Warith Scientific Journal

مقارنة التحليل التمييزي اللبي الحصين مع التحليل التمييزي اللبي باستعمال المحاكاة
Comparison Robust Kennel Discriminant Analysis with Kennel Discriminant Analysis by Simulation

* الباحث جعفر علي فرحان
جامعة كربلاء- كلية الإدارة والاقتصاد
Jaafar Ali Farhan
Statistics, College of Administration and Economics, Iraq
jafar.a@s.uokerbala.edu.iq

** أ.م.د. ايناس عبد الحافظ محمد
جامعة كربلاء- كلية الإدارة والاقتصاد
Enas Abdul Hafedh Mohammed
Statistics, College of Administration and Economics, Iraq
enas.albasri@uokerbala.edo.iq

المستخلص :
ان غالبية البيانات في العالم الحقيقي تنحرف عن الافتراضات المثالية التي تقدمها الأساليب الإحصائية القياسية ففي هذه المواقف يتم فيها انتهاك افتراض الحالة الطبيعية في البيانات (التوزيع الطبيعي للبيانات) ، او ان هنالك هياكل مجمعة غير خطية في مجموعة البيانات. فاننا نواجه مشكلة في التصنيف لايمكن للتحليل التمييزي التقليدي مواجهة هذه المشكلة فلابد من البحث عن طريقة حصينة تتعامل مع هذه المشكلة لذلك هدفت هذه الرسالة الى استعمال اسلوب التحليل التمييزي اللبي الحصين (Robust Kennel Discriminant analysis RKDA) في حلة وجود تلوث في البيانات ومقارنته مع التحليل التمييزي اللبي باستعمال معيار معدل خطأ التصنيف لاختيار افصل عرض حزمة عن طريق معيار للدقة اذ تبين ان اسلوب التحليل التمييزي اللبي الحصين افضلية على باقي الاساليب عند دوال الكثافة المنحرفة عن التوزيع الطبيعي بنسبة افضلية عالية.
الكلمات المفتاحية : تحليل تمييزي – حصانة – كثافة لبية- تقدر- دوال كثافة – توزيع طبيعي

Abstract:
The majority of data in the real world deviates from the ideal assumptions made by standard statistical methods. In these situations, the assumption of normality in the data (normal distribution of the data) is violated, or there are non-linear aggregate structures in the data set. We are facing a problem in classification. Traditional discriminant analysis cannot confront this problem. We must search for a robust method that deals with this problem. Therefore, this thesis aimed to use the Robust Kennel Discriminant analysis (RKDA) method in the case of contamination in the data and compare it with discriminant analysis. Using the classification error rate criterion to choose the best band width using an accuracy criterion, as it was shown that the robust kennel discriminant analysis method has an advantage over other methods when density functions deviate from the normal distribution with a high percentage of preference.
Keywords: *Discriminant analysis - Robustness - kernel density - estimation - density functions - normal distribution*

1. مقدمة (Introduction):

اكتسب تطبيق التحليل التمييزي (Discriminant analysis DA) اهتمامًا في مختلف مجالات العلوم الاجتماعية من الاقتصاد, العلوم الاجتماعية , التعليم , التمويل والهندسة وغيرها من العلوم. فإن مشكلة التحليل التمييزي هي واحدة من تعيين مشاهدة مجهولة إلى مجموعة ذات معدل خطأ منخفض. قد تكون الدالة أو الدوال المستعملة للتخصيص مماثلة لتلك المستعملة في تحليل التباين متعدد المتغيرات. لذلك يعد التحليل التمييزي أداة قيمة في ترسانة محللي البيانات وممارسي التعلم الآلي ، مما يوفر نهجًا مباشرًا وقابل للتفسير لمهام التصنيف وتقليل الأبعاد. إن قوته وتعدد استخداماته تجعله خيارًا شائعًا في مختلف المجالات ، مما يساهم في تقدم البحث والتطبيقات العملية.

يعد التحليل التمييزي الحصين (Robust Discriminant analysis RDA) امتدادًا للتحليل التمييزي التقليدي (DA) الذي يعالج قيود النهج القياسي ، خاصةً عند انتهاك افتراضات البيانات. تم تصميم RDA للتعامل مع المواقف التي قد لا تفي فيها البيانات بشكل كامل بافتراضات الحالة الطبيعية والامثلية التي تتطلبها DA. يمكن أن يكون التحليل التمييزي الحصين حساسًا للقيم الشاذة ، خاصةً عندما تنحرف البيانات عن الحالة الطبيعية. يستخدم قانون التمييز الحصين تقنيات إحصائية حصينه أقل تأثراً بالمشاهدات الشاذة ، مما يجعلها أكثر موثوقية في وجود القيم الشاذة. ولا يتطلب التحليل التمييزي الحصين الالتزام

