المجلة العراقية للعلوم الإحصائية

http://stats.uomosul.edu.iq

استخدام الشبكة المرنة لاختيار متغيرات السلاسل الزمنية عالية الابعاد لنموذج الانحدار الذاتي لحركة دودة الربداء الرشيقة

محمد خميس رشيد **اسامة بشير شكر**

قسم الإحصاء والمعلوماتية، كلية علوم الحاسوب والرياضيات، جامعة الموصل، الموصل، العراق.

معلومات النشر

تاريخ المقالة:
تاريخ الاستلام:10 تموز 2023
تاريخ القبول: 3 تشرين الثاني 2023
تاريخ القبول: 13 تشرين الثاني 2023
متاح على الإنترنت: 1 كانون الاول 2023

الكلمات الدالة:
اختيار متغيرات الانحدار الذاتي، السلاسل الزمنية عالية الابعاد، نموذج الانحدار الذاتي، الشبكة المرنة، النموذج الهجين Elastic-AR

المراسلة:
محمد خميس رشيد
albadrane2003@gmail.com

الخلاصة

ان عملية اختيار المتغيرات الإحصائية التي تحتوي على معلومات تتعلق بالتأثير على المتغير المعتمد لها دوراً اساسياً في النمذجة الاحصائية الدقيقة. في السلاسل الزمنية وعندما يكون هناك عدد كبير جداً من متغيرات الانحدار الذاتي Autoregressive (AR) في النموذج فإنه من المهم اختيار متغيرات الانحدار الذاتي المؤثرة فعلياً والمهمة من مجموعة كبيرة (عالية الابعاد) من المتغيرات الذاتية بتخلفات زمنية سابقة للحصول على نتائج أكثر دقة. تعد طريقة الشبكة المرنة Elastic Net من الأساليب التي تعمل على اختيار النموذج الافضل واجراء تقدير مشترك للنماذج الخطية مما يسهم في اختيار المتغيرات المؤثرة فعلياً واهمال ما دونها من مجموعة كبيرة جداً عالية الابعاد من المتغيرات الذاتية. في هذه الدراسة سيتم استخدام طريقة الشبكة المرنة لاختيار معلمات الانحدار الذاتي في نموذج السلاسل الزمنية وتقديرها. سيتم استخدام بيانات السلسلة الزمنية لحركة الربداء الرشيقة متمثلة بزوايا الظل للحركة الموجية للدودة Caenorhabditis Elegans (CE). تم اختيار نموذج السلسلة الزمنية أحادية المتغير لحركة الربداء الرشيقة عبر طريقة الشبكة المرنة ونماذج الانحدار الذاتي (Elastic-AR) الهجين بعد عمليات متعددة لاختيار متغيرات الانحدار الذاتي. ومن خلال النتائج فقد تطابقت المعلمات المختارة في نموذج الانحدار الذاتي AR مع النموذج الهجين Elastic-AR الى حد كبير مع تفوق واضح في نتائج الأسلوب الهجين ودقة عالية. ولذلك فمن الممكن استنتاج امكانية استخدام الأسلوب الهجين المقترح للحصول على أفضل نموذج للسلاسل الزمنية عالية الابعاد بأقل عدد من المتغيرات واكثرها تأثيراً مما يقلل من الجهد والتكاليف ويزيد من دقة النماذج.

DOI: 10.33899/IQJOSS.2023.0181216 , ©Authors, 2023, College of Computer Science and Mathematics , University of Mosul.
This is an open access article under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

1. مقدمة Introduction

يلعب اختيار المتغيرات دوراً اساسياً في النمذجة الاحصائية عندما يكون هناك عدد كبير من المتغيرات في النموذج بين مجموعة كبيرة من متغيرات الانحدار الذاتي ذات التخلفات الزمنية السابقة، حيث يهدف اختيار المتغيرات الى التخلص من المتغيرات التي لا تحتوي على معلومات متعلقة بالمتغير ذات التخلف الزمني الحالي وبالتالي تحسين دقة النموذج (1, 2).

إن طريقة Elastic-net تعتبر من الطرائق الجزائية المهمة لاختيار المتغيرات حيث انها توفر دقة تنبؤ جيدة، لأنها تقوم بتقليل وإزالة المعاملات والذي بدوره يؤدي الى تقليل التباين دون زيادة كبيرة في التحيز، وهذا جيد عندما يكون لدينا عدد قليل من المشاهدات وعدد كبير من المتغيرات. بالإضافة الى ان Elastic-net تساعد على تفسير النموذج من خلال ابعاد او التخلص من متغيرات الانحدار الذاتي غير ذات الصلة والتي لا ترتبط بمتغير الاستجابة.

