

مقارنة بين الخوارزمية الجينية Fast – MCD – Nested Extension والشبكة العصبية الارجاعية (الارتدادية) Back Propagation

أ.م.د. صباح منفي / كلية الادارة والاقتصاد / جامعة بغداد
م.فاطمة عبد الحميد / كلية الادارة والاقتصاد / جامعة بغداد

المستخلص :

تناول البحث استخدام الطرق اللامعلمية الحصينة في تقدير موقع الوسيط وذلك بالاعتماد على اصغر محدد لمصفوفة التباين والتباين المشترك لنموذج انحدار متعدد المتغيرات ، ونظرا لوجود القيم الشاذة وزيادة حجم العينة ووجود اكثر من بعد لنموذج انحدار متعدد المتغيرات لذا يكون من الصعوبة ايجاد موقع الوسيط . فقد تم استخدام الخوارزمية الجينية Fast – MCD- Nested Extension ومقارنتها مع الشبكة العصبية الارجاعية (الارتدادية) Back Propagation متعددة الطبقات من حيث دقة النتائج والسرعة في ايجاد موقع الوسيط ، اما افضل عينة يتم تحديدها من خلال الاعتماد على اقل مسافة (مهلنوبس Mahalanobis Distance) وقد اظهرت الدراسة كفاءة الشبكات العصبية في التقدير التي تمثلت باقل مسافة واصغر محدد لمصفوفة التباين والتباين المشترك ومقارنة التباين مع تباين العينة التي تم تحديدها باستخدام الخوارزمية الجينية Fast – MCD- Nested Extension .

اما الجانب التطبيقي فقد تم استخدام 9 انواع من المؤشرات الكيمياوية والفيزياوية المسببة لتلوث مياه الشرب ، فقد شمل البحث محافظات العراق كافة عدا اقليم كردستان ولعشرة اشهر لسنة 2013 ولحجم عينة 898 .

المصطلحات الرئيسية للبحث/ الخوارزمية الجينية - الشبكات العصبية - اللامعلمية - الارجاعية (الارتدادية) - متعدد المتغيرات - الحصينة .



المقدمة وهدف البحث

MCD واحدة من التقديرات الحصينة المستخدمة لتقدير موقع الوسيط والتشتت لبيانات متعدد المتغيرات التي لها n من المشاهدات و p من المتغيرات ، ونظرا لزيادة عدد المشاهدات ووجود القيم الشاذة (المتطرفة) والتي لا يمكن اهمالها لاهميتها يكون من الصعوبة ايجاد موقع الوسيط عندما يكون لدينا نموذج متعدد المتغيرات ، وذلك لوجود اكثر من متغير $P \geq 2$ وكذلك زيادة عدد المشاهدات. ففي عام ١٩٨٧ اقترح Roussey و Leroy تقدير موقع الوسيط بالاعتماد على الحد الادنى لمحدد مصفوفة التباين والتباين المشترك من خلال ايجاد h من المشاهدات .

لذا فان الهدف من هذا البحث هو استخدام الخوارزمية الجينية Fast – MCD - Nested Extension ومقارنتها مع الشبكة العصبية الارجاعية (الارتدادية) Back Propagation من حيث السرعة والدقة في ايجاد موقع الوسيط وذلك عند زيادة عدد المشاهدات .

1-1 مقدمة في الخوارزمية الجينية [1]

الخوارزمية هي مجموعة محددة من خطوات الحل التي تؤدي الى انجاز وظيفة معينة ويجب ان تتوفر فيها الشروط الاتية:-

- ١- المدخلات (INPUT) : صفر او اكثر من قيمة .
- ٢- المخرجات (OUTPUT) : قيمة واحدة على الاقل .
- ٣- الوضوح (DEFINITENESS) : كل خطوة في الخوارزمية واضحة المعاني وغير غامضة اي يجب ان تفهم من قبل الجميع .
- ٤- المحدودية (FINITENES) : كل خطوات الخوارزميات يمكن حلها في فترة زمنية محددة .
- ٥- المحلولة (EFFECTIVENESS) : كل خطوة تكون ممكنة الحل.

MINIMUM COVARIANCE DETERMINANT (MCD) ESTIMATER 1-2

[3]

اصغر محدد لمصفوفة التباين المشترك اقترح من قبل ROUSSEEW and LEROY في عام 1987 ، الهدف منه ايجاد موقع الوسيط عند زيادة حجم العينة من خلال ايجاد المشاهدات h والتي لها اقل محدد لمصفوفة التباين المشترك ، حيث يتم تقدير الموقع من معدل المشاهدات h وتقدير مصفوفة التباين المشترك وذلك من خلال تحديد نقطة التوقف كالاتي :-

$$G = \frac{n-h+1}{n}$$

حيث ان :-

G : تمثل نقطة التوقف

n : تمثل عدد المشاهدات

h : عدد صحيح يمثل الحجم وهو اعلى نقطة مساوية الى

$$h = \frac{n+p+1}{2}$$

$$\frac{n+p+1}{2} \leq h \leq n$$

اما افضل نقطة توقف نحصل عليها عندما تكون

$$h = 0.75 n$$

p : تمثل عدد المتغيرات

لذا يمكن كتابة الحل الامثل الى التقدير ب mcd كالاتي :-

$$\text{Minimize } K(X_{i1}, \dots, X_{ih}) = \det [\text{cov}(X_{i1}, \dots, X_{ih})]$$

