اتساق تابع قرار الجوارات الـ k –الأكثر قرباً باستخدام الطوريات العشوائية α–مزوجة

د. زياد قناية*

د. أحمد يونسو**

نور أزهري***

(تاريخ الإيداع 30 / 2 / 2021. قُبِل للنشر في 29 / 4 /2021)

□ ملخّص □

يعد التصنيف الإحصائي من المواضيع المتقدمة في الإحصاء بطرائقه المختلفة ومن هذه الطرق طريقة النواة، المدرج التكراري و الجوارات الـk الأكثر قرباً المستخدمة في هذا البحث، ويعد موضوع دراسة اتساق توابع القرار من المواضيع التي تشغل العديد من الباحثين،حيث تم في دراسات سابقة دراسة اتساق قاعدة اتساق قاعدة الجوارات الـ k الأكثر قرباً في الحالة المستقلة، وعندما تكون العينة المدروسة مرتبطة يصبح من الضروري استخدام مفهوم المزج باستخدام معاملات المزج المختلفة، لذلك نهدف في هذا البحث إلى اثبات اتساق تابع الجوارات الـ k – الأكثر قرباً في حالة الارتباط الضعيف أي سنثبت أن تابع الجوارات الـ k – الأكثر قرباً متسق عندما تكون العينة التدريبية مشاهدات لطورية عشوائية α مزوجة أو (مزوجة بقوة) ،كما توصلنا من خلال محاكاة عينة تجريبية لدوال مختلفة أن قيمة k الأفضل (عدد الجوارات الأفضل) هو ثلاث جوارات ، أوصت الدراسة بإجراء أبحاث لاحقة للحصول على أقوى تقارب لتوابع القرار لذلك كدراسة التقارب شبه الأكيد لتابع الجوارات الـ k – الأكثر قرباً تحت شروط المزج الأخرى.

الكلمات المفتاحية : الاتساق، الجوارات الـ k الأكثر قرباً، المزج، التصنيف الموجه، تابع القرار،طورية عشوائية.

* أستاذ مساعد، قسم الرياضيات، كلية العلوم، جامعة تشرين ، اللاذقية ، سورية.

** أستاذ مساعد، قسم الإحصاء الرياضي، كلية العلوم، جامعة دمشق ، دمشق ، سورية.

***طالب دراسات عليا دكتوراه ،قسم الرياضيات، كلية العلوم ، جامعة تشرين ، اللاذقية ، سورية.

Consistency the Decision Function of K-nearest Neighbors by Using α-mixing Random Stochastic

Dr. Ziyad Kanaya*
Dr. Ahmad Younso**
Nour Azhari***

(Received 30 / 2 / 2021. Accepted 29 / 4 /2021)

□ ABSTRACT □

The statistical classification is one of the advanced topics in the statistics in its many ways, including the Kernel method, the Histogram and the k-nearest neighbors used in this research, and the subject of the study of the consistency of decision functions is one of the topics that concern many researchers, in previous studies has been studied the consistency of the k-nearest neighbors in the independent case, and when the sample studied is dependent it becomes necessary to use the concept of mixing using different mixing coefficients Therefore, in this research, we aim to prove the consistency of the k-nearest neighbors in the case of weak correlation, i.e. we will prove that the follow-up of the k-nearest neighbors is consistent when the training sample is α-mixing random stochastic, or strong mixing, we came through a simulation of an experimental sample of different functions that the best k value (the number of best neighborhoods) is three neighborhoods, the study recommended subsequent research to obtain stronger convergence of the decision function as a study of the almost certain convergence of the k-nearest neighbors under other mixing conditions.

Keywords: consistency, k-nearest neighbors, mixing, supervised classification, decision functions, random stochastic .

*Associate Prof, Depart. Of Mathematics, Faculty of Science, Tishreen University, Lattakia, Syria.
**Associate Prof, Depart. Of Mathematical Statistics, Faculty of Science, Damascus University, Damascus, Syria.
***Postgraduate Student(Ph.D.), Depart. Of Mathematical Statistics, Tishreen University, Lattakia, Syria.

