تحليل الارتباط القويم اللاخطي باستخدام بعض دوال كيرنل

علي جواد كاظم•

الخلاصة :

يعد تحليل الارتباط القويم من تحليلات متعددة المتغيرات التقليديه حيث يعمل على ايجاد تحويل خطي لزوج من المتغيرات المتعددة بحيث يحقق هذا التحويل تعظيم معامل الارتباط ويجعله اعظم ما يمكن ومن وجهة النظر النظريه فان هذا التحويل يعمل على تعظيم المعلومات المستخلصه عن المعالم .

اذا كانت العلاقة بين المتغيرات غير خطيه فان تحليل الارتباط القويم لا يستطيع استخلاص معلومات مفيده عن تلك المعالم . في هذا البحث تم استخدام ما يدعى بخدعة كيرنل (Kernel trick) والتي تعمل على جعل البيانات ملائمه للتحليل وتم استخدام دوال مختلفة من نوع كيرنل منها دالة (Gaussian Kernel) ودالة (Uniform Kernel) ودالة(Triangle Kernel) .

المقدمة :

يفترض هذا التحليل وجود زوج من المتغيرات المتعدده $X \in R^{n_x}, Y \in R^{n_y}$ ويعمل على ايجاد زوج من التحويلات الخطيه التي تجعل معامل الارتباط بين المعالم المستخلصه اعظم ما يمكن وبافتراض ان X,Y لكل منهما متوسط مقداره صفر وتباين مقداره واحد فان التحويلات تكون كما ياتي[3]:

$$U = \langle a, x \rangle, V = \langle b, y \rangle$$

حيث ان :

$\langle a, x \rangle, \langle b, y \rangle$ تمثل حاصل الضرب الداخلي (inner product) بين a,x و b,y على التوالي.

والهدف هنا هو ايجاد التحويل المناسب الى a,b والذي يجعل معامل الارتباط اعظم ما يمكن. **وبمعنى اخر اذا كان لدينا متغير عشوائي يتكون من p من الأبعاد هو X ومتغير يتكون من q من الأبعاد هو Y ، هدفنا هو الحصول على تركيبات خطيه $a^T X$ ، $b^T Y$ من المتغيرات الاصلية تمتلك اعظم ارتباط ، ويمكن التعبير عن ذلك بصيغة رياضية بالشكل التالي [8] :**

• مدرس مساعد قسم الاحصاء/جامعة القادسيه

$$\rho = MAX\left|corr(a^T X, b^T Y)\right| \ldots\ldots (1)$$

ان المتغيرات الناتجة $U=a^TX$ ، $V=b^TY$ تدعى المتغيرات القويمة وان الارتباط القـويم الأول ρ يعرف كقيمه مطلقه للارتباط بين مجموعتين من المتغيرات القويمة وكما في العلاقـة رقم (1) ، ان المتغير القويم ذو الرتبة k ، $1 < k<Min(p,q)$ سيكون غير مرتبط مع كـل المتغيرات القويمة ذات الرتب الدنيا .

ويمكن ايجاد a^T, b^T كمتجه مميز مرافق لقيم مميزه عظمى [3] .

افرض ان Σ تمثل مصفوفة التبـاين المـشترك للمجتمـع للمتغيـر العـشوائي Z حيـث ان $Z = (X^T, Y^T)^T$ ويمكن وضع Σ بالشكل التالي :

$$\Sigma = \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix}$$

وقد نستخدم مصفوفة الارتباطات R

$$R = \begin{bmatrix} R_{XX} & R_{XY} \\ R_{YX} & R_{YY} \end{bmatrix}$$

حيث يتم حساب المتجه المميز b^T من العلاقه التاليه[7] :

$$(M - \lambda I)b = 0 \ldots\ldots (2)$$

بفرض ان

$$M = \Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$$

عند استخدام مصفوفة التباين المشترك .

او

$$M = R_{YY}^{-1}R_{YX}R_{XX}^{-1}R_{XY}$$

عند استخدام مصفوفة الارتباطات .