بافتراض الحالة الطبيعية ومصفوفات التغاير المتساوية للمجموعات المختلفة. إنه أكثر مرونة ويمكن أن يستوعب البيانات ذات التوزيعات غير العادية وهياكل التغاير المتباينة. ونظرًا لأن RDA أقل اعتمادًا على افتراضات توزيع البيانات ، يمكن تطبيقه بشكل فعال على نطاق أوسع من مجموعات البيانات في العالم الحقيقي ، خاصة تلك ذات التوزيعات المنحرفة أو ذات الذيل الثقيل.وعند التعامل مع أحجام عينات صغيرة ، يمكن أن تكون افتراضات DA التقليدية مشكلة لذلك يميل RDA إلى تحقيق نتائج أكثر استقرارًا ودقة في مثل هذه الحالات. يستخدم DA طرقًا حصينة لتقدير مصفوفات التغاير ، مثل مقدر الحد الأدنى من محدد التغاير (MCD) أو مقدر M. هذه الطرق أقل تأثرًا بالقيم المتطرفة وتوفر تقديرات تغاير أكثر موثوقية.

فبصورة عامة تكمن أهمية التحليل التمييزي الحصين في قدرته على تقديم نتائج أكثر موثوقية ودقة في وجود بيانات تنتهك افتراضات DA التقليدية. من خلال احتساب القيم الشاذة وتخفيف افتراضات التوزيع ، تقدم RDA بديلاً حصيناً ومرنًا لمهام التصنيف وتقليل الأبعاد ، مما يجعله أداة قيمة في مختلف المجالات حيث قد تكون البيانات عرضة لعدم اليقين والتنوع. ففي عام (2009) اقترح (Nudurupati) تحليل تمييزي غير معلمي أقل حساسية من التحليل التمييزي التقليدي للانحرافات عن الافتراضات المعتادة كاعتدالية البيانات باستعمال منهجية متابعة اسقاط المجموعات الداخلة في التصنيف [4], وفي عام (2017) اقترح (Li et al.) وآخرون التحليل التمييزي اللبي اللامعلمي المحلي ((local kernel nonparametric discriminant analysis)) (LKNDA) والتي تدمج التحليل التمييزي التقليدي مع الاحصاء اللامعلمي وتم مقارنة الطريقة مع طرائق التحليل التمييزي التقليدي باستعمال ست تجارب محاكاة والتي اثبتت جميعها ان طريقة (LKNDA) لها دقة تصنيف اعلى وتعد حلاً بديلاً للحالات التمييزية لاستخراج الميزات غير الخطية المعقدة [3]. وفي عام 2019 اقترح (Yu et al.) انموذج تحليل تمييزي بيزي لامعلمي جديد يقوم باختيار المتغير وتصنيفه ضمن إطار عمل بسيط. يتم تعيين مقدمات شجرة Polya للتوزيعات المجهولة لمجموعة مشروطة لحساب عدم الدقة [5]. وفي عام (2020) استعملت (سكينه) اســـــــلوبين في تصنيف (تمييز) البيانات وهما اسلوب التحليل التمييزي الخطي (LDA) و اسلوب التحليل التمييزي اللبي (KDA) وتوصلت التوصل الى ان اسلوب التحليل التمييزي الخطي هو الافضل لكونه يعطى اقل خطأ تصنيف للبيانات . اذ تم اجراء المقارنة بين الاسلوبين وفق معيار احتمال خطأ التصنيف (Misclassification) .[1] . في نفس عام (2020) اقترح (Obudho et al.) وآخرون دالة تمييز لبية حصينة لامعلمية تم بواسطتها معادلة مشكلة التصنيف في الحالات التي يتم فيها انتهاك شرط التوزيع الطبيعي للبيانات المستعملة اذ حساب معدلات التصنيف الخاطئ لمختلف مصفوفات النطاق الترددي وقارنوا الطريقة المقترحة مع دوال التصنيف المعلمية مثل التمييز الخطي والتمييز التربيعي باستعمال تجارب المحاكاة وتم التوصل الى ان الطريقة المقترحة تؤدي أداءً جيدًا من حيث معدلات التصنيف الخاطئ لمصنف تمييز اللب عند تحديد النطاق الترددي الصحيح بالمقارنة مع المصنفات الموجودة الأخرى المستعملة [2]. في عام (2022) استعمل (Gupta et al.) وآخرون التحليل التمييزي بأبعاد مختلفة مثل الخطي والتربيعي للتصنيف الثنائي لتحليل متلازمة تكيس المبايض ومقارنته باسلوب التعلم الآلي وطريقة تقليل الأبعاد الخاضعة للإشراف وتم التوصل الى ان باستخدام التحليل التمييزي يحقق دقة اعلى وتباين أقل مع دقة تدريب تصل إلى 97.37٪ ودقة اختبار 95.92٪ باستخدام التحليل التربيعي التمييزي مقارنة بباقي الطرائق[1] .