إذا كانت $h=n$ و $p=1$ البيانات تمثل Univariate Model

$$\{X_{i1}, \dots, X_{ih}\} \subset \{X_i, \dots, X_n\}$$

$$, \dots, X_n \in \mathbb{R}^d X_1$$

أذن موقع التقدير mcd يمثل T المعدل و S يمثل مصفوفة التباين المشترك .
 أما عندما $P \geq 2$ يكون Multivariate Model حيث يتم إيجاد المسافة (Mahalanobis Distance) (Mahanobis Distance) بالاعتماد على المتغيرات التوضيحية كالآتي :-

$$MD (X_i) = \sqrt{(X_i - T_0)' S_0^{-1} (X_i - T_0)}$$

$$X_n = \{X_1, \dots, X_n\}$$

$$H_1 \subset \{1, \dots, n\}$$

$$|H_1| = h$$

$$T_1 = \frac{1}{h} \sum_{i \in H_1} X_i$$

$$S_1 = \frac{1}{h} \sum_{i \in H_1} (X_i - T_1)(X_i - T_1)'$$

نجد المسافة (MD) Mahalanobis Distance الى

$$= \sqrt{(X_i - T_1)' S_1^{-1} (X_i - T_1) d_1(i)}$$

$$\{d_1(i): i \in H_2\} = \{(d_1)_{1:m}, \dots, (d_1)_{h:m}\} \quad i = 1, \dots, n$$

لايجاد المسافة الى H_2

$$(d_1)_{1:m} \leq (d_1)_{2:m} \leq \dots \leq (d_1)_{n:m}$$

نجد T_2 و S_2 المستندة على H_2

$$\leq |S_2| |S_1| \quad \text{إذا وفقد إذا}$$

$$|S_1| \neq 0 \quad \text{و} \quad S_2 = S_1 \quad \text{و} \quad T_2 = T_1$$

بالاعتماد على اصغر محدد لمصفوفة التباين المشترك نطبق الخوارزمية Step – C كالآتي :-

or pair (T_{old}, S_{old})

h-subset H_{old}

1- $d_{old}(i)$, $i = 1, \dots, n$ نجد المسافة

2- $d_{old}(\pi_1) \leq d_{old}(\pi_2) \leq \dots \leq d_{old}(\pi_n)$ نأخذ التباديل الى المسافة

3- $H_{new} = \{\pi_1, \dots, \pi_h\}$

4- $T_{new} = ave(H_{new})$

$$S_{new} = cov(H_{new})$$

لكل $d_{old}(i)$ نكرر Step – C لحين الحصول على $|S_2| = 0$ او $|S_2| = |S_1|$

عدا ذلك يتم ايجاد $|S_3|$ فتكون السلسلة $|S_3| \geq |S_2| \geq |S_1|$

قد تتكرر Step – C الى M من المرات لحين الحصول على $|S_m| = 0$ او $|S_m| = |S_{m-1}|$

اي لحين الحصول على التقارب (Converge) بعد ذلك نكرر Step – C الى

(T_m, S_m)

فتكون السلسلة $|S_m| \geq |S_{m-1}| \geq |S_1| \geq |S_2|$

ويكون الحل لأصغر محدد للتباين المشترك .

[4] FAST ALGORITHM – MCD1-2-1

الهدف من هذه الخوارزمية هو ايجاد h من المشاهدات خارج n ، الفكرة الرئيسية هو عدم تساوي محدد الرتبة . هناك اسلوبان لهذه الخوارزمية الاسلوب الاول (Selective iteration) تستخدم عندما تكون العينة صغيرة اما الاسلوب الثاني (Nested extension) تستخدم عند زيادة حجم العينة وسنأتي الى شرحها بالتفصيل لاحقا .

وقد اعطت هذه الخوارزمية دقة وسرعة للنتائج بالاضافة الى ذلك يمكن ايجاد اعلى مستوى (Hyper Plane) يحوي h عند زيادة عدد المشاهدات فهي تعتبر اداة لتحليل بيانات متعدد المتغيرات . لذا من الصعوبة ايجاد الموقع عندما تكون $P > 2$ (P- Variate) ، لذا تم الاعتماد على المسافة مهلنوبس (Mahalanobis Distance) المستندة لاعلى تقدير حصين لموقع متعدد المتغيرات ، والتي يتم فيها الاعتماد على المتغيرات التوضيحية .

FAST – MCD – NESTED EXTENSION 1-2-2

تستخدم عندما تكون حجم العينة n كبير حيث تجزأ العينة الى مجموعات جزئية منفصلة ويمكن ان تكون المجموعات الجزئية غير متساوية بالحجم ، اما عملية سحب العينات لتكوين المجموعات الجزئية تتم بشكل عشوائي ، بالاضافة الى ذلك لا تحتاج الى وقت عند التطبيق والخطوات التالية توضح طريقة العمل لهذه الخوارزمية كالآتي :-

نقسم حجم العينة n الى مجموعات جزئية منفصلة

1- وعشوائية

$$n = n_1 + n_2 + \dots + n_m$$

2- كل مجموعة جزئية تكرر كالآتي :-

$$R = \frac{n_i}{m}$$

حيث ان :-

R : تمثل عدد مرات التكرار
 n_i : تمثل حجم العينة و $i = 1, 2, \dots, m$
 m : تمثل عدد المجموعات

3- تكون H_1 مجموعة جزئية اولية حجم $h_{sub} \left(\frac{h}{n} \right) [n_{sub}]$

4- ننفذ الخطوة 2 من C- STEP الى n_{sub} ، h_{sub}

5- نأخذ افضل 10 نتائج الى (T_{sub}, S_{sub})