ان قيم المتجه المميز b^T والذي يرتبط بكل قيمه من القيم المميزه (λ) **الناتجه** من المعادله (2) تمثل اوزانا للمجموعه Y من المتغيرات .

وبعد حساب b^T **يمكن ان نحدد المتجه** a^T **المقابل للمجموعه X من المتغيرات من خلال العلاقه التاليه**[7]:

$$a^T = \frac{1}{\sqrt{\lambda}}R_{XX}^{-1}R_{XY}d^T$$

يستخدم تحليل الارتباط القويم عندما تكون x,y متغيرات تتبع التوزيع الطبيعي المشترك وحتى عندما يكون هذا الافتراض غير متحقق يمكن استخدامه في بعض الحالات وعندما يكون هدفنا هو الحصول على انحدار تلك المتغيرات فاننا نرغب في الحصول على قيم كبيره لمعاملات الارتباط . ولكن احيانا قد نحصل على قيم صغيره لتلك المعاملات وقد يكون ذلك عائدا الى احد السببين التاليين(3) :

1- عدم وجود أي علاقه بين X,Y .

2- هنالك علاقه غير خطيه قويه بين X,Y .

في الحاله الاولى لا يمكن الحصول على افضل مما تم الحصول عليه ولا يمكن اجراء أي تحسينات تذكر .اما في الحاله الثانيه فيمكن الحصول على العلاقه باستخدام بعض الطرق منها طريقة كيرنل (Kernel Method) في تحليل الارتباط القويم .

طريقة كيرنل في تحليل الارتباط القويم Kernel CCA(3)

في هذه الطريقه يتم تحويل X,Y في فضاء هلبرت* (Hilbert Space) حيث :

$$\phi_x(x) \in H_x, \phi_y(y) \in H_y$$

وباخذ حاصل الضرب الداخلي (inner product) للمعالم في فضاء هلبرت

$$a \in H_x, b \in H_y$$

نحصل على ما ياتي :

$$U = \langle a, \phi(x) \rangle, V = \langle b, \phi(y) \rangle$$

وبفرض وجود عينات $\{(x_i, y_i)\}_{i=1}^{N}$ يمكن ايجاد a,b وذلك عن طريق حل علاقة لاكرانج التاليه :

$$L_0 = E[(U - E[U])(V - E[V])] - \frac{\lambda_1}{2} E\left[(U - E[U])^2\right] - \frac{\lambda_2}{2} E\left[(V - E[V])^2\right]$$

ان علاقة لاكرانج لا تعمل بشكل جيد عندما تكون ابعاد فضاء هلبرت كبيره ولهذا يجب اضافة حد انتظام تربيعي (Quadratic Regularization Term) وذلك من اجل تحسين عمل علاقة لاكرانج وكما ياتي :

$$L = L_0 + \frac{\eta}{2}\left(\|a\|^2 + \|b\|^2\right)$$

حيث ان η تمثل ثابت انتظام (Regularization Constant).

ان معدل U يمكن حسابه من الصيغه التاليه :

$$E[U]=\frac{1}{N}\sum_{i}\langle a,\phi_x(x_i)\rangle$$

- ان فضاء هلبرت الابتدائي الكامل يدعى فضاء هلبرت. بعباره اخرى اذا كان X فضاء متجهات على الحقل F مع الضرب الداخلي $\langle,\rangle$ فان X يكون فضاء هلبرت اذا كان الفضاء المتري المتولد بواسطة المعيار $\|x\|^2=\langle x,x\rangle$ فضاء كاملا [1].

ويمكن توضيح فضاء هلبرت الابتدائي او ما يسمى احيانا بفضاء الضرب الداخلي من خلال التعريف التالي:

تعريف: ليكن X فضاء متجهات على الحقل F . الداله $\langle,\rangle: X\times X\to F$ تسمى بدالة الضرب الداخلي على X ، اذا تحققت البديهيات التاليه [1]:

1- $\langle x,x\rangle\geq 0$ لكل $x\in X$.