ان الديدان الاسطوانية تستخدم عادة في دراسة علم الوراثة ومن هذه الديدان دودة الربداء الرشيقة CE حيث ان حركة هذه الدودة مؤشر مفيد لفهم علم الوراثة لهذا النوع من الديدان. تم الحصول على بيانات لحركة هذه الدودة من أرشيف تصنيف السلاسل الزمنية[1]. البيانات هي عبارة عن أعداد تمثل حركة الدودة على لوح (agar plate)يحتوي على طعام بكتيري. تم تحديد هذه الحركة عن طريق التتبع باستخدام مقاطع فيديو يتم تسجيل طول الفيديو كإطارات ووقت كل حركة يتم تمثيلها بنقطة بداية ونقطة نهاية والوقت الذي تم قضاءه لعمل هذه الحركة حتى الحركة التي تليها على شكل مجموعة قيم سلسلة زمنية. حيث تمثل البيانات 6 ابعاد ل 5 انواع من جينات دودة الربداء الرشيقة (CE) وكل دودة تمثل سلسلة زمنية تحتوي على (17984) مشاهدة. اذ تمثل البيانات 6 ابعاد ل 5 سلالات strain من دودة الربداء الرشيقة (CE) وكل دودة تمثل سلسلة زمنية تحتوي على (17984) مشاهدة (3).

يقلل اختيار المتغيرات من ابعاد البيانات عن طريق تحديد مجوعة جزئية فقط من متغيرات الانحدار الذاتي لبناء النموذج. تبحث طريقة اختيار المتغيرات عن مجموعة فرعية من متغيرات الانحدار الذاتي والتي من خلالها نقوم بقياس التأثير على متغير الاستجابة. تتمثل الفوائد الرئيسية لاختيار المتغيرات في تحسين التنبؤ وتوفير تنبؤات أسرع وأكثر فاعلية من حيث التكلفة، وتوفير فهم أفضل لعملية توليد البيانات.

كما استخدمت طريقة الشبكة المرنة للتنبؤ بوقت التخزين عند درجة حرارة −20 درجة مئوية بناء على الملف التعريف الأيضي للبلازما واختيار وترتيب المستقلبات ذات التغيرات الزمنية العالية كما في (4). تم استخدام طريقة الشبكة المرنة وهي احدى طرائق التعلم الالي لتحليل البيانات الضخمة حيث تم استخدام هذه الطريقة لاختيار المتغيرات في السلاسل الزمنية والتنبؤ بالعائد الزائد في السوق الامريكية كما في (5). تم استخدام طريقة الشبكة المرنة لاختيار المتغيرات ودمجها مع معاملات الانحدار الجزئي للتنبؤ بالسلاسل الزمنية حيث يحل هذا النموذج عدم دقة اختيار المتغيرات وتقدير المعلمة كما في (6).

2. المواد والطرق

نموذج الانحدار الذاتي Auto Regressive (AR) Model

تعرف السلسلة الزمنية Time Series على أنها مجموعة من المشاهدات المتكونة بشكل متتابع وبترتيب زمني معين، سمتها الأساسية في عدم استقلاليتها، أي أنها مرتبطة زمنياً وتعتمد كل مشاهدة في السلسلة على سابقاتها (7, 8). يمكن استخدام الانحدار الذاتي للتعبير عن قيمة السلسلة الزمنية الحالية باستخدام دالة الانحدار الخطي لقيم السلاسل الزمنية السابقة بشكل عام يمكن كتابة الانحدار الذاتي من الرتبة p وكما في المعادلات ادناه.

$$\phi(B)x_t = a_t \Rightarrow (1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p)x_t = a_t \Rightarrow x_t = \phi_1 x_{t-1}+\phi_2 x_{t-2}+\cdots+\phi_p x_{t-p}+a_t \quad (1)$$

ان معلمات الشبكة المرنة β في المعادلة أعلاه قد تم ترميزها بالرمز ϕ_k وذلك تماشياً مع معظم المصادر العلمية الخاصة بالسلاسل الزمنية وفي هذه الحالة فأن $\beta = \phi$ لجميع معلمات متغيرات الانحدار الذاتي. وان a_t يمثل الخطأ العشوائي او (white noise) بمتوسط صفر وتابين.

طريقة الشبكة المرنة Elastic net method

في عام 2005 اقترح العالمان (Zou, H. and Hastie) طريقة elastic net وهي طريقة جديدة عبارة عن مزيج من حد الجزاء (L1−norm) الموجود في طريقة elastic−net والمقترح من قبل العالم (Tibshirani 1996) مع حد الجزاء في انحدار الحرف L2−norm والمقترح من قبل Hoerl and Kennard (1970) . حيث ان هذه تؤدي الى تقليل او انكماش المتغيرات وكذلك اختيار المتغيرات. ان حد الجزاء L1−norm من elastic net يقلل من عدد المتغيرات بتقليص قيم معاملات الانحدار الى الصفر. الجزء L2−norm من elastic net يتعامل(deal) مع الارتباط العالي بين متغيرات التوقع او المتغيرات التفسيرية (9, 10) ان صيغة elastic net هي كالاتي:

$$\hat{\beta}^{OLS} = \arg\min_{\beta} \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij} - \beta_j)^2 \quad (2)$$

$$\hat{\beta}^{ELNET} = \arg\min_{\beta} \left[\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij} - \beta_j)^2 + \lambda_1 \sum_{j=1}^{p}|\beta_j| + \lambda_2 \sum_{j=1}^{p}|\beta_j^2| \right] \quad (3)$$