مقدمة:

التصنيف الإحصائي عملية إحصائية يتم فيها توزيع البيانات لمجتمع إحصائي إلى مجموعات مختلفة بناء على معلومات كمية تستند إلى واحدة او اكثر من الخواص الأساسية لهذه البيانات أو أعضاء المجتمع الإحصائي . تستند عملية التصنيف هذه على خاصيات أصيلة في العناصر (التي قد تكون : رموزا أو متغيرات ....) و تستند على مجموعة تدريب، يستخدم التصنيف بشكل واسع في حل الكثير من المشكلات خاصة تلك التي تتعلق بالأعمال والطب والجرائم وغيرها من خلال تحليل مجموعة من البيانات ووضعها على شكل أصناف أو أقسام يمكن استخدامها فيما بعد لتصنيف البيانات مستقبلاً ، وهناك عدد من الطرق التي يمكن استخدامها في تصنيف البيانات باستخدام خوارزميات مختلفة مثل الخوارزميات الإحصائية Statistical Algorithms والشبكات العصبية Neural Network والخوارزميات الجينية Genetic Algorithms وطريقة الجار الأقرب Nearest Neighbor Method المستخدمة في هذا البحث وهي إحدى خوارزميات التعلم الآلي والتي تعمل بمشرف (موجه) وتعد خوارزمية الجار الأقرب من خوارزميات التصنيف التنبؤية والوصفية.

وللتصنيف نوعان رئيسيان إما تصنيف موجه Supervised Classification والذي يحتاج لوجود عينات تدريبية (Training Samples) وفي هذا النوع يتم مطابقة كل وحدة من الوحدات التجريبية مع العينات التدريبية باستخدام الحسابات الإحصائية ولهذا النوع العديد من الطرق منها طريقة التصنيف الأكثر احتمالية (Maximum Likelihood Classification) والتي تحسب مقدار الارتباط والاختلاف للعينات التدريبية بعضها مع بعض وعلى أساسها يتم توزيع الوحدات التجريبية غير المعروفة إلى تلك الفئات المعروفة[1]، أو تصنيف غير موجه Unsupervised Classification وتستخدم هذه الطريقة في حالة عدم توفر العينات التدريبية واساس عملها يعتمد على كون اي نوع من الاصناف الموجودة في الدراسة متكوناً من وحدات ذات قيم متقاربة مع بعضها، وتتضمن هذه الطريقة حسابات رياضية تختبر عددا كبيرا من الوحدات المجهولة وتقسيمها الى مجاميع معتمدة على القيمة الطيفية لكل وحدة من هذه الوحدات، وهناك تقنيات احصائية عديدة متوفرة بشكل برامج جاهزة بالإمكان استخدامها مع الحاسبات الالكترونية ومن ابرز هذه التقنيات عملية التحليل العنقودي Cluster Analysis. [2]

من أجل إجراء التصنيف الموجه نحتاج إلى بناء تابع قرار تعتمد كفاءته على قربه من دالة قرار أمثلية تسمى دالة قرار بايز (أي أن يكون الخطأ المرتكب أقرب ما يمكن إلى خطأ بايز) وسندرس في هذا البحث اتساق تابع قرار الجوارات الـ k– الأكثر قرباً عندما تكون العينة التدريبية مشاهدات لطورية عشوائية مزوجة.

مشكلة البحث:

تم بناء معظم نظريات الاحصاء التقليدي اعتماداً على متحولات عشوائية مستقلة وذلك بالاستفادة من مبرهنات النهاية المركزية وقانون الأعداد الكبيرة، إلا أنه في بعض الدراسات كالسلاسل الزمنية وبعض المقدرات التابعية يكون شرط الاستقلال غير محقق فقد تكون المتحولات العشوائية المدروسة مرتبطة وفق صيغة معينة ، لذلك سندرس أسرة من الطوريات العشوائية المؤلفة من متحولات عشوائية مرتبطة احتمالياً وفق مفهوم معين يسمى المزج mixing ووفق هذا المفهوم تميل متحولات الطورية الى الاستقلال مع تباعد الزمن والتكرارات .ففي السلاسل الزمنية على سبيل المثال (دراسة تغيرات درجة الحرارة،دراسات الطقس...) يلاحظ كلما زادت الفروق الزمنية أو التكرارات بين المشاهدات قل تأثير المشاهدات السابقة على المشاهدات الحالية أو المستقبلية حتى أنه من أجل قفزات كبيرة للزمن أو المشاهدات

تميل الى أن تصبح مستقلة عن المشاهدات السابقة ومن هنا انطلقت العديد من الدراسات حول مفهوم المزج الذي له أشكال عديدة تتمثل مشكلة البحث في دراسة اتساق تابع قرار الجوارات الـk – الأكثر قرباً من تابع قرار بايز وذلك في حالة الارتباط أي **عندما تكون العينة التدريبية مشاهدات لطورية عشوائية مزوجة.**

أهمية البحث و أهدافه:

يهدف البحث الى دراسة اتساق قاعدة تابع قرار الجوارات الـk – الأكثر قرباً من تابع قرار بايز وذلك في حالة الارتباط أي عندما تكون العينة التدريبية مشاهدات لطورية عشوائية مزوجة وذلك كتعميم لحالة الاتساق المثبت في الحالة المستقلة [3] ثم إجراء محاكاة باستخدام البرنامج الإحصائي R.