2- $\langle x,x\rangle=0$ اذا وفقط اذا كان $x=X$.

3- $\langle x,y\rangle=\overline{\langle y,x\rangle}$ حيث $\overline{\langle x,y\rangle}$ تعني العدد المرافق للعدد $\langle x,y\rangle$.

4- $\langle \alpha x+\beta y,z\rangle=\alpha\langle x,z\rangle+\beta\langle y,z\rangle$ لكل $x,y,z\in X$ لكل $\alpha\beta\in F$.

فضاء هلبرت الابتدائي (pre-Hilbert Space) او فضاء الضرب الداخلي هو الثنائي $(X,\langle\ \rangle)$ حيث X فضاء متجهات على الحقل F ,$\langle,\rangle$ضرب داخلي على X .

وكذلك معدل UV يمكن حسابه بالشكل التالي:

$$E[UV]=\frac{1}{N}\sum_{i,j}\langle a,\phi_x(x_i)\rangle\langle b,\phi_y\langle y_i\rangle\rangle$$

وباشتقاق L بالنسبه الى a ومساواة المشتقه الى الصفر نحصل على ما ياتي:

$$a=\sum_{i}\alpha_i\phi_x(x_i)$$

حيث ان α_i تمثل ثابت (Scalar).

ونتيجه للخطوه السابقه فان :

$$U = \sum_{i} \alpha_i \langle \phi_x(x_i), \phi_x(x) \rangle$$

والتي يمكن حسابها عن طريق حاصل الضرب الداخلي .ان ما يسمى بخدعة كيرنل (Kernel Trick) يتلخص باستخدام دالة كيرنل $K_x(X_1, X_2)$ بدلا من حاصل الضرب الداخلي بين $\phi_x(X_1), \phi_x(X_2)$ وبذلك لسنا بحاجه الى صيغة $\phi_x(X)$ نحتاج فقط الى تحديد K_x والتي تكون معرفه موجبه متماثله (2). ويمكن اعادة كتابة L من حسب وجهة نظر طريقة كيرنل كما ياتي :

افرض ان

$$\beta = (\beta_1, \beta_2, \dots, \beta_N)^T \text{ ، } \alpha = (\alpha_1, \alpha_2, \dots, \alpha_N)^T$$

$$(K_x)_{ij} = K_x(x_i, x_j) \text{ ، } (K_y)_{ij} = K_y(y_i, y_j)$$

فان :

$$L = \alpha^T M \beta - \frac{\lambda_1}{2} \alpha^T L \alpha - \frac{\lambda_2}{2} \beta^T N \beta$$

حيث ان :

$$M = \frac{1}{N} K_x^T J K_y$$

$$L = \frac{1}{N} K_x^T J K_x + \eta_1 K_x$$

$$N = \frac{1}{N} K_y^T J K_y + \eta_2 K_y$$

$$J = I - \frac{1}{N} 1 1^T$$

$$1 = (1,1,\dots,1)^T$$

$$\eta_1 = \frac{\eta}{\lambda_1}, \eta_2 = \frac{\eta}{\lambda_2}$$

واذا كانت $\eta > 0$ فان كل من N, L سيكون مصفوفة معرفة موجبه (Positive definite matrix) ويمكن اثبات ان $\lambda_1 = \lambda_2 = \lambda$ ونتيجة لذلك يكون لدينا ما يدعى بمشكلة قيمة مميزه عامه (Generalized eigenvalue problem) لـ α, β وكما ياتي:

$$M\beta = \lambda L\alpha$$

$$M^T\alpha = \lambda N\beta$$

والتي يمكن حلها كمشكلة قيمة مميزه عامة[2].