[1] تم الحصول على البيانات من أرشيف تصنيف السلاسل الزمنية UEA&UCR. إذ تم سحب البيانات من الموقع بتاريخ 2022/9/21
http://www.timeseriesclassification.com/description.php?Dataset=EigenWorms

حيث ان $\|\beta\|_2^2 = \sum_{j=1}^{p} \beta_j^2$ يمثل L2−norm مربع المتجه β وان $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$ يمثل L1−norm للمتجه β . وان λ_1, λ_2 تمثل معلمات الضبط التي تتحكم بمدى تقليل معلمات الانحدار وهي قيم غير سالبة $\lambda_1, \lambda_2 \geq 0$ والتي يتم تحديدها تلقائياً باستخدام طريقة التحقق المتقاطع CV . وهناك صيغة أخرى elastic net عندما تكون قيمة α بين (0،1) وكما يلي:

$$\hat{\beta}^{ELNET} = \min_{\beta_0, \beta} \left(\frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \beta_0 - x_i^T \beta \right)^2 + \lambda P_\alpha (\beta) \right) \quad (4)$$

حيث ان

$$P_\alpha (\beta) = \frac{(1-\alpha)}{2} \|\beta\|_2^2 + \alpha \|\beta\|_1 = \sum_{j=1}^{p} \left(\frac{(1-\alpha)}{2} \beta_j^2 + \alpha |\beta_j| \right) \quad (5)$$

عندما تكون قيمة $\alpha = 1$ فأن قيمة Elastic−net تساوي LASSO ، وعندما تتجه قيمة α نحو الصفر فان Elastic−net تقترب من انحدار الحرف. اما بالنسبة لقيم α الأخرى فأن حد الجزاء يقع بين L1−norm للمتجه ومربع L2−norm للمتجه. ومن الممكن تلخيص اهم مميزات طريقة الشبكة المرنة بما يلي.

• تؤدي طريقة الشبكة المرنة الى اختيار المتغيرات وتنظيمها (regularization) في ان واحد.

• تكون تقنية الشبكة المرنة هي الأنسب عندما يكون عدد المتغيرات كبير جداً.

3.تقدير معلمة الضبط Tuning Parameter Estimation

ان تقدير معلمة الضبط λ مهم وذلك لتأثيره بشكل كبير على اداء الطرق الجزائية حيث يلعب دوراً مهماً في اختيار المتغيرات لان قيمته تحدد عدد المتغيرات المختارة في النموذج ومقدار التحيز المفروض على معاملات الانحدار المقدرة (12, 13). ان واحدة من أكثر الطرق المستخدمة على نطاق واسع لتقدير معلمة الضبط هي معيار معلومات بيز ((BIC Bayesian Information Criterion وطريقة التحقق المتقاطع Cross−Validation (CV) method.

التحقق المتقاطع CV هي طريقة لاختيار النموذج من خلال تقسيم البيانات (مرة واحدة على الاقل) فيتم استخدام جزء من البيانات (مجموعة التدريب) لتدريب الخوارزمية ويتم استخدام الجزء المتبقي (مجموعة الاختبار) لتقدير خطأ الخوارزمية واختيار النموذج المقابل لأصغر خطأ مقدر. لذلك يتم استخدام CV لتقييم اداء التنبؤ لنموذج التعلم الالي الاحصائي statistical learning model. تضمن هذه الطريقة ان البيانات المستخدمة لتدريب النموذج مستقلة عن مجموعة بيانات الاختبار التي يتم فيها تقييم اداء التنبؤ. ويقصد بعملية CV K−fold اجراء عملية CV من خلال تقسيم البيانات الى K من المجموعات واستخدام أحدها للاختبار فيما تستخدم المجموعات المتبقية (K−1) كبيانات تدريب. بهذه الطريقة نحصل على عدة تقديرات مختلفة لخطأ التنبؤ واختيار اقل تلك الأخطاء لتحقيق الامثلية. يتم استخدام CV في تحليل البيانات للتحقق من صحة النماذج المنفذة حيث يكون الهدف الرئيسي هو التنبؤ وتقدير اداء التنبؤ لنموذج التعلم الالي الاحصائي. بعبارة اخرى يقيّم CV مدى جودة الآلة الإحصائية (14). ولغرض توظيف هذا الأسلوب (CV) في تقدير معلمة الضبط (λ) سوف يتم حساب معدل خطأ التنبؤ لكل قيمة من القيم المفروضة. يمكن تمثيل هذا الأسلوب رياضياً بالشكل الاتي:

$$k - cv_{(\lambda)} = \frac{1}{k} \sum_{i=1}^{k} (y_i - \hat{y}_{i(\lambda)}^{-k(i)})^2 \quad (6)$$