6- جمع (دمج) $n_{merged} = n$

7- دمج المجموعة يكرر 50 مرة ونجد (T_{sub}, S_{sub})

8- ننفذ الخطوة 2 من C- STEP باستعمال $\left(\frac{h}{n} \right) [n_{merged}]$ ، h_{merged} ، n_{merged}

9- نجد افضل 10 (T_{merged}, S_{merged})

10- بعد اكمال المجموعات الجزئية ، كرر m_{full} لافضل النتائج

11- نكرر C- STEP باستعمال n و h

12- احتفظ بافضل اخر نتيجة ل (T_{full}, S_{full})

13- m_{full} والعدد من خطوات C- STEP يعتمد على حجم مجموعة البيانات

1-3 مقدمة في الشبكات العصبية [6] [8]

الشبكات العصبية يمكن ان تعرف بأنها نظم او تقنية لمعالجة المعلومات منبثقة من دراسة المخ والنظام العصبي للانسان ، او هو نظام معالجة معلومات يحاكي اسلوب الشبكات العصبية الحيوية وله مميزات اداء



مقارنة بين الخوارزمية الجينية Fast – MCD – Nested Extension والشبكة العصبية الارتجاعية [الارتدادية] Back Propagation

معينة أو خوارزميات أو نماذج رياضية لمعالجة المعلومات .

تتكون الشبكة العصبية من مجموعة من العصبونات أو الوحدات أو الخلايا أو العقد موزعة في طبقات وهي :-

1- طبقة الدخل

2- طبقة الخرج

3- طبقة المعالجة

ترتبط العقد مع بعضها البعض بروابط اتصال وترفق بها قيم عددية (اوزان) وتعتبر المعلومات الاساسية لحل المشكلة وفيها يبدأ الحل حيث تستقبل طبق الدخل الاشارة وتعمل على توزيعها الى الطبقات الاخرى حسب نوع الاشارة .

- الخلية العصبية تستقبل الاشارة وتعالجها ونحصل على مخرجات

Input → process → output

- الخلية العصبية تستمد مميزاتا من الخلية العصبية الحيوية .

- كل عصبون له Active معين ناتج من عملية المعالجة .

- الروابط هي عبارة عن connection واوزان و sum وهي العملية الاساسية في العصبون ونعني بها المحاكاة للخلية العصبية الحية فالعصبون الصناعي هو محاكاة للعصبون الحي .

اما ال sum هو جمع اشارة الدخل مضروبة في الاوزان $\sum x_i w$

$x \Rightarrow$ cell /unit/node

$w \Rightarrow$ weight on connection

- العصبون في طبقة الدخل لا تتم عليه اي عملية معالجة بل ادخال فقط .

- المحاكاة هي العملية الاساسية للعصبون وهناك عمليات اخرى اساسية لعملية المعالجة في الشبكة العصبية هي :-

1- neurons / node

2- connection /link

3- weights

4- summation

5- activation function

اهم نقطة في الشبكات العصبية هي الاوزان وهي الحل الامثل لحل المشكلة بالاضافة الى معمارية الشبكة Activation Function الميزة الاساسية للتفريق بين انواع الشبكات العصبية ، شبكة عصبية وحيدة

الطبقة Linear وشبكة عصبية متعددة الطبقات Non Linear .

1-4 الشبكات متعددة الطبقات MultiLayer Networks

[7] [2] [5]

تمتلك هذه الشبكات طبقة واحدة او اكثر من العقد المخفية (Hidden Nodes) هذه العقد لا تمثل وحدات ادخال او اخراج ، ولها استخدامات عديدة مثل تمييز النماذج، تمييز الاصوات، تحليل البصمة الصوتية، في خوارزميات الضغط والخوارزميات الجينية... الخ .

وتستخدم الدالة السببية (sigmoid function) كدالة تحويلية وهناك ثلاث انواع من الدوال :-

$$1- F(x) = \tanh(x B)$$

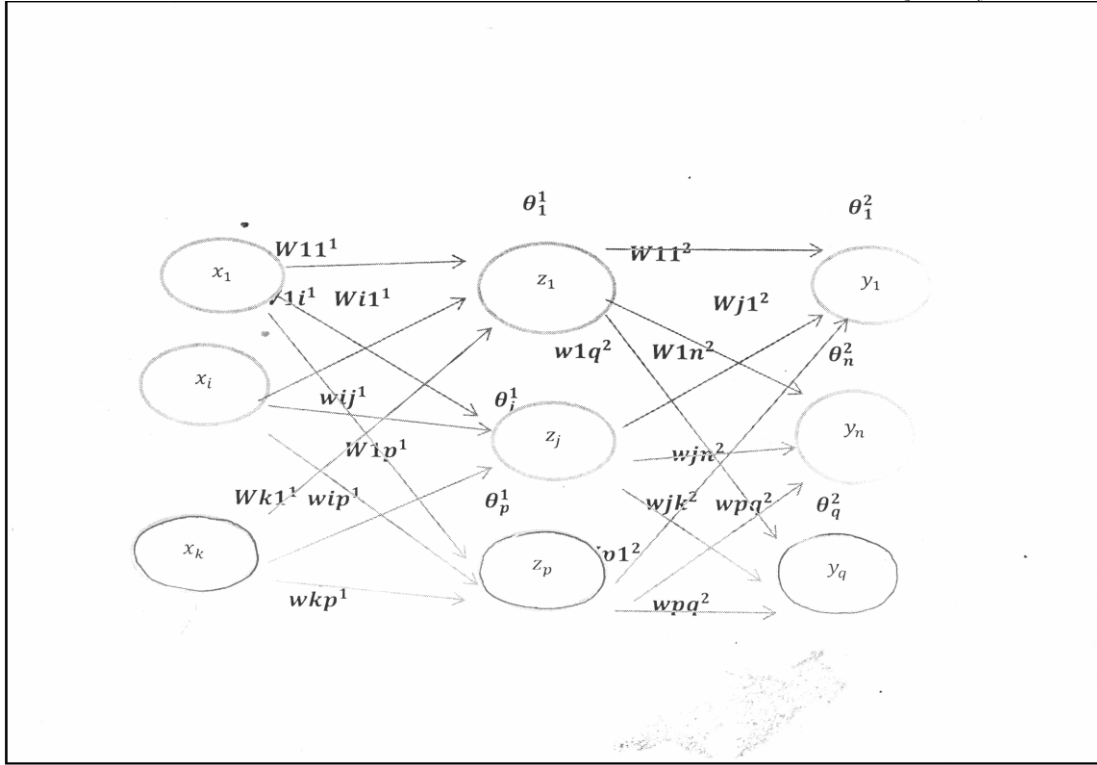
$$2- F(x) = \frac{e^{Bx} - e^{-Bx}}{e^{Bx} + e^{-Bx}}$$