المحاكاة:

لقد تم استخدام المحاكاة لغرض إجراء المقارنة بين مقدرات الارتباط القويم المحسوبة وفق الطريقه التقليديه ومقدرات الارتباط القويم المحسوبه وفق طريقة كيرنل باستخدام ثلاث دوال من نوع كيرنل هي دالة (Gaussian Kernal) ودالة (Uniform Kernel) ودالة (Triangle Kernel) وذلك بهدف معرفة أفضلية المقدرات وذلك بالاعتماد على كون الارتباط القويم يمثل اكبر الارتباطات للحكم على المقدر الأفضل وتم كتابة برنامج باستخدام لغة (V.B) خاص بالتجربة لتحقيق هذا الهدف حيث تم تكرارها (1000) مرة لغرض الوصول إلى نتائج مُقنعة ويمكن أدراج خطوات أجراء المحاكاة وكالتالي :

1ـ نولد قيم للمتغيرات (x_1, x_2) لتكوين تركيبة خطية هي UC_i والتي يتم استخدامها في حساب الارتباط القويم وفق الطريقة التقليدية وتركيبه خطيه أخرى هي UK_i والتي يتم استخدامها في حساب الارتباط القويم وفق طريقة كيرنل وبأحجام عينات (n=10,20,30,40,50,60,70,80,90) أي لكل متغير من متغيرات x نولد n من المشاهدات المذكورة حيث إن :

$$UC_i = a_1x_1 + a_2x_2$$

$$UK_i = a_1K(x_1) + a_2K(x_2)$$

وان صيغ دوال كيرنل المستخدمة هي الصيغ التالية :

a) $$K(x) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2})$$

وهذه الدالة تدعى دالة (Gaussian Kernal).

b) $K(x) = 1/2, |x| \le 1$

وهذه الدالة تدعى دالة(Uniform Kernel).

c) $K(x) = (1 - |x|), |x| \le 1$

وهذه الداله تدعى دالة (Triangle Kernel).

وان :

$$X = (\theta)\sin 3\theta + \varepsilon_1$$

حيث يتم توليد قيم θ على اساس انها تتبع التوزيع المنتظم المستمرعلى الفتره $[-\pi, \pi]$.

ε_1 : يمثل حد الخطا العشوائي ويتم توليده على أساس انه يتبع التوزيع الطبيعي القياسي في الحالة الأولى ويتبع التوزيع الاسي بالمعلمة λ=1/3 في الحالة الثانية أما في الحالة الثالثة فيتم توليده على أساس انه يتبع توزيع مربع كاي بالمعلمة n.

2ـ يتم توليد متغيرات (y_1, y_2) لتكوين تركيبة خطية هي VC_i والتي يتم استخدامها في حساب الارتباط القويم وفق الطريقة التقليدية وتركيبه خطيه أخرى هي VK_i والتي يتم استخدامها في حساب الارتباط القويم وفق طريقة كيرنل وبأحجام عينات (n=10,20,30,40,50,60,70,80,90) أي لكل متغير من متغيرات y نولد n من المشاهدات المذكورة حيث إن :

$$VC_i = b_1 y_1 + b_2 y_2$$

$$VK_i = b_1 K(y_1) + b_2 K(y_2)$$

وان $K(y)$ تأخذ نفس الصيغ السابقة وان الصيغة التي يتم الحصول على قيم y منها هي الصيغة التالية:

$$Y = e^{\theta/4}(\cos 2\theta)(\sin 2\theta) + \varepsilon_2$$

حيث يتم توليد قيم θ على اساس انها تتبع التوزيع المنتظم المستمر على الفتره $[-\pi, \pi]$.

ونولد قيم للمتغير ε_2 مرة بالتوزيع الطبيعي القياسي ومرة أخرى بالتوزيع الاسي بالمعلمة $\lambda=1/3$ ومره ثالثه بتوزيع مربع كآي بدرجة حرية (n) .

3ـ يتم حساب معاملات الارتباط بين UC_i و VC_i حيث ان اكبرها يمثل معامل الارتباط القويم المحسوب وفق الطريقه التقليديه ويتم حساب معاملات الارتباط بين UK_i و VK_i حيث ان اكبرها يمثل معامل الارتباط القويم المحسوب وفق طريقة كيرنل حسب الداله المستخدمه أي يتم حساب معامل الارتباط القويم نوع كيرنل لدوال كيرنل الثلاث .