حيث ان ($\hat{y}_{i(\lambda)}^{-k(i)}$) تمثل متغير الاستجابة المناسب عندما تكون المشاهدة (i) تنتمي الى بيانات التحقيق ما دامت هي القيمة الثابتة لقيمة (λ). وبما ان هناك اكثر من قيمة لمعلمة الضبط ، سوف يتم اختيار افضل قيمة والتي تقابل اصغر معدل لخطأ التنبؤ13. وبالشكل الرياضي التالي:

$$\lambda_{optimal} = \underset{r=1,2,\ldots,R}{\operatorname{argmin}} \, \mathrm{k - CV}_{(\lambda_r)} \quad (7)$$

سيتم استخدام معيار RMSE للمفاضلة ويمكن كتابته بالشكل الاتي (15):

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (e_i)^2} \quad (8)$$

حيث ان e_i تمثل خطأ التنبؤ، وان n تمثل عدد المشاهدات.

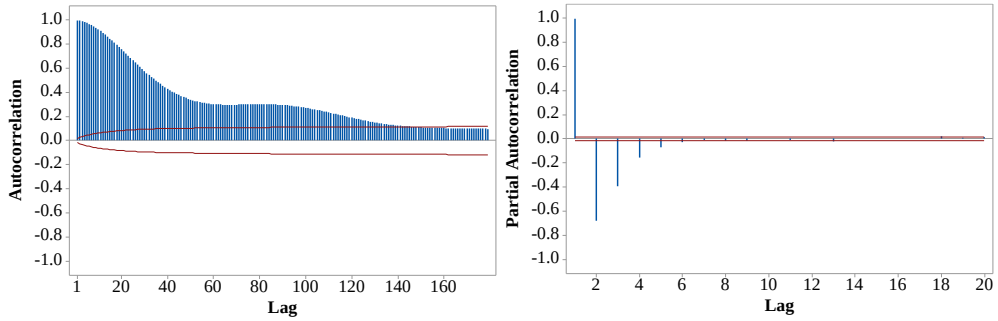
4.نتائج البيانات المستخدمة في الدراسة

ان البيانات المستخدمة في البحث تمثل حركة دودة الربداء الرشيقة (CE) ان هذه الحركة تكون حسب سرعتها يتم تسجيل السرعة بشكل سلبي عندما يتحرك جزء من الجسم تجاه الذيل (على عكس الرأس)،معدل سرعة الدودة هي على الأقل تساوي 5% من الطول لكل ثانية في كل إطار حيث ان الدودة تحافظ على هذه السرعة بشكل مستمر فتقوم الدودة بتعويض التأخر الحاصل بسبب التوقف من خلال زيادة سرعتها عن طريق زيادة حدة زوايا الحركة لحين الوصول الى معدل السرعة المطلوبة حيث تم تسجيل حركة الدودة (الحركة تكون الى الامام او الى الخلف) مـع الوقت كسلسلة زمنية لحركة ممثلة بزوايا ظل للحركة الموجية ،كل مشاهدة لهذه السلسلة الزمنية عبارة عن إطار مسجل (0.5) ثانية من فيديو (2.5) ساعة لحركة CE.في كل ثانية تنتقل الدودة بسرعة (5%) من طولها ويجب ان نحافظ على هذه السرعة بشكل شبه مستمر مع انقطاعات مسموح بها على الأكثر (0.25) ثانية مما قد يولد حركات متناقضة اثناء وبعد التوقف مثل انسحاب الرأس وانقباضات في الجسم وضوضاء في حركة اجزائها المختلفة . 3وبناء على ما تقدم فأنه من الممكن الاستفادة من السلوك الحركي للدودة من خلال علاقة الزاوية مع السرعة لذلك عندما تكون الزاوية حادة تكون السرعة ستكون أكبر وعندما تكون الزاوية منفرجة أي عندما تكون الزاوية موجبة تكون الحركة أبطأ. حيث تمثل البيانات 6 ابعاد ل 5 انواع من جينات دودة الربداء الرشيقة (CE) وكل دودة تمثل سلسلة زمنية تحتوي على (17984) مشاهدة او متغير. تم سحب عينة من سلسلتين زمنيتين (كل سلسلة تمثل حركة دودة مستقلة) من نفس السلالة وهي السلالة المرجعية (N2) وهي السلالة التي تهتم بها معظم الدراسات. تتضمن المكونات الرئيسية للجانب التطبيقي في هذه الدراسة ما يأتي:

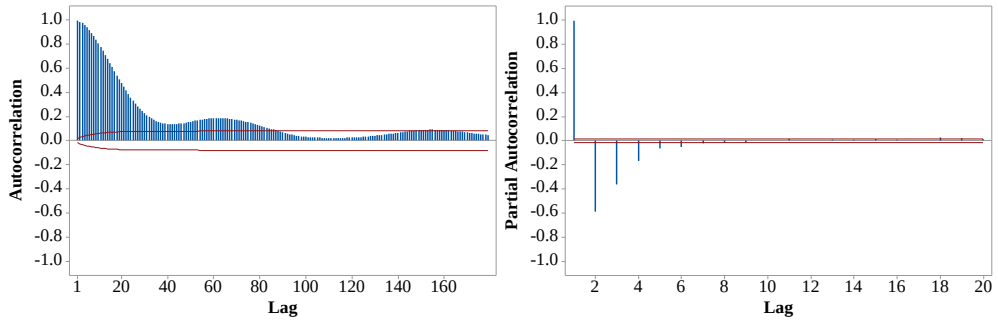
1. تحديد نموذج AR المناسب.
2. حساب ثلاث طرق elastic net بالاعتماد على التخلفات الزمنية لنموذج AR وهي 1000،500،100.
3. حساب قيمة RMSE لنموذج AR و elastic net لتوقعات العينة.
4. مقارنة نتائج الدقة لنموذج AR و elastic net لتحديد النموذج الذي سيقدم نتائج افضل واي معلمات غير صفرية سيتم تضمينها في كل نموذج.

5.نموذج الانحدار الذاتي Auto-Regressive AR Model

بعد اختيار السلسلتين بشكل عشوائي من البيانات قمنا برسم هاتين السلسلتين باستخدام برنامج (Minitab) تم رسم دالتي (ACF) و (PACF) للسلسلتين لتحديد المعلمات المعنوية وادخالها في نموذج AR للحصول على قيم المعلمات المعنوية والتي تمثل عدد من القيم التي سوف نستخدمها في بناء نموذج ال (AR). الشكل ادناه يوضح الدالتين ACF وPACF للعينة الأولى للسلسلة الاصلية.



الشكل رقم 1: يوضح رسم دالتي ACF و PACF للعينة الأولى.



الشكل رقم 2: يوضح رسم دالتي ACF و PACF للعينة الثانية.

نلاحظ من الاشكال السابقة الشكل رقم 11 والشكل رقم 2 يتضح ان أفضل النماذج بغض النظر عن إستقرارية السلاسل الزمنية هي AR(6) لكلا العينتين وذلك بعد ملاحظة معنوية التخلفات الزمنية الستة الأولى لدالة ACF في كلا الشكلين حيث لا حاجة الى استخدام معايير المفاضلة AIC و BIC للمقارنة بين النماذج ,من الرسم يتضح ان الرتبة المناسبة للنموذج هي AR(6).

الجدول التالي يبين سلوك دالة الارتباط الذاتي ACF ودالة الارتباط الجزئي PACF من خلال الرسم

النموذج	دالة الارتباط الذاتي (**ACF**)	دالة الارتباط الجزئي (**PACF**)
AR(**p**)	تقترب من الصفر ببطيء وتساوي الصفر بعد الارتباط الذاتي (**q**)	تساوي الصفر فجأة بعد الفجوة الزمنية (**p**).
MA(**q**)	تساوي الصفر فجأة بعد الفجوة الزمنية (**q**).	تقترب من الصفر ببطيء.
ARIMA(**p,d,q**)	تقترب من الصفر ببطيء وتنقطع الفجوة الزمنية (**q**) بعد أخذ (**d**) من الفروق.	تقترب من الصفر ببطيء وتنقطع الفجوة الزمنية (**p**) بعد أ (**d**) من الفروق.

اما بالنسبة لمقدرات النموذج AR(6) لكلا العينتين فمن غير الممكن الحصول عليها بدقة عالية من خلال البرامج الإحصائية التقليدية ولذلك تم اللجوء الى matlab واستخدام الايعاز ((ar(y,6)) عندما y هو متغير هدف الذي يمثل السلسلة الاصلية.

وعند تطبيق عدة ايعازات تم حساب MSE للنموذج والحصول على ال X وال Y وبعد ذلك حصلنا على مقدر elastic net وتم استخدام طريقة التحقق المتقاطع (cross validation) لتحديد قيم معلمة الضبط. وبعد تطبيق ايعاز طريقة Elastic-net على البيانات ولجميع قيم(AR) المختلفة والتي تحتوي على عدة متغيرات وهي (p=100, p=500,p=1000) تم الحصول على النتائج والجدول التالي يوضح قيم المعلمات المعنوية للعينتين الأولى والثانية عندما تكون قيم الانحدار الذاتي AR(6).

الجدول رقم 1: يبين المعلمات المعنوية لنموذج الانحدار الذاتي AR(6) للعينة الأولى والثانية.

p	AR(6) sample 1	AR(6) sample 2
1	− 1.338	−1.294
2	0.0297	0.02602
3	0.1884	0.1439
4	0.0666	0.08593
5	0.0283	0.0036
6	0.0281	0.05235
MSE	0.00837	0.021896

6.طريقة الشبكة المرنة Elastic-net method

ان الاطار العام لخوارزمية تنفيذ Elastic-net method يتضمن تنفيذ عدة خطوات متسلسلة وكما يلي.