الإطار النظري:

الطوريات العشوائية المزوجة بالمفهوم α $(\alpha - mixing)$ [4,5,6]:

لتكن $(Z_i, i \geq 1)$ طورية عشوائية معرفة على فضاء احتمالي $(\Omega, \mathcal{F}, \mathrm{P})$ وتأخذ قيمها في فضاء قيوس$(\Omega`, \mathcal{F}`)$، عندها نقول عن الطورية $(Z_i, i \geq 1)$ أنها α – مزوجة (أو مزوجة بقوة) إذا تحقق الشرط التالي:

من أجل الجبرين التامين $\mathcal{F}_1^{\ell}, \mathcal{F}_{\ell+n}^{+\infty}$ المولدان بـ $(Z_t, t = 1, \dots, k), (Z_t, t = \ell + n, \dots)$ على الترتيب يكون:

$$\alpha(n) = \sup_{\ell \geq 1} \sup_{A \in \mathcal{F}_1^{\ell}, B \in \mathcal{F}_{\ell+n}^{+\infty}} |P(A \cap B) - P(A)P(B)| \xrightarrow[n \to \infty]{} 0 \qquad (1)$$

لاحظ أنه كلما اقتربت $\alpha(n)$ من الصفر كلما نزعت الطورية نحو الاستقلال وخصوصاً عندما $\alpha(n) = 0, n \geq 0$ ويمكن التحقق من أن العديد من نماذج $ARMA$ **تحقق شرط المزج (1) تحت شروط معينة [7] فمثلاً** $AR(1)$ تعرف بالشكل $X_t = \rho X_{t-1} + \varepsilon_t \ ; \ |\rho| < 1$.

قاعدة الجوارات الـ k – الأكثر قرباً [3]:

لتكن $\{(X_i, Y_i), i \geq 1\}$ طورية عشوائية مستقرة بقوة (بمعنى أن الخصائص الاحتمالية للمتجهات تبقى ثابتة عند تغير الزمن بمقدار ثابت) معرفة على فضاء احتمالي $(\Omega, \mathcal{F}, \mathrm{P})$ وتأخذ قيمها في الفضاء $\mathbb{R}^d \times \{0,1\}$، في التصنيف الموجه يسمى X_i متجه السمات (الخصائص) ويسمى Y_i الصف الموافق لـ X_i ونسعى من خلال هذا التصنيف إلى التنبؤ بالصنف Y_j لمتجه السمات X_j الموافق لمشاهدة جديدة ليست من ضمن العينة.

بما أن الطورية مستقرة بقوة فرضاً يمكن اعتبار التوزيع الاحتمالي لأي شعاع (X_i, Y_i) مطابق لتوزع الشعاع (X, Y) والذي يعرف جيداً من خلال μ و η حيث μ القياس الاحتمالي لـX حيث $\mu(A) = P(x \in A)$ وη تابع انحدار Y على X عندما X يأخذ القيمة x أي من خلال $\eta(x) = E(Y/X = x)$.

من أجل إجراء التصنيف الموجه نحتاج إلى بناء تابع قرار $g: \mathbb{R}^d \to \{0,1\}$ بحيث $g(X)$ تقابل صنف X وعندها نرتكب خطأ عندما $Y \neq g(X)$ حيث Y هو الصنف الفعلي لـ X.

نرمز لاحتمال الخطأ لتابع القرار g بـ $L(g) = \mathbb{p}(Y \neq g(X))$ وإن احتمال هذا الخطأ يكون أصغرياً من أجل تابع القرار g^* المعرف بالشكل [3]:

$$g^*(x) = \begin{cases} 0 & if \ P(Y = 0/X = x) \geq P(Y = 1/X = x) \\ 1 & otherwise \end{cases}$$

يسمى$g^*(x)$ تابع قرار بايز ونرمز لاحتمال الخطأ الموافق بالرمز $L^* = L(g^*)$ ويسمى خطأ بايز.