2. مشكلة البحث (Problem of the Article)

غالبًا ما تنحرف بيانات العالم الحقيقي عن الافتراضات المثالية التي تقدمها الأساليب الإحصائية القياسية. ففي هذه المواقف يتم فيها انتهاك افتراض الحالة الطبيعية في البيانات (التوزيع الطبيعي للبيانات) ، او ان هنالك هياكل مجمعة غير خطية في مجموعة البيانات. فاننا نواجه مشكلة تواجه مشكلة في التصنيف لايمكن للتحليل التمييزي التقليدي مواجهة هذه المشكلة فلابد من البحث عن طريقة حصينة تتعامل مع هذه المشكلة. يعترف قانون التمييز الحصين (RDA) بهذه الحقيقة ويقدم نهجًا أكثر واقعية للتحليل.

3. هدف البحث (Aim of the Article)

يهدف البحث الى استعمال التحليل التمييزي اللبي الحصين في حالة وجود تلوث في البيانات ومقارنته مع التحليل التمييزي اللبي التقليدي باستعمال معيار معدل خطأ التصنيف لاختيار افصل عرض حزمة ومن ثم مقارنة الطرائق باستعمال معيار للدقة.

4. معدل خطأ التصنيف (Misclassification rate)

هذا المعدل هو النسبة المئوية للنقاط التي يتم تعيينها إلى المجموعة غير الصحيحة بناءً على قاعدة تمييز. ويُعرف باسم معدل الخطأ في التصنيف ، وهو مقياس تقييم شائع يستخدم لتقييم أداء نموذج التصنيف. يمثل نسبة الحالات المصنفة خطأً إلى إجمالي عدد الحالات في مجموعة البيانات.

ذكر الباحثان (Hall and Wand) عام (1988) ان ايجاد افضل عرض الحزمة بشكل مباشر تكون عن طريق معدل خطأ التصنيف كالاتي :[8][9]

$$Pr(\ is\ classifed\ correcctly) = 1 - MR\ \}$$

$$= E_Y[1 - \{Y\ is\ classified\ correctly\}]\ \}$$

$$= E_X\left[E_Y[1\{Y\ is\ classified\ correctly\}]\right] /x_1, x_1, \dots x_v\}$$

$$= \frac{TP+TN}{TP+FP+TN+FN} \quad \dots (1)$$

اذ ان : E_Y التوقع بالنسبة للـ Y او $x_1, x_1, \dots x_v$ التوقع بالنسبة للـ E_Y ,$\sum_{j=1}^{v} \pi_j f_j$ او $\pi_1 f_1, \pi_2 f_2, \dots, \pi_v f_v$, TP (True Positive) من المتوقع أن تكون الملاحظة إيجابية وهي في الواقع إيجابية., TN (True Negative) من المتوقع أن تكون الملاحظة سلبية وهي في الواقع سلبية. FP (False Positive) من المتوقع أن تكون الملاحظة إيجابية وهي سلبية في الواقع. FN (False Positive) من المتوقع أن تكون الملاحظة سلبية وهي في الواقع إيجابية.[9]

5. التحليل التمييزي (Discriminant Analysis)

ويمز له اختصاراً (DA) وهو من ألاساليب المهمة في تحليل البيانات متعددة المتغيرات ، اذ أنه يعتمد على نوع المشكلة ونوع البيانات سواء كانت هذه البيانات (كمية او نوعية) وله عدة تطبيقات عملية مهمة ، اذ يستعمل في مختلف المجالات من أهمها الزراعية , الطبية كتصنيف الأمراض ومعرفة شدة الإصابة بها.