1- استخدام متغيرات الانحدار الذاتي بعدد p ممثلة بالتخلفات الزمنية اعتمادا على (الجدول رقم 1) لتحديد متغيرات الادخال لنموذج Elastic-net.

2- ادخال متغير الهدف والذي هو السلسلة الزمنية الاصلية.

3- بناء افضل نموذج elastic-net باستخدام بيانات السلاسل الزمنية لمتغيرات الادخال والهدف وكذلك بتحديد استخدام أسلوب التحقق المتقاطع Cross validation (CV) مع K-Fold الافتراضية عندما K=10. باستخدام الايعاز في برنامج Matlab وذلك عندما يكون عدد التخلفات 100 و 500 و 1000 على التوالي.

4- استخدام النموذج في الخطوة السابقة للتنبؤ Forecasting بالبيانات بعد تحديد القيم الافتراضية للمعلمات.

5- حساب دقة طريقة Elastic-net للتنبؤ عن طريق مقياس RMSE.

وكانت قيم المعلمات غير الصفرية non-zero ودقة التنبؤ من خلال RMSE.

الجدول رقم 2: يوضح معلمات العينة الاولى والتي تم اختيارها باستخدام طريقة elastic net لقيم (p) المختلفة.

الجدول رقم 2: يبين المعلمات التي تم اختيارها باستخدام الشبكة المرنة للعينة الأولى.

i	P=100	i	P=500	i	P=1000
x_1	1.098675	x_1	1.099693	x_1	1.096554
x_2	0.091877	x_2	0.090119	x_2	0.091966
x_4	−0.05436	x_4	−0.05438	x_4	−0.05258
x_5	− 0.063310	x_5	−0.06364	x_5	−0.06306
x_6	−0.04509	x_6	−0.04546	x_6	−0.04484
x_7	−0.01914	x_7	−0.01876	x_7	−0.01874
x_8	−0.00587	x_8	−0.00549	x_8	−0.00635
⋮	⋮	⋮	⋮	⋮	⋮
x_{100}	−0.00005	x_{488}	0.000102	x_{994}	−0.00011
RMSE	0.100011	RMSE	0.099987	RMSE	0.100245

الجدول رقم 3 يوضح قيم المعلمات للعينة الثانية والتي تم اختيارها باستخدام طريقة Elastic−net لقيم (p) المختلفة.

الجدول رقم 3: يبين المعلمات التي تم اختيارها باستخدام الشبكة المرنة للعينة الثانية.

i	**P=100**	**i**	**P=500**	**i**	**P=1000**
x_1	1.2215065	x_1	1.2192381	x_1	1.2159731
x_3	−0.0703729	x_3	−0.0694169	x_3	−0.0673852
x_4	−0.0874861	x_4	−0.0872702	x_4	−0.0868157
x_5	−0.0246816	x_5	−0.0237702	x_5	−0.0234088
x_6	−0.0327317	x_6	−0.0332219	x_6	−0.0343359
x_7	−0.0030242	x_7	−0.0019683	x_7	−0.0013307
x_9	−0.0089524	x_9	−0.0097497	x_9	−0.0081757
⋮		⋮		⋮	
x_{100}	0.000455	x_{493}	0.0003212	x_{1000}	−0.0000471
RMSE	0.1502367	RMSE	0.1494121	RMSE	0.14954713

مناقشة

في الجدول رقم 2 وعند استخدام طريقة Elastic−net للعينة الأولى عندما تكون قيمة (p=100) نلاحظ ان المتغيرات التي تم اختيارها هي (x1,x2,x4,…,x100) التي تعد هي المتغيرات المهمة حسب هذه الطريقة كما استبعدت المتغيرات المتبقية. ويمكن مقارنتها مع متغيرات AR(6) المعنوية كما في الجدول رقم 1 وهي (x1,x2,x3,x4,x5,x6). من خلال المقارنة يتبين ان طريقة Elastic−net قد استبعدت المتغير (x3) المعنوي في الجدول رقم 1 كذلك اختارت المتغيرات (x7,x8,…,x100) كمتغيرات مهمة في حين انها لم تكن معنوية في AR(6) في الجدول رقم 1.

في الجدول رقم 2 وعند استخدام طريقة Elastic−net للعينة الأولى عندما تكون قيمة (p=500) نلاحظ ان المتغيرات التي تم اختيارها هي (x1,x2,x4,…,x488) التي تعد هي المتغيرات المهمة حسب هذه الطريقة كما استبعدت المتغيرات المتبقية. ويمكن مقارنتها مع متغيرات AR(6) المعنوية كما في الجدول رقم 1 وهي (x1,x2,x3,x4,x5,x6). من خلال المقارنة يتبين ان طريقة Elastic−net قد استبعدت المتغير (x3) المعنوي في الجدول رقم 1 كذلك اختارت المتغيرات (x7,x8,…, ,x488) كمتغيرات مهمة في حين انها لم تكن معنوية في AR(6) في الجدول رقم 1.