لسوء الحظ تابع قرار بايز غير قابل للاستخدام مباشرة في التصنيف لأنه يعتمد على التوزيع الاحتمالي لـ (X,Y) والذي هو في الغالب غير معلوم لذلك نسعى إلى تقدير $g^*(x)$ من خلال عينة عشوائية $D_n = \{(X_i,Y_i), i = 1,\dots,n\}$ تسمى عينة تدريبية وبفضل هذه العينة التدريبية نعرف تابع قرار الجوارات الـ k – الأكثر قرباً $g_n(x)$ كما يلي حيثk عدد صحيح موجب تماماً أي $k \geq 1$:[3]

$$g_n(x) = \begin{cases} 0 & if \ \sum_{i=1}^{n} w_{n_i} Y_i \leq \frac{1}{2} \\ 1 & otherwise \end{cases}$$

حيث $w_{n_i} = w_{n_i}(x, D_n)$ يساوي $\frac{1}{k}$ عندما X_i واحدة من الـk الأقرب لـ x في D_n و يساوي الصفر خلاف ذلك حيث $k = k(n)$ متتالية من الأعداد الصحيحة الموجبة تماماً التي تحقق من أجل $n \to \infty$ الشرطين: $k \to \infty$, $\frac{k}{n} \to 0$ (فرضيات تقليدية في الحالة المستقلة[3]).

لنفرض أن X له كثافة احتمالية f حيث يمكن من خلال ذلك تجنب وقوع ربطات ناتجة عن تساوي بعد مجاورتين عن x ، لنأخذ :

$$\eta_n(x) = \sum_{i=1}^{n} w_{n_i} Y_i$$

يسمى $\eta_n(x)$التقدير الـk مجاورة أكثر قرباً لدالة الانحدار $\eta(x)$ ولنرمز بـ$g_n(x)$ بتابع القرار التجريبي واحتمال خطأ التصنيف له $L_n = L(g_n)$ ويسمى الخطأ التجريبي حيث يمكن أن نكتب:

$$g_n(x) = \begin{cases} 0 & if \ \eta_n(x) \leq \frac{1}{2} \\ 1 & otherwise \end{cases}$$

أفضل ما نتوقعه من $g_n(x)$ هو أن يكون الخطأ أقرب ما يكون إلى خطأ بايز L^* وهذا يعبر عنه من خلال دراسة أشكال الاتساق المختلفة لـ $g_n(x)$ نحو تابع قرار بايز $g^*(x)$.

تعريف الاتساق: نقول عن تابع القرار $g_n(x)$ أنه متسق إذا تحقق الشرط:

$$E(L_n) \underset{n\to\infty}{\longrightarrow} L^*$$

أي أن$g_n(x)$ متسق عندما يكون متوسط الخطأ التجريبي $E(L_n)$ يتقارب عددياً نحو L^* .

سنثبت في هذه الورقة أن تابع الجوارات الـ k – الأكثر قرباً متسق عندما تكون العينة التدريبية مشاهدات لطورية عشوائية مزوجة .

قبل أن نتناول النتيجة الرئيسية لنقدم التمهيديات التالية:

تمهيدية(1) [8]: ليكن z_1, z_2 متحولين عشوائيين مستمرين بقيم حقيقية ولنفرض أنهما محدودان عندئذ:

$$|cov(z_1,z_2)| \leq 4\|z_1\|_\infty \|z_2\|_\infty \propto \{\sigma(z_1),\sigma(z_2)\}$$

حيث :

$$\propto \{\sigma(z_1),\sigma(z_2)\} = \sup_{\substack{A\in\sigma(z_1)\\ B\in\sigma(z_2)}} |P(A\cap B) - P(A)P(B)|$$

معامل المزج بين الجبرين التامين $\sigma(z_1),\sigma(z_2)$ المولدين بـ z_1, z_2 على الترتيب.

تمهيدية(2) [9]: لتكن المجموعة :

$$B_a(x^{\backslash}) = \left\{x \in \mathbb{R}^d; \mu(S_{x,\|x-x^{\backslash}\|}) \leq a\right\}$$

عندها من أجل كل $x^{\backslash} \in \mathbb{R}^d$:

$$\mu\left(B_a(x^{\backslash})\right) \leq a\gamma_d$$

حيث γ_d العدد الأصغري للمخاريط المتمركزة في المبدأ وبزاوية $\frac{\pi}{6}$ وتغطي $\mathbb{R}^d$.

النتيجة الرئيسية:

مبرهنة: بفرض أن العينة التدريبية D_n مشاهدات لطورية عشوائية α–مزوجة تحقق الشرط :

$\alpha(t) = O\left(t^{-\theta}\right)$ حيث $\theta > 1$ إذا تحقق من أجل $n \to \infty$ الشرطين:

$$\frac{k}{\sqrt{n}} \to \infty$$
$$k \to \infty$$

حيث هنا الشروط على k أضعف من الشروط المقترحة في [8] عندئذ:

$$E(L_n) = \underset{n\to\infty}{\longrightarrow} L^*$$

قبل البدء بإثبات المبرهنة لنعرف بعض الرموز المهمة في عملية البرهان:

لنرمز بـ $\rho_n = \rho_n(x)$ لحل المعادلة $(*)$ $\frac{k}{n} = \mu\left(S_{x,\rho_n}\right)$ حيث إن الحل موجود لأن X يملك تابع كثافة ولنعرف أيضاً :

$$\widehat{\eta_n}(x) = \frac{1}{k}\sum_{i=1}^{n} Y_i \mathbb{I}_{(X_i \in S_{x,\rho_n})}$$

حيث $\mathbb{I}_A$ التابع المميز للمجموعة A حيث:

$$\mathbb{I}_A(w) = \begin{cases} 1 & ; w \in A \\ 0 & ; w \notin A \end{cases}$$

اثبات المبرهنة:

بحسب المبرهنة في [3] يكفي أن نثبت أن :

$$E\int_{\mathbb{R}^d} |\eta(x) - \eta_n(x)|\mu(dx) \underset{n\to\infty}{\longrightarrow} 0$$

لكن:

$$|\eta(x) - \eta_n(x)| \leq |\eta(x) - E\widehat{\eta_n}(x)| + |E\widehat{\eta_n}(x) - \eta_n(x)| \quad (i)$$

بما أن $\frac{k}{n} \to 0$ فرضاً فإنه وباستخدام $(*)$ يتحقق $\rho_n \to 0$ عندما $n \to \infty$ وبحسب مبرهنة لوبيغ نحصل من أجل $n \to \infty$:

$$E\widehat{\eta_n}(x) = \frac{1}{\mu\left(S_{x,\rho_n}\right)}\int_{S_{x,\rho_n}} E\left(Y/X = x^{\backslash}\right)\,\mu\left(dx^{\backslash}\right)$$
$$\to E(Y/X = x) = \eta(x) \quad \forall x \; mod\; \mathcal{M}$$

بما أن $|Y| \leq 1$ نحصل حسب مبرهنة التقارب الراجح عندما $n \to \infty$:

$$\int_{\mathbb{R}^d} |\eta(x) - E\widehat{\eta_n}(x)|\mu(dx) \underset{n\to\infty}{\longrightarrow} 0 \qquad (ii)$$

بحسب (i) و (ii) يكفي أن نثبت أنه عندما $n \to \infty$:

$$E\int_{\mathbb{R}^d}|E\widehat{\eta_n}(x)-\eta_n(x)|\mu(dx)\xrightarrow[n\to\infty]{}0$$

ولدينا المتراجحة التالية:

$$E\int_{\mathbb{R}^d}|E\widehat{\eta_n}(x)-\eta_n(x)|\mu(dx)$$

$$\leq E\int_{\mathbb{R}^d}|E\widehat{\eta_n}(x)-\widehat{\eta_n}(x)|\mu(dx)+E\int_{\mathbb{R}^d}|\widehat{\eta_n}(x)-\eta_n(x)|\mu(dx)$$

لنثبت أن الحدين في يمين المتراجحة يتقاربان نحو الصفر عندما $n \to \infty$.

من أجل الحد الأول وحسب متراجحة كوشي شوارتز :

$$E\int_{\mathbb{R}^d}|E\widehat{\eta_n}(x)-\widehat{\eta_n}(x)|\mu(dx)\leq\int_{\mathbb{R}^d}\sqrt{E\left(E\widehat{\eta_n}(x)-\widehat{\eta_n}(x)\right)^2}\mu(dx)$$

$$=\int_{\mathbb{R}^d}\sqrt{Var(\widehat{\eta_n}(x))}\mu(dx)$$

$$\leq\int_{\mathbb{R}^d}\sqrt{\frac{n}{k^2}Var\left(Y\mathbb{I}_{(X\in S_{x,\rho_n})}\right)+C_n(x)}\mu(dx)$$

حيث :

$$C_n(x)=\frac{1}{k^2}\sum_{i\neq j}\left|cov\left(Y_i\mathbb{I}_{(X_i\in S_{x,\rho_n})},Y_j\mathbb{I}_{(X_j\in S_{x,\rho_n})}\right)\right|$$

من جهة أخرى لدينا:

$$\frac{n}{k^2}Var\left(Y\mathbb{I}_{(X\in S_{x,\rho_n})}\right)\leq\frac{n}{k^2}E\left(\mathbb{I}_{(X\in S_{x,\rho_n})}\right)=\frac{n}{k^2}\mu(S_{x,\rho_n})=\frac{1}{k}$$

