرات التكرار عندما يكون صغيرا يزيد من الخطا بشكل ملحوظ وهذا مايوضحه الجزء الثاني من الشكل (9) حيث تم تقسيم حجم الجيل الى عشرين مجالا، لذلك كان أفضل وأسرع تكرار للخوارزمية الجينية المتوازية عندا 50 جيل والموضح في الجزء الثالث من الشكل (9) .



الشكل رقم(9) تأثير عدد مرات تكرار الخوارزمية في نتيجة الحل

ISSN:1813 – 1662

4- الاستنتاجات والتوصيات

الخوارزميات الجينية المتوازية أثبتت نجاحها للوصول إلى الحل الأمثل بسرعة وكفاءة عاليتين بتقنية المدير/العامل وبدالة معيارالمعلمات بايزي (BIC) للكشف عن البيانات المتطرفة في نماذج الانحدار الخطي المتعدد. أعطت النتائج قيمة مثلى ل 'BIC مساوية (6.6037-) في الكشف عن كافة البيانات المتطرفة في البيانات الأصلية بحيث وصلت نسبة تقليل الاخطاء الى 98% . تطبيق الانحدار الخطي المتعدد على البيانات قبل وبعد اكتشاف البيانات المتطرفة كان له دورفي قياس كفاءة هذه الطريقة حيث وصلت قيمة معامل التحديد المعدل الى 0.732 قبل اكتشاف البيانات المتطرفة في حين وصلت قيمة المعامل الى0.974 بعد اكتشاف كافة البيانات المتطرفة.اوصي باستخدام تلك البيانات لبناء نموذج تنبؤي باستخدام الشبكات العصبية الاصطناعية .استخدام الخوارزميات الجينية المتوازية في الكشف عن البيانات المتطرفة قبل بناء نماذج الانحدار المتعدد.استخدام التقنية المتوازية في ايجاد الحلول لاي مشكلة اذا كان عامل السرعة مهم لحلها.

6- المصادر

[1] Ozlem , A. Serdar, K., "Genetic Algorithms Based Outlier Detection Using Bayesian Information Criterion in Multiple Regression Model Having Multicollinearity Problem ",G.U. Journal of science, 2009.

[2] Amidan, B., Ferryman, T., Cooley, S., "Data Outlier Detection Using the Chebyshew Theorem, In: Aerospace", IEEE Aerospace Conference Proceedings, IEEE, Piscataway NJ USA, 3814-3819 (2005).

[3] P. Borovska ,"Solving the Travelling Salesman Problem in Parallel by Genetic Algorithm on Multicomputer Cluster", Computer Systems Department,Technical University of Sofia, 2006.

[4] Salena,A.,Mozammel,K.,"Multiple –Case Outlier Detection in Multiple Linear Regression Model Using Quantum-Inspired Evolutionary Algorithm", J. computer ,Vol. 5,No. 12, Dec. 2010.

[5] S.Tongchim,Prabhas ," Parallel Genetic Algorithm for Finite-State Machine Synthesis From Input/Output Sequences", Department of Computer Engineering,Chulalongkorn University Bangkok 10330, Thailand ,email g41stc@cp.eng.chula.ac.th, 2000

[6] Ozlem, A. Serdar, K, "Genetic Algorithms For Outlier Detection in Multiple Regression with different information criteria ", Journal of Statistical Computation and simulation , Vol. 81 ,Issue ,1, 2011.

[7] Matthew,S.S.,Elizabeth,H.S.,"Bayesian Outlier Detection with Dirichlet Process Mixtures", 6,No. 4,PP. 665-690,2012

[8] J. Tolvi , "Genetic Algorithms For Outlier Detection and Variable Selection in Linear Regression Models ",S.Computing, Vol. 8, No. 8, 527-533 , 7, Oct. 2003.

[9] Abdoun,O.,Abouchabaka,J."A Comparative Study of Adaptive Crossover Operators for Genetic Algorithm to Resolve the Traveling Salesman Problem", International Journal of Computer Applications, Vol. 31,No. 11, Oct. 2011.

[10] Ishibuchi, H., Nakashima, T., Nii, M., "Genetic Algorithm Based Instance and Feature Selection",In: Liu H, Motoda H, Instance Selection and Construction for Data Mining, Kluwer Academic, (2001).

[11] Jo-Anne, T., Aaron, D., "Automatic Outlier Detection :A Bayesian Approach", IEEE International Conference (ICRA), Apr.2007.

[12] Hancong, L., Sirish, S.,Weu, J.," On-Line Outlier detection and Data Cleaning", Comp. & Chemical Engineering ,1635-1647, 20Jan. 2004

[13] Dragoljub,P.,Aleksandar, L.,"Incremental Local outlier Detection for Data Streams",IEEE Symposium on computational Intelligence and Data Mining (CIDM),Ap.2007.

[14] Schwarz,G.," Estimating the Dimension of a model", The annals of Statistics, 461-464, 1978.

[15] الحجار، م .عبيد، م" . (2003) .الخوارزميات الجينية ."إشراف العرفي، هادي.مشروع دبلوم، جامعة دمشق، كلية الهندسة الميكانيكية و الكهربائية .

Using Parallel Genetic Algorithm for outlier detection Based on BIC

Shahlla A. AbdAlKader
Dept. of Computer Systems, Technical Institute , Mosul , Iraq
shahla_ak71@yahoo.com

(Received: 8 / 6 / 2011 ---- Accepted: 2 / 4 / 2012)

Abstract

The genetic algorithm is one of the intelligent research techniques, which is used for finding approximate solutions for the problems that require optimal solutions.

The objective of this research is to use the parallel genetic algorithm using the technique of Manager/ Worker in parallel computers with Bayesian Information Criterion (BIC) to detect the outlier data in the multiple linear regression model.

Weather forecast data was used including daily temperature, sunshine, wind speed and the day in a year to prepare them in order to construct a predictive model of the daily water vaporization of the cultivated lands in Mosul city for years 1999-2001. The results showed that the BIC Function, as a fitness function in parallel genetic is of high efficiency in detecting the outlier data with least errors, as percentage of error minimizing was 98% . The study that the genetic algorithm factors (selection, mutation, crossover and generations) have an evident impact on minimizing those errors. The Manager/ Worker technique in parallel computer, which operates with the servers of the distributed computers, was of high efficiency in increasing the speed in this type of intelligent genetic algorithms. Moreover, the multiple linear regression method was used to evaluate the results and the adjusted R square (R^2) was also used in judging the results and the value of the adjusted R square was 0.732 before detecting the outlier data and it increased to be 0.974 after detecting them .

Keywords: Parallel Genetic Algorithm, outlier, BIC, manager/worker