4ـ يتم المقارنه بين معاملات الارتباط القويم المحسوبه وفق طريقه كيرنل وبالدوال الثلاث ومعامل الارتباط القويم المحسوب وفق الطريقه التقليديه .

تحليل النتائج :

بعد أجراء تجربة المحاكاة تم الحصول على النتائج التالية:

1- من الجدول زقم (1) والذي يمثل قيم معاملات الارتباط القويم المحسوبه باستخدام طريقة كيرنل وبالدوال الثلاث وكذلك معاملات الارتباط القويم المحسوبه وفق الطريقه التقليديه عندما يكون $\varepsilon_1, \varepsilon_2$ يتبع كل منهما التوزيع الطبيعي القياسي نلاحظ ان معاملات الارتباط القويم المحسوبه وفق طريقه كيرنل كانت اكبر من تلك المحسوبه وفق الطريقه التقليديه ولجميع حجوم العينات وكانت اكبر تلك الارتباطات باستخدام دالة (Gaussian Kernal) ومن ثم باستخدام دالة (Uniform Kernel) وبعد ذلك باستخدام دالة(Triangle Kernel) .

ونلاحظ بشكل عام ان الارتباطات المحسوبه وفق الطريقه التقليديه كانت صغيره جدا .

2- نلاحظ من الجدول رقم (2) والذي يمثل قيم معاملات الارتباط القويم المحسوبه باستخدام طريقة كيرنل وبالدوال الثلاث وكذلك معاملات الارتباط القويم المحسوبه وفق الطريقه التقليديه عندما يكون $\varepsilon_1, \varepsilon_2$ يتبع كل منهما التوزيع الاسي بالمعلمه $\lambda=1/3$ ان معاملات الارتباط القويم المحسوبه وفق طريقه كيرنل كانت اكبر من تلك المحسوبه وفق الطريقه التقليديه عند استخدام دالة (Gaussian Kernel) وكذلك عند استخدام داله (Uniform Kernel) ولجميع حجوم العينات بينما كانت تلك الارتباطات اقل مما كانت عليه ارتباطات الطريقه التقليديه عند استخدام دالة(Triangle Kernel) .

ونلاحظ بشكل عام ان الارتباطات المحسوبه وفق الطريقه التقليديه كانت اكبر مما كانت عليه في الجدول رقم (1).

3- من الجدول رقم (3) والذي يمثل قيم معاملات الارتباط القويم المحسوبه باستخدام طريقة كيرنل وبالدوال الثلاث وكذلك معاملات الارتباط القويم المحسوبه وفق الطريقه التقليديه عندما يكون $\varepsilon_1, \varepsilon_2$ يتبع كل منهما توزيع مربع كاي بالمعلمه n نلاحظ ان معاملات الارتباط القويم المحسوبه وفق طريقه كيرنل كانت اكبر من تلك المحسوبه وفق الطريقه التقليديه ولجميع حجوم العينات وكانت اكبر تلك الارتباطات باستخدام دالة (Gaussian Kernal) ومن ثم باستخدام دالة (Uniform Kernel) وبعد ذلك باستخدام دالة(Triangle Kernel) . ونلاحظ بشكل عام ان الارتباطات المحسوبه وفق الطريقه التقليديه كانت اكبر مما كانت عليه في الجدول رقم (1) ، ولكنها كانت اقل مما كانت عليه في الجدول رقم (2) عند حجوم العينات المتشابهه.

الاستنتاجات:

1- عندما يكون $\varepsilon_1, \varepsilon_2$ يتبع كل منهما التوزيع الطبيعي القياسي او التوزيع الاسي بالمعلمه $\lambda=1/3$ او توزيع مربع كاي بالمعلمه n نلاحظ ان معاملات الارتباط القويم المحسوبه وفق طريقه كيرنل كانت اكبر من تلك المحسوبه وفق الطريقه التقليديه ولجميع حجوم العينات فيما عدا في حالة التوزيع الاسي حيث كانت الارتباطات القويمه المحسوبه باستخدام طريقه كيرنل اقل من الطريقة التقليديه عند استخدام دالة (Triangle Kernel) .