يعتمد اسلوب التحليل الاحصائي لمتعدد المتغيرات على الظواهر التي لها ابعاد ومتغيرات متعددة لوصفها وتحليلها. ان التحليل التمييزي يهتم بمسألة التمييز بين مجموعتين او اكثر والتي تشترك فيما بينها بمجموعة من الصفات والخصائص بدرجة مختلفة وذلك باستعمال دالة خاصة تسمى الدالة التمييزية (Discriminant Function). تتبعها عملية التصنيف وهي عملية تلي عملية تكوين الدالة التمييزية، إذ يتم الاعتماد على هذه الدالة في التنبوء وتصنيف المفردة الجديدة لإحدى المجموعات قيد الدراسة بأقل خطأ تصنيف ممكن.[6]

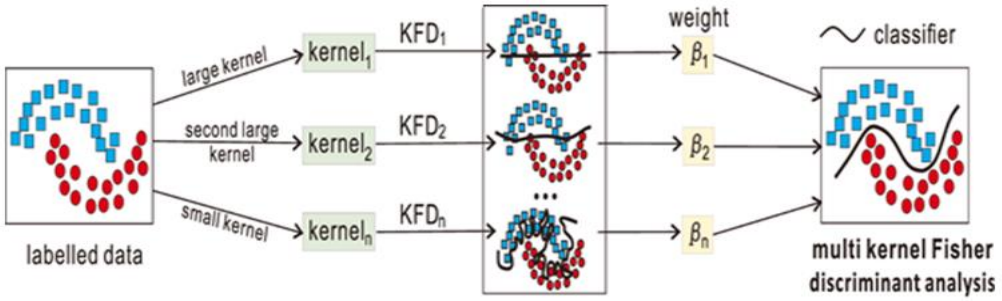
يعتمد نموذج التحليل التمييزي على الوصول إلى دالة التمايز التي تعمل على تعظيم الفروق بين متوسط المجموعات وتقليل التشابه في أخطاء التصنيف في الوقت ذاته، وذلك من خلال إيجاد تجميعات خطية لمجموعة من المتغيرات. كثيراً ما يطلق على المتغيرات الكمية في التحليل التمييزي متغيرات مستقلة أو منبئة،ويشار أيضاً لمتغير عضوية الجماعة بالمتغير التابع أو المتغير المحكي التصنيفي. والمتغيرات التابعة في التحليل التمييزي يطلق عليها متغيرات مستقلة في تحليل التباين، كما أن المتغيرات التابعة في تحليل التباين هي المتغيرات المستقلة في التحليل التمييزي ولا تحتاج بيانات التحليل التمييزي لأن تكون معيارية، أي أن يكون لها وسط صفري وتباين يساوي الواحد، وذلك لأن نتيجة تحليل التمييزي لا تتأثر بكثرة بتغيير مفردات المتغيرات[7].

6. التحليل التمييزي اللبي (Kernel Discriminant Analysis)

ان التحليل التمييز اللبي ((KDA) Kernel discriminant analysis) تقنية قوية لمهام التصنيف ، خاصة عند التعامل مع هياكل البيانات المعقدة وغير الخطية. ومع ذلك ، فإنه يتطلب التحديد الدقيق لوظيفة kernel وضبط المعلمة لتحقيق الأداء الأمثل. بالإضافة إلى ذلك ، يمكن أن تكون KDA مكلفة من الناحية الحسابية بالنسبة لمجموعات البيانات الكبيرة ، لأنها تتضمن حساب مصفوفة النواة والقيمة الذاتية. اذ يعد دالة تقدير الكثافة اللبية اسلوب مهم في تحليل البيانات التمييزية، اذ ان استعمال تقديرات الكثافة اللبية في التحليل التمييزي معروف تماماً من قبل الباحثين للتعرف على الانماط الاحصائية و يتم الاعتماد على تقديرات الكثافة اللبية لبناء طريقة التصنيف اللامعلمية ، و تسمى بالتحليل التمييزي اللبي (KDA).[2]

الفكرة الأساسية وراء (KDA) هي ايجاد فضاء فرعي مميز يزيد من التباين بين الفئة ويقلل من التباين داخل الفئة في مساحة الميزة ذات الأبعاد الأعلى الناتجة عن دالة اللب. فبذلك يمكن لـ (KDA) التعامل بفعالية مع البيانات غير القابلة للفصل الخطي وتحسين إمكانية فصل الفئات المختلفة.[4]

والشكل (1) يوضح مبدأ (KDA):



الشكل (1) توضيح التحليل التمييزي اللبي

ان اداء تقديرات الكثافة اللبية تعتمد بشكل كبير على الأختيار المناسب لحجم التمهيد وهي معلمة مصفوفة عرض الحزمة (H). اذ ان القيمة المثلى لعرض الحزمة هي التي جعل متوسط مربع الخـــــــطأ (Mean Sequare Erorr) اقل مايمكن ، إذ يكون التركيز الرئيسي على تقليل نسبة خطأ التصنيف ($\widehat{MR}$) ومن ناحية اخرى الطرائق الاساسية لأختيارمعلمة مصفوفة عرض الحزمة (H) التي تحاول تقليل نسبة خطأ التصنيف المقدرة .