$$3- F(x) = \frac{1}{1 + e^{-Bx}} \quad B > 0$$

وقبل البدء باستخدام الشبكات العصبية لابد من تحديد الاتي :-

1- العتبة Threshold

2- الدالة التفعيلية Activation function والشكل الاتي يوضح معمارية الشبكات متعددة الطبقات :-



اما افضل خوارزمية مستخدمة في تطبيق الشبكات العصبية متعددة الطبقات هي خوارزمية الانبعاث الخلفي Back Propagation Algorithm يتم في هذه الخوارزمية ايجاد الخطأ بين المخرج المطلوب والفعلي وتعديل الاوزان بالرجوع به من الطبقة الاخيرة الى الطبقات الخفية ثم الى طبقة المدخلات ، حيث يستخدم التعليم المراقب لتقليل الخطأ باتجاه الصفر ، خلال عملية التعليم يتم تدريب الشبكة باستخدام الدالة السببية (الدالة التفعيلية) Activation functions والتي تكون قابلة للتفاضل (Differentiation) بالاضافة ان لا تكون ثنائية القيمة وان تكون قادرة على النمذجة اللاخطية ، وتوجد انواع مختلفة من الدوال التفعيلية :-

- 1- الدالة الخطية (Linear function) وتستعمل عادتا في خلايا طبقة المخرجات .
 - 2- Sigmoid function تستعمل في الطبقات الاخرى .
- من خواص الانبعاث الخلفي (الارتدادية) تقليل الخطأ وقابليتها للتعامل مع البيانات المشوشة وكذلك قدرتها على التعامل مع الدوال الخطية واللاخطية القابلة للاشتقاق .
- اما خطوات خوارزمية الانبعاث الخلفي (الارتدادية) هي كالآتي :-
- 1- اعطاء قيم عشوائية للاوزان w .
 - 2- تعريف معامل التعليم (قيمة معامل التعلم الدقيقة a=0.5) .
 - 3- ادخال مصفوفة تمثل مدخلات ومخرجات التي تم التدريب عليها .
 - 4 - تطبيق عملية الانتشار الامامي لتحديد مخرجات الشبكة
- $$z_p = \sum_k w_{kp}^1 x_k$$
- x_k : تمثل المدخلات التي تم التدريب عليها .

w_{kp}^1 : تمثل الاوزان العشوائية التي تربط طبقة المدخلات والعقد الخفية .
بعد مرور هذه القيمة على الخلية والممثلة في الدالة السببية sigmoid function يكون مخرج كل خلية في

$$y_q = \sum_p w_{pq}^2 \theta_p^1$$

$$\theta_p^1$$

$$\theta_q^2 = \frac{1}{1 + e^{-y_q}}$$

5 - مقارنة المخرجات الفعلية مع المخرجات المطلوبة وتحديد قيمة الخطأ

$$\delta_1 = (t_k - \theta_q^2) \theta_q^2 (1 - \theta_q^2)$$

t_k : تمثل مخرج الخلية .

δ_1 : الخطأ .

تعديل الاوزان التي تربط الطبقة الخفية وطبقة المدخلات في الطبقة الثانية

$$w_{pq} \text{ new} = w_{pq} \text{ old} + a \delta_1 \theta_q^2$$

a=0.5 تمثل معامل التعلم .

6 - بعد الوصول الى الطبقة الخفية نعيد العمليات السابقة في تعديل الاوزان بعد تحديد قيمة الخطأ

$$\delta_2 = \theta_q^2 (1 - \theta_q^2) \sum_q w_{pq}^2 \delta_1$$

وبالتالي يتم تحديد الاوزان الجديدة بين طبقة المدخلات والطبقة الخفية

$$\text{new} = w_{kp} w_{kp} \text{ old} + a \delta_2 x_k$$

تكرر هذه الخطوات لعدة مرات وذلك للحصول على اقل نسبة للخطأ .