من جهة أخرى لدينا بحسب التمهيدية1:

$$C_n(x)\leq\frac{4}{k^2}\sum_{i\neq j}\alpha(|i-j|)\leq\frac{4n}{k^2}\sum_{i=1}^{\infty}\alpha(i)\leq\frac{4n}{k^2}C\sum_{I=1}^{\infty}i^{-\theta},\theta>1$$

$$\leq C\frac{4n}{k^2}\xrightarrow[n\to\infty]{}0$$

لأن $\frac{k}{\sqrt{n}} \to \infty$ فرضاً.

$$\Longrightarrow E\int_{\mathbb{R}^d}|E\widehat{\eta_n}(x)-\widehat{\eta_n}(x)|\mu(dx)\xrightarrow[n\to\infty]{}0$$

الآن سنبرهن أن الحد الثاني في المتراجحة المذكورة يسعى للصفر ولأجل ذلك لنرمز $X_{(k)}$ للمجاورة الـk الأقرب لـx وليكن $r_n=r_n(x)=\left\|X_{(k)}(x)-x\right\|$عندا واضح أن :

$$|\widehat{\eta_n}(x) - \eta_n(x)| = \left|\frac{1}{k}\sum_{i=1}^{n} Y_i \mathbb{I}_{(X_i \in S_{x,\rho_n})} - \frac{1}{k}\sum_{i=1}^{n} Y_i \mathbb{I}_{(X_i \in S_{x,r_n})}\right|$$
$$\leq \frac{1}{k}\sum_{i=1}^{n} \left|\mathbb{I}_{(X_i \in S_{x,\rho_n})} - \mathbb{I}_{(X_i \in S_{x,r_n})}\right| \leq \left|\frac{1}{k}\sum_{i=1}^{n} \mathbb{I}_{(X_i \in S_{x,\rho_n})} - 1\right|$$
$$= |\widehat{\eta_n}(x) - E\widehat{\eta_n}(x)|$$

حيث $\widehat{\eta_n}(x) = \frac{1}{k}\sum_{i=1}^{n} \mathbb{I}_{(X_i \in S_{x,\rho_n})}$

وعليه يكفي أن نثبت أنه عندما $n \to \infty$ فإن:

$$E \int_{\mathbb{R}^d} |\widetilde{\eta_n}(x) - E\eta_n(x)| \mu(dx) \xrightarrow[n\to\infty]{} 0$$

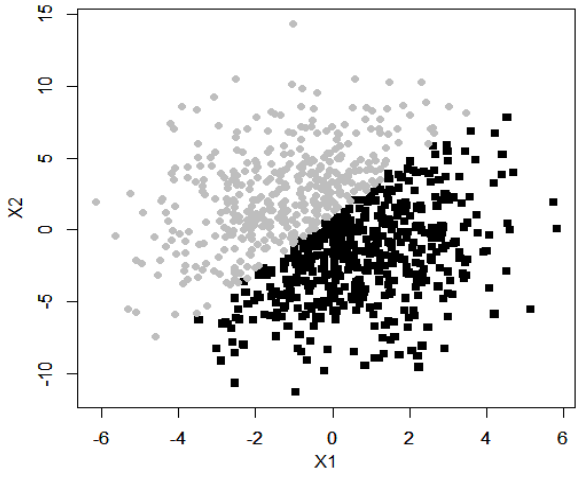
لاحظ أن $\widetilde{\eta_n}(x) = \widehat{\eta_n}(x)$ عندما $Y_i = 1$ من أجل $i = 1, \ldots, n$ نتيجة لذلك الإثبات مشابه تماماً لما قمنا به سابقاً حول $\eta_n(x)$.

محاكاة النتائج:

تم استخدام أسلوب المحاكاة بواسطة برنامج تم كتابته بلغة R [10,11] حيث نبين من خلاله آلية التصنيف وفق الخطوات التالية:

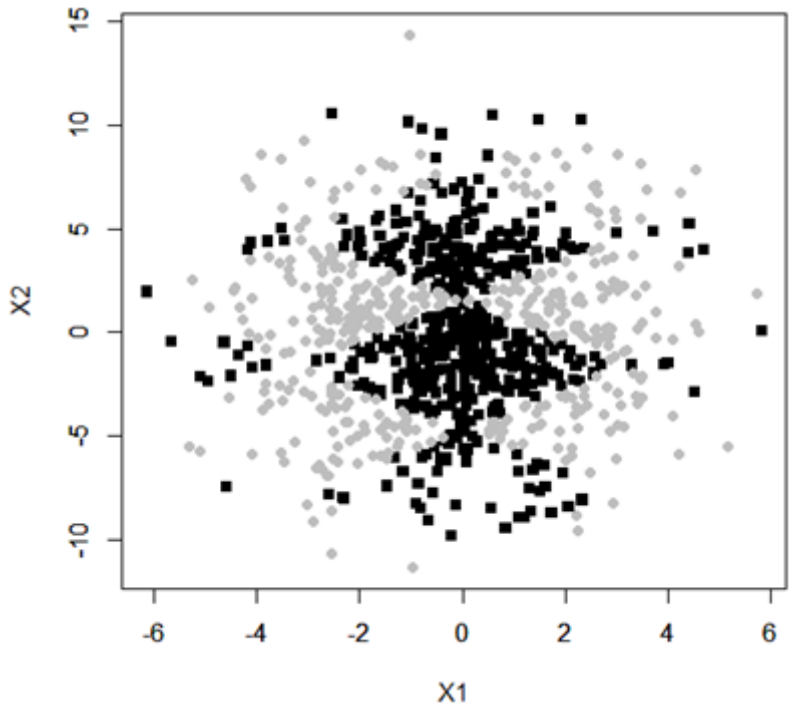
1. تابع لحساب مصفوفة المسافات بين الأدلة.
2. تابع لحساب مصفوفة التباين والتغاير.
3. محاكاة عينة حجمها ألف مرتبطة.
4. تقسيم العينة إلى عينة تدريبية حجمها 900 وعينة اختبار حجمه 100.
5. تابع لحساب قاعدة التصنيف لقيمة واحدة.
6. تابع لحساب قاعدة التصنيف لمجموعة قيم.
7. رسم العينة التدريبية لدوال مختلفة.
8. حساب مجموع مربعات الخطأ لعينة الاختبار من أجل قيم مختلفة لـ k.

-من أجل دوال بسيطة (مثلاً كمعادلة مستقيم) لاحظ دقة التصنيف:



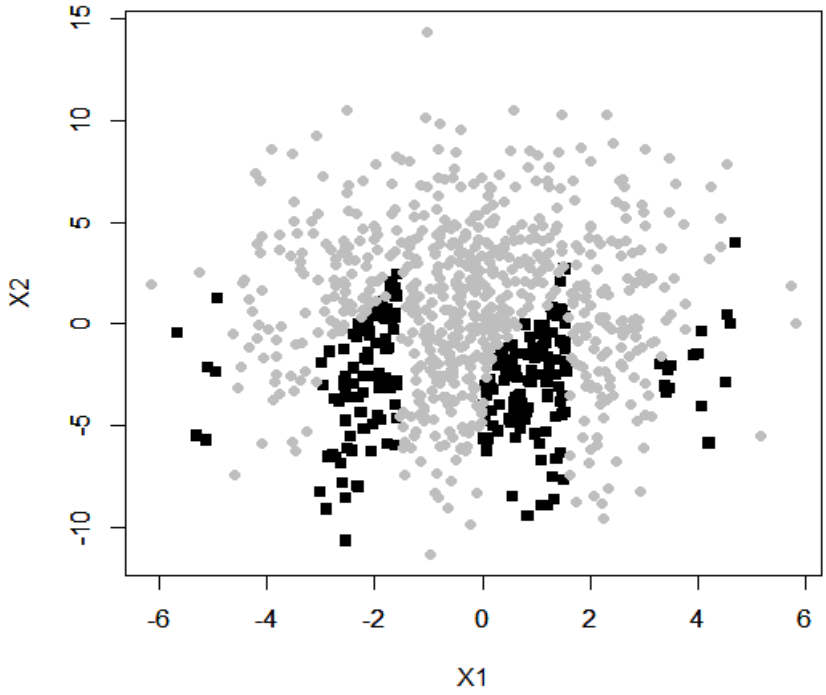
الشكل (1) تصنيف عينة بدوال بسيطة (معادلة مستقيم)

– من أجل دوال أكثر تعقيداً sin و cos وعينة حجمها 1000 مقسمة كما ذكرنا سابقاً يكون من أجل القرارين الشكل:



الشكل (2) تصنيف عينة بدوال معقدة (sin و cos)

– من أجل دوال افتراضية tan و exp وعينة حجمها 1000 مقسمة كما ذكرنا سابقاً يكون من أجل القرارين الشكل :



الشكل (3) تصنيف عينة بدوال معقدة (tan و exp)

وتكون قيم مجموع مربعات الخطأ من أجل قيم مختلفة لـ k والدوال السابقة كما هو موضح بالجدول(1) التالي:

الجدول(1) قيم مجموع مربعات الخطأ لقيم مختلفة لـk

معادلة مستقيم	k	1	3	5	7	9	33
	MSE	1	3	2	2	2	1
sin و cos	k	1	3	5	7	9	33
	MSE	6	5	4	5	6	10
tan و exp	k	1	3	5	7	9	33
	MSE	7	5	7	8	8	16

حيث نلاحظ أن أفضل قيمة لـ k في الحالة البسيطة 1 أو 33 وفي الحالة الأكثر تعقيداً (sin و cos) كانت عندما k=5 وعندما (tan و exp) كانت عندما k=3 وبالتالي يمكن اعتبار القيمة المثلى لـk هي 3 .