الى جانب ذلك فأن مشكلة التصنيف سيكون اكثر اهمية ووضوحاً بان يكون معلمة عرض الحزمة مختلفة عند مقارنه مع مختلف دوال الكثافة المصنفة (اي وجود عدة مجموعات ولكل مجموعة توجد تقدير كثافة، وبعد اجراء التصنيف نقوم بمقارنة بين دوال الكثافة تسمى دوال الكثافة المصنفة) واختيارها بشكل امثل لتقدير الكثافة تعتمد بشكل كبير على الكثافة المصنفة ($\hat{f}_j$) و الاحتمال السابق لها ($\hat{\pi}_j$) ولذلك في مشكلة تعدد الطبقات (استعمال عدة متغيرات) .

علاوة على عن ذلك فان حسن اختيارمعلمة عرض الحزمة ينبغي ايضاً ان يعتمد على المشاهدة الخاصة لتصنيفها ونتيجة لذلك وعوضاً عن التركيز على عرض الحزمة مثلى الواحدة لتقدير الكثافة يكون اكثر فائدة ، اذاً تم الاعتماد على نتائج المستويات المختلفة من التمهيد لتقديردالة الكثافة اللبية .

لنفرض انه لدينا مجموعات مؤفلة من N من النقاط لكل منها متجه ميزات بـ n بعد أي أن:

$$X = \{x_1, x_2, \dots x_n\}$$

اذ أن $X_j \in R^d$ وان $Y = \{y_1, y_2, \dots y_n\}$ تمثل الاصناف المقابلة اذ ان $y_i \in \{1,2, \dots c\}$ والتي تمثل انتماء الصفة لكل مجموعة نقاط i^{th} .

اختيار دالة لبية معينة $k(x,x')$ والتي تعبر عن مقياس التشابه بين نقاط البيانات xx' في فضاء الميزة الاصلي ومنه الدوال الشائعة هي الدالة الغاوسية (*Gaussain*) والدالة متعددة الحدود (*Multinomial*) والدالة السينية (*Sigmoid*) ... الخ .

ومن ثم نحسب المصفوفة اللبية K كالآتي:

$$K(i,j) = k(x_i, x_j)$$

و هي مقياس التشابه بين النقط x_i, x_j

ومن ثم نقوم بتوسيط مصفوفة اللب K للتأكد من خاصية *Kernel Trick properity* وكالآتي:

$$H = I_N - \left(\frac{1}{N}\right) * 1_N * 1_N^T \quad \dots (2)$$

اذ ان: I_N مصفوفة الوحدة بالبعد N

1_N متجه عامودي بالواحدات بالبعد N فان مصفوفة اللب المركزية تحسب كالآتي:

$$KC = HKH'$$

اذ ان KC تمثل القيم الذاتية (*Eigen Values*) على مصفوفة اللب المركيزة للحصول على القيم الذاتية (λ) والمتجهات الذاتية المقابلة لها (a) ونقوم بترتيبها تنازلياً.

بعد ذلك نقوم باستخراج الميزات بتحديد المتجهات الذاتية k الأعلى المقابلة لأكبر متجه مميز لتشكيل فضاء فرعي جديد مخفض الأبعاد. وتمثل هذه المتجهات الذاتية $a_k \in R^N$, $a_1, a_2, \dots a_k$ الاتجهات في مساحة الميزة ذات الابعاد الاعلى التي تفصل بين الفئات بشكل افضل.

ثم نقوم بعملية التصنيف (*Classification*) لتنصيف نقاط البيانات الجديدة باسقاطها على مساحة الكيزة ذات الابعاد المخفضة باستعمال المتجهات الذاتية (*Eigen Vectors*) المحددة . ثم نطبق خوارزمية تصنيف متجه آلآت الدعم (*Support Vector Machine*) على مساحة الميزة المخفض لعمل تنبؤات.

تتمثل الخطوة الرئيسة في لانموذج الرياضي في استخدام دالة اللب للعمل ضمنًا في مساحة ميزة ذات أبعاد أعلى دون حساب متجهات الميزات المحولة وهذا يسمح لـ *KDA* بالتعامل مع البيانات غير الخطية القابلة للفصل وتحقيق أداء أفضل مقارنة بـ *LDA* في مثل هذه الحالات. يعد اختيار دالة اللب وعدد المتجهات الذاتية العليا (k) معلمات تشعبية أساسية تؤثر على أداء نموذج *KDA*. غالبًا ما يتم استخدام الضبط الصحيح والتحقق من الصحة لتحديد القيم المثلى لهذه المعلمات.[10]

7. التحليل التمييزي اللبي الحصين (Robust Kernel Discriminant Analysis)