الجانب التطبيقي

المقدمة

في هذا الفصل سوف يتم تطبيق ما تم التطرق اليه في الجانب النظري باستخدام بيانات تشمل انواع المؤشرات الكيماوية والفيزيائية المسببة لتلوث مياه الشرب ، حيث ان زيادة في نسبة المؤشرات يؤدي الى تلوث الماء، وقد تم استخدام تسع انواع من المؤشرات لكافة محافظات العراق عدا اقليم كردستان . اما المقارنة فقد كانت بين الخوارزمية الجينية Fast – MCD – Nested Extension والشبكة العصبية متعددة الطبقات الارجاعية (الارتدادية) Back Propagation من حيث السرعة والدقة في ايجاد موقع الوسيط عندما يكون لدينا نموذج انحدار متعدد المتغيرات اللامعلمي .

1- جمع البيانات

اعتمدت الدراسة على الاحصاءات البيئية للعراق لسنة 2013 الصادرة من الجهاز المركزي للاحصاء مقسمة الى (10) أشهر وتشمل المحافظات كافة عدا اقليم كردستان وقد تم استخدام (9) انواع من المؤشرات الكيماوية والفيزيائية المسببة لتلوث مياه الشرب موضحة في الجدول رقم (1) والذي يبين فيه المؤشرات الداخلة في مياه الشرب مع الرمز ووحدة القياس المطابقة للمواصفات العراقية حيث ان الزيادة في الحدود الدنيا للقياس تسبب تلوث في مياه الشرب .

جدول رقم (1) : المؤشرات الداخلة في مياه الشرب مع الرمز الحدود الدنيا للقياس ووحدة القياس للمواصفات العراقية .



مقارنة بين الخوارزمية الجينية Fast – MCD – Nested Extension والشبكة العصبية الارتدادية [الارتدادية] Back Propagation

المؤشر	الرمز	وحدة القياس	الحدود الدنيا للقياس
العكورة	Turbidity	NTU	5
العسرة الكلية	T.H	mg/ L	500
القاعدية	ALK.	mg/ L	0.2
الأملاح الذائبة الكلية	T.D.S	mg/ L	1000
الكوريدات	Cl	mg/ L	350
الكالسيوم	Ca	mg/ L	150
المغنسيوم	Mg	mg/ L	100
الصوديوم	Na	mg/ L	200
الكبريتات	SO4	mg/ L	400

المصدر: الإحصاءات البيئية للعراق لسنة 2013 الصادر من الجهاز المركزي للإحصاء .

2- التقديرات باستخدام الخوارزمية الجينية MCD :-

2-1 Fast – MCD – Nested Extension

$$Y_i = f_i(X) \quad , (i=1, \dots, 10)$$

حيث ان :-

Y_i : يمثل المتغير المعتمد (نسبة المؤشرات المسموح او غير المسموح بها) .

X : يمثل المتغير المستقل (المؤشرات الداخلة في تلوث مياه الشرب) .

$$X=(X_1, \dots, X_9)$$

$$x_{ijk} \quad , i=1,2,\dots,10 \quad X_i =$$

$$J=1,2,\dots,ni$$

$$K=1,2,\dots,9$$

ويمكن اعادة كتابتها كالآتي:-

$$Y_1=(y_{11}, \dots, y_{1n1}) , Y_2=(y_{21}, \dots, y_{2n2}) , \dots, Y_{10}=(y_{101}, \dots, y_{10n10})$$

$$X_1=(x_{11k}, \dots, x_{1nik}) , X_2=(x_{21k}, \dots, x_{2nik}) , \dots, X_9=(x_{91k}, \dots, x_{9nik})$$

لايجاد تقدير موقع الوسيط باستخدام اصغر محدد لمصفوفة التباين المشترك ولتطبيق الخوارزمية الجينية (Nested – FAST MCD) تتم في المرحلة الاولى تجزئة حجم العينة (N=898) الى عشرة مجموعات جزئية (m=10) كالآتي :-

$$N=n_1+n_2+\dots+n_{10}$$

$$(n_1=98 , n_2=90 , n_3=84 , n_4=87 , n_5=93 , n_6=91 , n_7=72 , n_8=87 , n_9=96 ,$$

$$n_{10}=100)$$

مقسمة حسب الاشهر وكل شهر يشمل (16) محافظة عدا اقليم كردستان .

يتم التقدير الى h من المشاهدات حيث ان :-

$$h=(N+p+1)/2$$

اما التكرار (R_i) يمكن ايجاده كالآتي :-

$$R_i = (n_i/m)$$

ثم يتم ايجاد h_{sub} حيث ان :-

$$h_{sub} = (n_i * (h/N))$$

يتم ايجاد المسافة (Mahalanobis distance) كالآتي :-

$$= \sqrt{(x_i - T_{sub})' S_{sub}^{-1} (x_i - T_{sub})} md_{1(i)}$$

$$i=1,2,\dots,n$$

$$= \frac{1}{h_{sub}} \sum_{i=1}^{n_i} x_i T_{sub}$$

$$= \frac{1}{h_{sub}} \sum_{i=1}^{n_i} (x_i - T_{sub}) (x_i - T_{sub})' S_{sub}$$

وبالاعتماد على التكرار R_i حيث يتم في كل تكرار حذف صف ويتم ايجاد المسافة و (T_{sub}, S_{sub}) تكرر المرحلة الاولى لجميع العينات الجزئية ثم نرتب المسافة تصاعديا $(md_{1(1:n)}, md_{1(2:n)}, \dots)$ ويتم اختيار افضل (10) نتائج الى (T_{sub}, S_{sub}) لاصغر مسافة واصغر محدد الى مصفوفة التباين المشترك وفيما يلي نتائج المرحلة الاولى :-

n_i	R_i	h_{sub}
n_1	10	50
n_2	9	46
n_3	8	42
n_4	9	44
n_5	9	47
n_6	9	46
n_7	7	36
n_8	9	44
n_9	10	49
n_{10}	10	51