وقمنا بكتابة تابع لتصنيف قيمة جديدة:

```
CLASS=function(x,X1,X2,Y,k){
XX=cbind(X1,X2)
d=rep(0,length(X1))
for (i in 1:length(X1)){
d[i]=sqrt(sum((x-XX[i,])^2))}
ds=sort(d)
D=ds[1:k]
m=max(D)
dc=(d<=m)
gn=Y[dc]
gn=sum(gn)/k
Class=(gn>0.5)
Class;}
```

كما قمنا بكتابة تابع لتصنيف مجموعة قيم جديدة:

```
CLASS=function(x,X1,X2,Y,k){
XX=cbind(X1,X2)
d=rep(0,length(X1))
for (i in 1:length(X1)){
d[i]=sqrt(sum((x-XX[i,])^2))}
ds=sort(d)
D=ds[1:k]
m=max(D)
dc=(d<=m)
gn=Y[dc]
gn=sum(gn)/k
Class=(gn>0.5)
Class;
}
MCLASS=function(x,X1,X2,Y,k){
n=nrow(x)
h=rep(0,n)
```

```
for (i in 1:n){
h[i]=CLASS(x[i,],X1,X2,Y,k)}
h;}
```

وبالتالي من أجل أي قيمة جديدة ولتكن مثلاً 0.5 وعند k=12 مثلاً نستخدم التابع:

CLASS(0.5,X1,X2,Y,12)

تكون النتيجة true

و من أجل أي قيمة جديدة ولتكن مثلاً 6 وعند k=12 مثلاً نستخدم التابع:

CLASS(6,X1,X2,Y,12)

تكون النتيجة false

الاستنتاجات والتوصيات:

توصلنا في هذا البحث إلى اثبات اتساق تابع الجوارات الـ k – الأكثر قرباً في حالة الارتباط الضعيف أي أن تابع الجوارات الـ k – الأكثر قرباً متسق عندما تكون العينة التدريبية مشاهدات لطورية عشوائية مزوجة كما توصلنا من خلال محاكاة عينة تجريبية لدوال مختلفة أن قيمة k الأفضل (عدد الجوارات الأفضل) هو ثلاث جوارات ، ونسعى في دراسات لاحقة للحصول على أقوى تقارب لتوابع القرار لذلك يمكن دراسة التقارب شبه الأكيد لتابع الجوارات الـ k – الأكثر قرباً تحت شروط المزج الأخرى.

References:

[1]Li, N., Martin, A., & Estival, R. *Heterogeneous information fusion: combination of multiple supervised and unsupervised classification methods based on belief functions(2020).*

[2]Ueda, R. M., Souza, A. M.,&Menezes,R. M. C. P. *How macroeconomic variables affect admission and dismissal in the Brazilian electro-electronic sector: A VAR-based model and cluster analysis. Physica A: Statistical Mechanics and Its Applications,* (2020) *124872.*

[3]Luc Devroye ,Laszlo Gyorfi,Gabor Lugosi, *A Probabilistic Theoryof Pattern Recognition,* (1996) Springer-Verlag New York, Inc,P26-27.P78-79.

[4]Rosenblatt, M. Remarks on some non-parametric estimates of the density function. Annals Math. Statist. 27, (1956) 832-837.

[5]E. Rio, *Th´ eorie asymptotique des processus al´ eatoires faiblement d´ ependants*. Springer Verlag, Berlin Heidelberg (2000).p40.

[6] M. Rosenblatt, *A central limit theorem and a strong mixing condition*. *Proc. Nat. Acad. Sci., USA* 42 (1956) 43–47.

[7] Bandyopadhyay, Soutir; *A NOTE ON STRONG MIXING***,** Department of Statistics, Iowa State University, April 21, 2006.

[8]H.C.P. Berbee, *Random walks with stationary increments and renewal theory*. *Math. Cent. Tract.* Amsterdam (1979).

[9]Devroye, L. and Gyorfi, L. *Nonparametric Density Estimation: The L1 View.* John Wiley, New York(1985).

[10]CRAWLEY J. M. *The R book. 2nd. ed*., John Wiley & Sons, Ltd.,(2013) 1060.

[11]COHEN, Y.; COHEN, J.Y.*Statistics and Data with R: An Applied Approach Through Examples*. A John Wiley & Sons, Ltd. (2008).