يعد التحليل التمييزي اللبي الحصين (Robust Kernel Discriminant Analysis RKDA) امتدادًا للتحليل التمييزي اللبي (KDA)) الذي يهدف إلى تحسين أداء وحصانة KDA ، خاصة عند التعامل مع البيانات التي بها ضوضاء أو التالفة. تتمثل الفكرة الرئيسية وراء RKDA في تقديم الحصانة لخطوة تحويل البيانات التي تقوم بها KDA ، مما يسمح لها بالتعامل مع القيم الشاذة والعينات الضوضائية بشكل أكثر فعالية.[14][9]

يتضمن معيار KDA حساب مصفوفات لانتشار (داخل الطبقة وبين الطبقة) في مساحة النواة ، متبوعًا بحل مشكلة القيمة الذاتية المعممة للحصول على اتجاهات الإسقاط. ومع ذلك ، عند وجود قيم متطرفة أو عينات ضوضائية في البيانات ، يمكن أن تتأثر مصفوفات التبعثر بهذه النقاط الشاذة ، مما يؤدي إلى اتجاهات إسقاط دون المستوى الأمثل.[11]

في RKDA ، يتم استخدام متغير قوي لتقدير المصفوفة المبعثرة لتقليل تأثير القيم الشاذة. أحد الأساليب الشائعة هو استخدام تقدير مصفوفة التغاير الحصين ، مثل محدد التغاير الأدنى (Minimum Covaruance Determeniant) أو مقدر هوبر ، لحساب مصفوفات التبعثر. هذه التقديرات الحصينة أقل حساسية للقيم الشاذة ويمكن أن تنتج اتجاهات إسقاط أكثر موثوقية.

تتشابه الخطوات العامة للتحليل التمييزي الحصين مع KDA القياسي ، مع التعديل فقط في حساب مصفوفة التبعثر وكالآتي:

1. ادخال البيانات (Data Input) :
   X: البيانات الأصلية مع العينات الموجودة في الصفوف والمعالم في الأعمدة.
   Y: تسميات الفئة المقابلة لكل عينة في X.
2. دالة اللب (Kernel function):
   يتم دالة kernel مناسبة (على سبيل المثال ، Gaussian ، متعدد الحدود) تقيس التشابه بين عينتين في مساحة الميزة الأصلية.
3. مصفوفة اللب (Kernel Matrix) :
   يتم حساب مصفوفة اللب ، حيث تمثل K (i,j)، التشابه بين العينات X (i)، و X (j)
4. توسيط البيانات في مساحة اللب:
   تحسب مصفوفة التمركز H ، والتي تضمن أن البيانات في مساحة النواة لها متوسط صفري.
5. مصفوفات التشتت الحصينة:
   نستخدم تقدير مصفوفة التغاير الحصينة (Huber) لحساب مصفوفة التشتت داخل الفئة Sw والمصفوفة المبعثرة بين الفئة Sb في مساحة النواة باستخدام البيانات المركزية.
6. حل مشكلة القيمة الذاتية العمومية:
   نبحث عن القيم الذاتية والمتجهات الذاتية لمسألة القيمة الذاتية العمومية
   Sb * alpha = lambda * Sw * alpha.
7. حدد أعلى المتجهات الذاتية:
   يتم اختيار أعلى متجهات ذاتية لـ k المطابقة لأكبر قيم ذاتية لـ k لتشكيل مصفوفة الإسقاط.
8. مشروع البيانات:
   نقوم بإسقاط البيانات الأصلية X في مساحة الميزة الجديدة باستخدام مصفوفة الإسقاط W_rkda.
9. التصنيف:
   تطبيق مصنف (على سبيل المثال ، k-Nearest Neighbours ، دعم آلة المتجه (Support Vector Machine)) على البيانات المتوقعة من أجل التصنيف.

من خلال دمج تقدير مصفوفة التبعثر الحصينة ، يمكن لـ RKDA معالجة البيانات التي بها الضوضاء أو الملوثة بشكل أكثر فعالية من طريقة KDA القياسية . [13][9]

8. طريقة التحقق المتقاطع الممهد (Smoothed Cross –Validation)

تم تنفيذ طريقة التحقق المتقاطع التي اقترحها Stone بواسطة (Jaksa, Nejad, 2017) لتقسيم البيانات إلى ثلاث مجموعات: التدريب والاختبار والتحقق من الصحة. تستعمل مجموعة التدريب لضبط الاوزان ، بينما تستعمل مجموعة الاختبار للتحقق من أداء الانموذج في مراحل مختلفة من التدريب ولتحديد وقت إيقاف التدريب لتجنب فرط الملائمة. يتم استخدام مجموعة التحقق من الصحة لتقدير أداء الشبكة المدربة في البيئة المنتشرة. [12]