اما افضل (10) نتائج لاصغر مسافة واصغر محدد الى مصفوفة التباين المشترك والمرتبة تصاعديا تمثلت في مشاهدات العينة n_7 والعينة n_3 ، يتم الدمج بين العينات الجزئية كالآتي :-

$$= n_7 + n_3 A_1$$

$$= 156$$

$$= n_4 + n_8 A_2$$

$$= 174$$

$$= n_1 + n_2 A_3$$

$$= 188$$

$$= n_5 + n_6 A_4$$

$$= 184$$

$$= n_9 + n_{10} A_5$$

$$= 196$$

نعيد خطوات الخوارزمية الجينية التي تم تطبيقها في المرحلة الاولى بالاستناد الى العينات الجزئية التي تم

دمجها لايجاد اصغر مسافة افضل (10) تمثلت في مشاهدات العينة الجزئية A_1 الناتجة من الدمج بين العنيتين الجزئيتين (n_7, n_3) ، نكرر الدمج مرة اخرى بين العينات الجزئية كالاتي :-

$$= A_1 + A_2 + A_3 AA_1$$

$$= 518$$

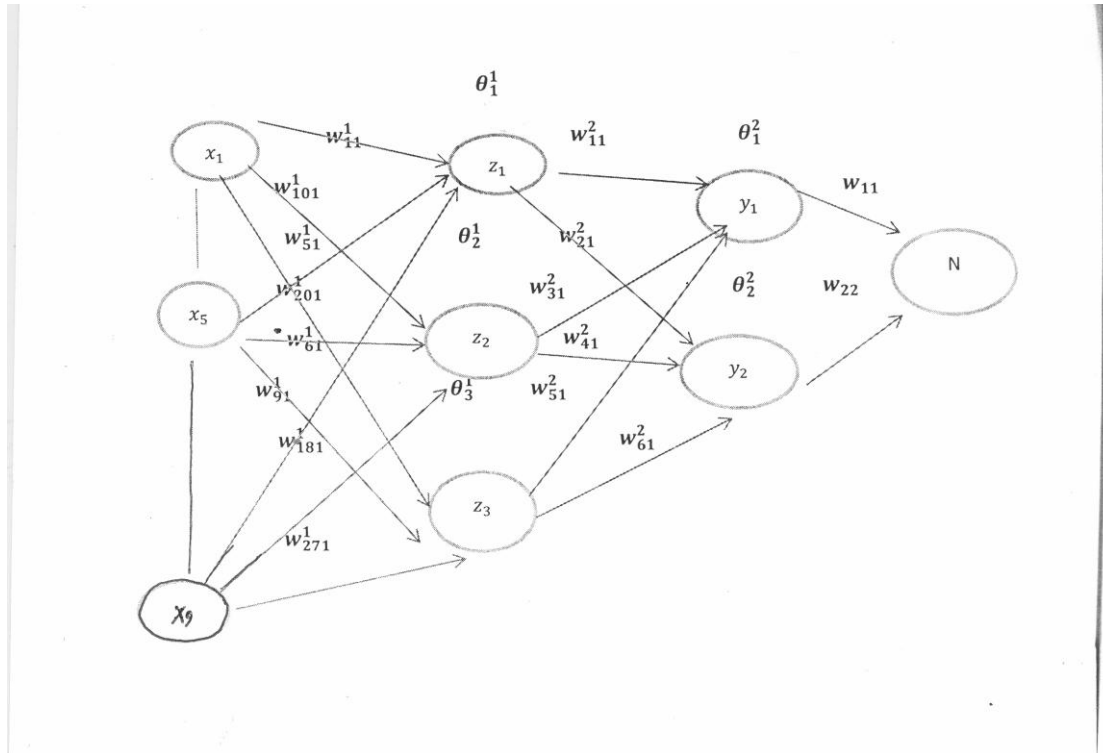
$$AA_2 = A_4 + A_5$$

$$= 380$$

ثم نعيد خطوات الخوارزمية بالاستناد الى العينات المدمجة اعلاه ، اما في المرحلة الاخيرة يتم الدمج بين AA_1 و AA_2 واعادة خطوات الخوارزمية الجينية ثم يتم اختيار افضل (10) لديها اصغر مسافة واصغر محدد لمصفوفة التباين المشترك بين العشرة المحددة في المرحلة الاخيرة ، اما قيمة التكرار $(R_i = 10)$ و $(h_{sub} = 454)$ وقيمة المسافة $(md=367386.285592203)$ والمحدد الى مصفوفة التباين المشترك $(\det(S) = 5.44263195032355e+35)$.