2- كانت الارتباطات القويمه المحسوبه وفق طريقة كيرنل باستخدام دالة(Gaussian Kernal) تمثل اكبر الارتباطات في جميع الحالات وعند جميع حجوم العينات.

المصادر:

1- المياحي ، نوري فرحان وعلي حسين بتور (2005)."مقدمه في التحليل الدالي ".مؤسسة النبراس للطباعه والنشر والتوزيع ، النجف الاشرف،العراق.

2-Anderson, T.W (1984)."An Introduction to Multivariate statistical analysis".Jonwiley and sons.

3-Akaho,S.(2001)"A kernel method for canonical correlation analysis". International meeting of psychometric society, Osaka.

4- Donald F. Morrison (1988)."Multivariate statistical methods". Second edition, McGraw. Hill series in probability and statistics.

5- Florian,M.(2003) "Canonical Correlation Analysis With Kernels". Computational Diagnostics Group Seminar .Berlin.

6-Lai,p.l and C.fyfe(2000)."Kernel and nonlinear canonical correlation analysis " . International Journal Of Neural Systems 10(5) , 365-377.

7- Lindeman,R,H(1980)."Introduction to bivariate and multivariate analysis".Scott-Foresman company,Illinois,U.S.A.
8- Romanazzi,m.(1992)"Influence in canonical correlation analysis", Psychometrika,57,237-259.

الملحق

جدول رقم (1)
يمثل الارتباطات القويمه المحسوبه وفق الطريقتين وبالدوال المختلفه عندما تتبع

الاخطاء $\varepsilon_1, \varepsilon_2$

التوزيع الطبيعي القياسي

الطريقه	الداله المستخدمه	n=10	n=20	n=30	n=40	n=50	n=60	n=70	n=80	n=90
Kernel CCA	Gaussian	0.82	0.80	0.79	0.78	0.78	0.77	0.77	0.77	0.77
	Uniform	0.63	0.62	0.61	0.59	0.58	0.57	0.57	0.57	0.58
	Triangle	0.55	0.52	0.50	0.47	0.46	0.45	0.45	0.46	0.46
CCA	Classical	0.10	0.03	0.02	0.002	0.02	0.04	0.04	0.04	0.03

جدول رقم (2)
يمثل الارتباطات القويمه المحسوبه وفق الطريقتين وبالدوال المختلفه عندما تتبع

الاخطاء $\varepsilon_1, \varepsilon_2$

التوزيع الاسي بالمعلمه $\lambda=1/3$

الطريقه	الداله المستخدمه	n=10	n=20	n=30	n=40	n=50	n=60	n=70	n=80	n=90
Kernel CCA	Gaussian	0.73	0.72	0.71	0.71	0.71	0.71	0.72	0.72	0.72
	Uniform	0.56	0.54	0.54	0.53	0.52	0.52	0.53	0.54	0.54
	Triangle	0.47	0.44	0.43	0.42	0.41	0.42	0.42	0.42	0.42
CCA	Classical	0.50	0.47	0.45	0.44	0.44	0.44	0.44	0.44	0.44

جدول رقم (3)

يمثل الارتباطات القويمه المحسوبه وفق الطريقتين وبالدوال المختلفـــه عنـــدما تتبـــع الاخطاء $\varepsilon_1, \varepsilon_2$

توزيع مربع كاي بالمعلمه n

الطريقه	الداله المستخدمه	n=10	n=20	n=30	n=40	n=50	n=60	n=70	n=80	n=90
Kernel CCA	Gaussian	0.75	0.74	0.72	0.72	0.71	0.70	0.71	0.71	0.71
	Uniform	0.62	0.59	0.57	0.56	0.55	0.54	0.55	0.55	0.55
	Triangle	0.52	0.51	0.48	0.46	0.45	0.44	0.45	0.45	0.45
CCA	Classical	0.43	0.40	0.39	0.36	0.35	0.34	0.34	0.34	0.34