في الجدول رقم 2 وعند استخدام طريقة Elastic–net للعينة الأولى عندما تكون قيمة (p=1000) نلاحظ ان القيم التي تم اختيارها هي (,…, x1,x2,x4,x5 x994,) التي تعد هي المتغيرات المهمة حسب هذه الطريقة كما استبعدت المتغيرات المتبقية. ويمكن مقارنتها مع متغيرات AR(6) المعنوية كما في الجدول رقم 1 وهي (x1,x2,x3,x4,x5,x6). من خلال المقارنة يتبين ان طريقة Elastic net قد استبعدت المتغير (x3) المعنوي في الجدول رقم 1 كذلك اختارت المتغيرات (x994, …,x7,x8) كمتغيرات مهمة في حين انها لم تكن معنوية في AR(6) في الجدول رقم 1.

في الجدول رقم 3 وعند استخدام طريقة Elastic–net للعينة الثانية عندما تكون قيمة (p=100) نلاحظ ان القيم التي تم اختيارها هي (x100,…,x9,x7,x6,x5,x4,x3,x1) التي تعد هي المتغيرات المهمة حسب هذه الطريقة كما استبعدت المتغيرات المتبقية. ويمكن مقارنتها مع متغيرات AR(6) المعنوية كما في الجدول رقم 1 وهي (x1,x2,x3,x4,x5,x6). من خلال المقارنة يتبين ان طريقة Elastic–net قد استبعدت المتغير (x2) المعنوي في الجدول رقم 1 كذلك اختارت المتغيرات (x84,…,x9,x7) كمتغيرات مهمة في حين انها لم تكن معنوية في AR(6) في الجدول رقم 1.

في الجدول رقم 3 وعند استخدام طريقة Elastic–net للعينة الثانية عندما كانت قيمة (p=500) نلاحظ ان القيم التي تم اختيارها هي (x484,…,x9,x7,x5,x4,x3,x1) التي تعد هي المتغيرات المهمة حسب هذه الطريقة كما استبعدت المتغيرات المتبقية. ويمكن مقارنتها مع متغيرات AR(6) المعنوية كما في الجدول رقم 1 وهي (x1,x2,x3,x4,x5,x6). من خلال المقارنة يتبين ان طريقة Elastic–net قد استبعدت المتغير (x2) المعنوي في الجدول رقم 1 كذلك اختارت المتغيرات (x484,…,x9,x7) كمتغيرات مهمة في حين انها لم تكن معنوية في AR(6) في الجدول رقم 1.

في الجدول رقم 3 وعند استخدام طريقة Elastic–net للعينة الثانية عندما تكون قيمة (p=1000) نلاحظ ان القيم التي تم اختيارها هي (x1000,…,x7,,x6,x5,x4,x3,x1) التي تعد هي المتغيرات المهمة حسب هذه الطريقة كما استبعدت المتغيرات المتبقية. ويمكن مقارنتها مع متغيرات AR(6) المعنوية كما في الجدول رقم 1 وهي (x1,x2,x3,x4,x5,x6). من خلال المقارنة يتبين ان طريقة Elastic–net قد استبعدت المتغير (x2) المعنوي في الجدول رقم 1 كذلك اختارت المتغيرات (x1000,…,x9,x7) كمتغيرات مهمة في حين انها لم تكن معنوية في AR(6) في الجدول رقم 1.

7.الاستنتاجات Conclusion

في هذه الدراسة تم استخدام الأسلوب Elastic–net كاسلوب مقترح لتحسين اختيار المعلمات غير الصفرية المثلى من بين عدد كبير من المعلمات والتي ربما من الصعب ان يستوعبها أسلوب تقليدي مثل AR عندما تكون البيانات للسلسلة الزمنية لدودة الرداء الرشيقة CE مع نتائج تنبؤية قريبة جداً ودقيقة جداً لكلا الاسلوبين Elastic–net و AR. تم استخدام عينتين من البيانات وأوضحت النتائج اتفاق كبير على المعلمات المهمة لأسلوب Elastic–net مع AR كأسلوب تقليدي مع تفوق واضح للأسلوب الهجين Elastic–AR في نمذجة السلاسل الزمنية والتنبؤ بها. تم استخدام معيار RMSE لبيان جودة التنبؤ والذي عكس تقارب كبير في دقة التنبؤ رغم اختلاف المعلمات المختارة من قبل كلا الاسلوبين المقترح والتقليدي. ان أسلوب Elastic–AR الهجين قد عكس افضلية مطلقة وحقيقية مع اعداد مختلفة من متغيرات الانحدار الذاتي حيث اختار أمثل المتغيرات التي تعبر فعلياً عن بيانات الدراسة. من الممكن استنتاج إمكانية استخدام Elastic–AR الهجين كطريقة مثلى مع بيانات السلاسل الزمنية عالية الابعاد لبيانات أحد أنواع الديدان الاسطوانية والتي تحمل بصفتها عدد كبير جدا من المشاهدات ممثلة كمتغيرات انحدار ذاتي. ومن خلال النتائج التي تم الحصول عليها فقد كانت المعلمات المختارة في نموذج الانحدار الذاتي AR متطابقة مع النموذج الهجين Elastic–AR الى حد كبير مع تفوق في نتائج الأسلوب الهجين ودقة عالية. ولذلك فمن الممكن استنتاج امكانية استخدام الأسلوب الهجين Elastic–AR المقترح للحصول على أفضل نموذج للسلاسل الزمنية عالية الابعاد.