2-2-3- الشبكات العصبية متعددة الطبقات Multilayer Networks

لتطبيق الشبكات متعددة الطبقات لابد من تحديد عدد الطبقات وعدد العصبون neuron لكل طبقة بالاضافة الى تحديد الاوزان التي سيتم توليدها والشكل التالي يوضح مخطط الشبكة التي سيتم التدريب عليها :-





مقارنة بين الخوارزمية الجينية Fast – MCD – Nested Extension والشبكة العصبية الارتجاعية [الارتدادية] Back Propagation

x_n : تمثل المدخلات (المؤشرات الداخلة في تلوث مياه الشرب) و $n=1,2,\dots,9$

z_p : الطبقة الخفية و $p=1,2,3$

y_q : الطبقة الخفية و $q=1,2$

w_{kp}^1 : الاوزان العشوائية في الطبقة الاولى و $k=1,2,\dots,27$

w_{jq}^2 : الاوزان العشوائية في الطبقة الثانية و $j=1,2,\dots,6$

w_{11} و w_{22} الاوزان العشوائية

θ_p^1 : العصبون في الطبقة الاولى الذي يربط طبقة المدخلات مع الطبقة الخفية

θ_q^2 : العصبون في الطبقة الثانية

تم استخدام خوارزمية الانبعاث الخلفي (الارتدادية) والتعليم المراقب لتدريب الشبكة العصبية متعددة الطبقات، ففي المرحلة الاولى تم توليد الاوزان العشوائية التي تدخل طبقة المدخلات وفيها يتم ضرب الاوزان مع كل عنصر دخل x_i ثم يتم ايجاد مجموع المدخلات للخلية الاولى في الطبقة الخفية كالآتي :-

$$= \sum_k w_{kp}^1 x_i z_p$$

يتم تمرير المدخلات وصولاً الى طبقة المخرجات وذلك باستخدام الدالة السببية

$$\theta_p^1 = \frac{1}{1+e^{-z_p}}$$

والجدول التالي يوضح الاوزان العشوائية التي تم توليدها في الطبقة الاولى :

الاوزان العشوائية في الطبقة الاولى
0.8234578283
0.6948286229
0.3170994800
0.9502220488
0.0344460805
0.4387443596
0.3815584570
0.7655167881
0.7951999011
0.1868726045
0.4897643957
0.4455862007
0.6463130101
0.7093648308
0.7546866819
0.2760250769
0.6797026768
0.6550980039
0.1626117351
0.1189976815
0.4983640519
0.9597439585
0.3403857266
0.5852677509
0.7546866819
0.7512670593
0.2550951154

يتم بعدها توليد اوزان الطبقة الثانية حيث تضرب الاوزان مع المخرجات التي تم الحصول عليها من الطبقة الاولى وتجميعها في الطبقة الخفية كالآتي :-

$$y_q = \sum_p w_{pq}^2 \theta_p^1$$

يتم تمرير مدخلات الطبقة الخفية باستخدام الدالة السيبيية الآتية :

$$\theta_q^2 = \frac{1}{1+e^{-y_q}}$$

الجدول التالي يوضح الاوزان العشوائية في الطبقة الثانية :-

اوزان الطبقة الثانية
٠.٥٦٤٥٠٣٦٣٨٩
٠.٧٧٤٢٧٤٩٦٢٠
٠.٩٤٩٤٤٩٨٣٩٨
١.٠٣٤٤٨٩٦٦٤٥
٠.٦٠٥٧٦٢١١٧٢
٠.٢١٣٨٢٢٦٨٢٢

يتم بعدها تحديد قيمة الخطأ وذلك من خلال مقارنة المخرجات الفعلية مع المخرجات المطلوبة (t_k) حيث ان قيمة المخرجات الفعلية يتم ايجادها من ضرب الاوزان العشوائية w_{11} و w_{22} مع θ_q^2 وايجاد مجموع

المدخلات لخلية طبقة المخرجات N وبهذا يكون المخرج الفعلي للشبكة $O = \frac{1}{1+e^{-N}}$ و $N = w_{11} \theta_1^2 + w_{22} \theta_2^2$

اما قيمة الاوزان العشوائية التي تم توليدها هي كالآتي :-

$$= 1.0062466544w_{11}$$

$$= 1.1013133923w_{22}$$

لتعديل الاوزان نبدأ اولا بتحديد قيمة الخطأ

$$\delta_1 = (t_k - O) * O * (1 - O)$$

بعد تحديد قيمة الخطأ يتم تعديل الاوزان التي تربط بين الطبقة الخفية وطبقة المخرجات من المعادلة التالية :-

$$w_{11}^{new} = w_{11}^{old} + a\delta_1 \theta_1^2$$

$$w_{22}^{new} = w_{22}^{old} + a\delta_1 \theta_2^2$$

اما قيمة معامل التعلم $a=0.5$

ثم الانتقال في هذه المرحلة من طبقة المخرجات الى الطبقة الخفية .
ولتعديل اوزان الطبقة الثانية يتم اولا ايجاد قيمة الخطأ

$$\delta_{11} = \theta_1^2 (1 - \theta_1^2) w_{11}^{new} \delta_1$$

$$\delta_{22} = \theta_2^2 (1 - \theta_2^2) w_{22}^{new} \delta_1$$

ثم تعدل الاوزان في الطبقة الثانية كالآتي :-

$$w_{11}^2^{new} = w_{11}^2^{old} + a\delta_{11} \theta_1^1$$

$$w_{21}^2^{new} = w_{21}^2^{old} + a\delta_{22} \theta_1^1$$

$$w_{31}^2^{new} = w_{31}^2^{old} + a\delta_{11} \theta_2^1$$

$$w_{41}^2^{new} = w_{41}^2^{old} + a\delta_{22} \theta_2^1$$

$$w_{51}^2^{new} = w_{51}^2^{old} + a\delta_{11} \theta_3^1$$

$$w_{61}^2^{new} = w_{61}^2^{old} + a\delta_{22} \theta_3^1$$

والان يمكن تعديل الازان في طبقة المدخلات بعد ايجاد قيمة الخطأ بين الطبقة الخفية الاولى وطبقة المدخلات كالاتي :-

$$\delta_1^1 = \theta_1^1 (1 - \theta_1^1) (w_{11} \text{new} \delta_{11} + w_{21} \text{new} \delta_{22})$$

$$\delta_2^1 = \theta_2^1 (1 - \theta_2^1) (w_{31} \text{new} \delta_{11} + w_{41} \text{new} \delta_{22})$$

$$\delta_3^1 = \theta_3^1 (1 - \theta_3^1) (w_{51} \text{new} \delta_{11} + w_{61} \text{new} \delta_{22})$$

اما الازان يمكن ايجادها من المعادلات التالية :-

$$w_{11} \text{new} = w_{11} \text{old} + a \delta_1^1 x_1$$

:

$$w_{91} \text{new} = w_{91} \text{old} + a \delta_1^1 x_9$$

$$w_{101} \text{new} = w_{101} \text{old} + a \delta_2^1 x_1$$

:

$$w_{181} \text{new} = w_{181} \text{old} + a \delta_2^1 x_9$$

$$w_{191} \text{new} = w_{191} \text{old} + a \delta_3^1 x_1$$

:

$$w_{271} \text{new} = w_{271} \text{old} + a \delta_3^1 x_9$$

تم تكرار خطوات خوارزمية الشبكة العصبية (50) مرة للحصول على اقل نسبة للخطأ بعدها يتم ايجاد مصفوفة التباين والتباين المشترك للعينات ثم تحديد افضل عينة من خلال ايجاد اصغر مسافة Mahalanobis Distance وقد تمثلت في مشاهدات العينة (n₆) .

3- الاستنتاجات

لغرض المقارنة بين الخوارزمية الجينية Fast – MCD-Nested Extension والشبكات العصبية متعددة الطبقات حيث اظهرت الدراسة السرعة والدقة في ايجاد موقع الوسيط لبيانات نموذج متعدد المتغيرات وباستخدام الطرق الحصينة والذي اعتمد فيه البحث على اصغر محدد لمصفوفة التباين والتباين المشترك من خلال ايجاد اصغر مسافة Mahalanobis Distance تمثل في الشبكات العصبية من خلال تمرير الازان العشوائية التي يتم توليدها في المراحل المختلفة وتعديل الازان للحصول على اقل نسبة للخطأ وقد تم تكرار الخوارزمية (50) مرة ، والجدول التالي يوضح النتائج مبين فية التباين والمسافة والمحدد لأفضل عينة محددة كالاتي :-

	الخوارزمية الجينية Fast-MCD-Nested Extension	الشبكات العصبية
Var(x)	1.5136e+03	23.9724
	3.4616e+05	4.5581e+03
	3.84612e+04	241.2565
	1.6140e+06	1.6062e+05
	1.1721e+05	3.5979e+03
	1.8988e+04	429.5391
	4.7462e+03	524.9195
	2.5011e+04	1.4867e+03
	1.3355e+05	4.1446e+03
	Mahalanobis Distance المسافة	3.8515e+05
Determinant المحدد	1.4037e+35	7.5644e+23



المصادر

- 1- طعمة، د-حسن ياسين والاستاذ المساعد هدرستم محمد شعبان والباحث حسن ثابت رشيد "تحليل وتصميم الخوارزميات، 1-159 .
- 2- صباح منفي رضا، "تحديد هوية المتكلم باستخدام الشبكات العصبية " كلية الادارة والاقتصاد ، جامعة بغداد ، 2004 .
- 3- Chakraborty,B.and Chaudhuri,p."On The Use Genetic Algorithm With Elitism in Robust and Nonparametric Multivariate Analysis",Austrian Journal of Statistics ,Vol.32,2003,p.13-27.
- 4- P.J.Rousseeuw and K.Van Driessen.,“Afast algorithm for the minimum covariance determinant estimator”,Technometrics,Vol.41,No.3,August,1999,p.212-223.
- 5- Johnston , R.D.,”Are speech Recognitionstill 98% Accurate or has the time come Repeal HYDE Law “,BT Tehnoly ,vol_14,No.1,pp 165-176,January , 1996 .
- 6- Kinnerbrock , W.,”Neural Network”, Suneel Galgotia , India ,1995 .
- 7-Pandya,A.S.,and Marcy,R.B.,”Pattern Recognition with Neural Network in c^{++} “,CRC press,Florida 1996 .
- 8-Wu,J.,”Neural Network and Simulation Methods”, Marcel Dakker ,Inc ,New York,1994.



Compared with Genetic Algorithm Fast – MCD – Nested Extension and Neural Network Multilayer Back propagation

Abstract

The study using Nonparametric methods for robust to estimate a location and scatter it is depending minimum covariance determinant of multivariate regression model , due to the presence of outlier values and increase the sample size and presence of more than after the model regression multivariate therefore be difficult to find a median location .

It has been the use of genetic algorithm Fast – MCD – Nested Extension and compared with neural Network Back Propagation of multilayer in terms of accuracy of the results and speed in finding median location ,while the best sample to be determined by relying on less distance (Mahalanobis distance)has the study showed the efficiency of neural network in application which represented less distance and smaller determinant matrix of variance covariance compared with sample of Fast – MCD – Nested Extension .

As for practical side has been 9 kinds of chemical and physical indicators for water pollution, the research covered all the provinces of Iraq except Kardistan region and ten month of the year in 2013 and sample size of 898 .

Keywords : Genetic Algorithm , Neural Network , Back Propagation , Multilayer , Robust , Nonparametric , Multivariate