References

1. Zhang, Y., R. Li, and C.-L. Tsai. Regularization parameter selections via generalized information criterion. Journal of the American Statistical Association 2010; 105: 312-323. DOI.
2. Konrath, S., L. Fahrmeir, and T. Kneib. Bayesian smoothing, shrinkage and variable selection in hazard regression. Robustness and Complex Data Structures: Festschrift in Honour of Ursula Gather 2013: 149-170. DOI.
3. Yemini, E., et al. A database of c. elegans behavioral phenotypes. Nature Methods; 10: 877. DOI.
4. Gonzales, G.B. and S.J.S.R. De Saeger. Elastic net regularized regression for time-series analysis of plasma metabolome stability under sub-optimal freezing condition. 2018; 8: 3659. DOI.
5. Rapach, D.E., G.J.M.l.f.a.m.N.d. Zhou, and f. applications. Time-series and cross-sectional stock return forecasting: New machine learning methods. 2020: 1-33. DOI.
6. Xing, Y., D. Li, and C.J.A.S.C. Li. Time series prediction via elastic net regularization integrating partial autocorrelation. 2022; 129: 109640. DOI.
7. Brockwell, P.J. and R.A. Davis. Time series: theory and methods. ed.: Springer science & business media. 2009.
8. Liu, L.-M. Time Series Analysis and Forecasting. 2nd ed. Illinois, USA: Scientific Computing Associates Corp. 2006.
9. Zou, H. and T. Hastie. Regularization and variable selection via the elastic net (vol B 67, pg 301, 2005). Journal of the Royal Statistical Society Series B-Statistical Methodology 2005; 67: 768-768. DOI.
10. Al-Jawarneh, A.S. and M. Ismail. Elastic-Net Regression based on Empirical Mode Decomposition for Multivariate Predictors. Pertanika Journal of Science & Technology 2021; 29. DOI.
11. Masselot, P., et al. EMD-regression for modelling multi-scale relationships, and application to weather-related cardiovascular mortality. Science of The Total Environment 2018; 612: 1018-1029. DOI.
12. Androulakis, E., C. Koukouvinos, and K. Mylona. Tuning parameter estimation in penalized least squares methodology. Communications in Statistics-Simulation and Computation 2011; 40: 1444-1457. DOI.
13. Li, Y., L. Dicker, and S.D. Zhao. The Dantzig selector for censored linear regression models. Statistica Sinica 2014; 24: 251. DOI.
14. Zhang, Y. and Y. Yang. Cross-validation for selecting a model selection procedure. Journal of Econometrics 2015-95 :187 ; .112DOI.
15. Hyndman, R.J. and A.B. Koehler. Another look at measures of forecast accuracy. International Journal of Forecasting 2006; 22: 679-688. DOI: http://dx.doi.org/10.1016/j.ijforecast.2006.03.001.

Using Elastic-Net for High Dimensional Time Variables Selection of Autoregressive Model Series of Caenorhabditis Elegans Motion

Mohammed Khames Rasheed Osamah Basheer Shukur
Department of Statistics and Informatics, College of Computer Science and Mathematics, University of Mosul, Mosul, Iraq

Abstract : The process of selecting statistical variables that contain information related to the effect on the dependent variable has a fundamental role in accurate statistical modeling. In time series and when there are a large number of an autoregressive (AR) variables in the model, selecting the most effective AR variables among a large set of autoregressive variables (high dimensional) with prior time lags is important for more accurate results. The elastic net method is one of the methods used for selecting the best model and conducting a joint estimation of the linear models, which contributes to the selection of the actually influencing variables and ignoring others among a very large set (high dimensional) of autoregressive variables. In this study, the elastic network method will be used to select and estimate the autoregressive parameters in the time series model. Caenorhabditis elegans (CE) will use time series data for the movement of Caenorhabditis elegans, represented by tangent angles of the wave motion. The univariate time-series model of CE movement was selected via the elastic network method and the hybrid (Elastic-AR) autoregressive model after multi-processes of selecting autoregressive variables. According to the results, the selected parameters in the AR model matched the Elastic-AR hybrid model, with clear superiority in the results of the hybrid method and with high accuracy. Therefore, it is possible to conclude the possibility of using the proposed hybrid method to obtain the best model for the high-dimensional time series dataset with the least number and the most influential of variables, which reduces effort and costs and increases the accuracy of these models**.**

Keywords: selecting autoregressive variables, time series high dimensional, autoregressive model AR, Elastic-Net, Elastic-AR hybrid model.