تهدف طرق التحقق المتقاطع إلى توفير تقديرات حصينة لأداء الانموذج من خلال تقسيم البيانات بطرائق مختلفة لتدريب النموذج والتحقق من صحته. وهي طريقة جديدة للتحقق من الصحة (Validity) وهي تقنية إعادة أخذ العينات المستعملة لتقييم أداء الانموذج التنبئي والتخفيف من مخاطر الملائمة المفرطة (Overfitting). يتضمن تقسيم مجموعة البيانات إلى مجموعات فرعية متعددة ، وتدريب النموذج على بعض هذه المجموعات الفرعية (مجموعات التدريب) ، ثم تقييم أدائها على المجموعة الفرعية المتبقية (مجموعة التحقق من الصحة). تتكرر العملية عدة مرات للحصول على متوسط مقياس الأداء.

تم تقديم طريقة التحقق المتقاطع لأحادي المتغير(SCV) من قبل الباحثون Hall , Marron and Park عام (1992) مع ممهد تجريبي افضل g المستقل عن h. اما بعد ذلك قدم ممهد SCV لمتعدد المتغيرات من قبل الباحثون (Sain , Baggerly) (and Scott (1994 . عمل بها في البداية من صيغة معدله بشكل قليل من طريقة LSCV في المعادلة (3) المعروفة باسم صيغة Leave – in – diagonal للحصول على عينات البيانات التي ليس لها قيم متكرره. [11]

$$Lscv(H) = n^{-1}(4\pi)^{-d/2}|H|^{-1/2} + n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}(K_{2H} - 2K_H + K_0)(X_i - X_j) \quad \ldots(3)$$

اذ ان :

$K_0 \rightarrow$ Dirac delta function

لتشكيل (SCV) قبل تمهيد فروقات البيانات ($X_i - X_j$) بواسطة K_{2G} , اي استبدال ($X_i - X_j$) بالإلتواء مع K_{2G} ($X_i - X_j$) :

$$SCV(H) = n^{-1}(4\pi)^{-d/2}|H|^{-1/2} + n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}(K_{2H+2G} - 2K_{H+2G} + K_{2G})(X_i - X_j) \quad \ldots(4)$$

اذ ان :

G: تمثل مصفوفة عرض الحزمة التجريبي

8. المحاكاة (Simulation)

تمت عملية بناء انموذج محاكاة لدراسة سلوك طرائق التحليل التمييزي المدروس باعتماد معايير معينة في تقدير النماذج المفترضة اذ تم اولا تحديد احجام العينات وتعد هذه الخطوة من الخطوات المهمة التي يعتمد عليها لتنفيذ باقي خطوات تجارب المحاكاة والتي تضمنت اختيار احجام عينات التدريب (Trianning Sample size) اذ تم تحديد احجام عينات التدريب الافتراضية لغرض اجراء طرائق التحليل التمييزي المستعملة في هذه الرسالة والتي هي n=100, 500, 1000, 5000 , وكذلك اختيار احجام عينات الاختبار (Test Sample size) اذ تم تحديد حجم عينة الاختبار الافتراضية لغرض اجراء طرائق التحليل التمييزي المستعملة في هذه الرسالة والتي هي k=1000, 2000, 3000, 5000 . وتم توليد العينات (Samples Generating) باستعمال مجوعتين لغرض اختبار طرائق التحليل التمييزي اذ تم توليد بيانات المجموعة الاولى والمجموعة الثانية بإستعمال طريقة (Box-Muller) بالاضافة إلى الدالة المكتبية (Randn) في برنامج ماتلاب اذ تستخدم لتوليد متغير عشوائي ثنائي يتبع التوزيع الطبيعي القياســـــــــي N_2 (0, 1)، وتعتمد طريقة (Box – Muller) على الأسلوب الآتي:

1. توليد عددين عشوائيين مستقلين U_1، U_2 بحيث يتبعان التوزيع المنتظم للفترة (1، 0) حيث يتم توليد متجه من هذين العددين بحجم العينة المطلوبة (n) أي إن:

$$U_i = rand(1, n) \quad \ldots(5)$$

2. يمكن تحويل هذين العددين إلى التوزيع الطبيعي القياسي وفقاً لما يأتي:

$$X_1 = (-2\ln(U_1))^{1/2}\cos(2\pi U_2) \quad \ldots(6)$$

$$X_2 = (-2\ln(U_1))^{1/2}\sin(2\pi U_2) \quad \ldots(7)$$

حيث أن X_1، X_2 متغيران عشوائيان طبيعيان مستقلان وبذلك فان الدالة المشتركة لهما هي:

$$f(x_1, x_2) = \frac{1}{2\pi}e^{-\frac{1}{2}(x_1^2 + x_2^2)} \quad \ldots(8)$$

3. تم توليد عدد من المتغيرات التي لها توزيع طبيعي قياسي متعدد متغيرات في كل مجموعة باستعمال الخوارزمية الآتية في ماتلاب:

1) تحديد حجم العينة (n) في كل متغير كالآتي:

```
numSamples = n;
```

2) تحديد عدد المتغيرات في كل مجموعة وكالآتي:

```
numVariables = K; % Change this to the desired number of variables
```

3) تحديد متجه المتوسطات الصفري بعدد المتغيرات المطلوبة وكالآتي:

```
mu = zeros(1, numVariables);
```

4) تحديد مصفوفة التبيان-التباين المشترك للتوزيع الطبيعي القياسي متعدد المتغيرات وكالآتي:

```
covMatrix = eye(numVariables);
```

5) توليد المتغيرات العشوائية القياسية الطبيعية المتعددة في كل مجموعة كالآتي:

```
X = mvnrnd(mu, covMatrix, numSamples);
```

وكذاك تم اختيار مصفوفة عرض الحزمة (Bandwidth Selection) باختيار مصفوفة عرض الحزمة القطرية باستعمال طريقة العبور الشرعي وايضاً اختيار دالة الكثافة اللبية الهدف (Target Kernel Density Selection) اذ تم اختيار اربعة دوال كثافة لبية لتحقيق صيغة التمييز اللبي البيزية مع توزيعتها الاولية الافتراضية بحيث تكون بعض من هذه الدوال الهدف تبتعد عن التوزيع الطبيعي لغرض انتهاك افتراض التوزيع الطبيعي لاختبار طرائق التقدير عند هذه الدوال الهدف وكالآتي:[9]

1) دالة الهدف الكاوسية D : والتي تكون كالآتي:

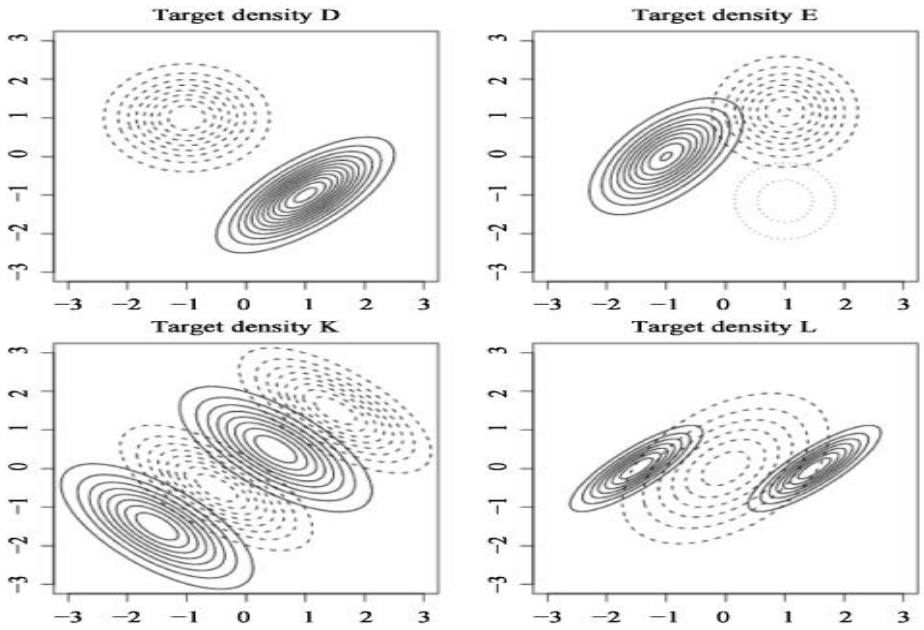
$$D=\begin{cases}\pi_1=\frac{1}{2}\ f_1\sim N_2\left(\begin{bmatrix}1\\-1\end{bmatrix},\begin{bmatrix}\frac{4}{9} & \frac{14}{45}\\ \frac{14}{45} & \frac{4}{9}\end{bmatrix}\right)\\ \pi_2=\frac{1}{2}\ f_2\sim N_2\left(\begin{bmatrix}-1\\1\end{bmatrix},\begin{bmatrix}\frac{4}{9} & 0\\ 0 & \frac{4}{9}\end{bmatrix}\right)\end{cases} \quad \ldots. (9)$$

2) دالة الهدف الكاوسية E: والتي تكون كالآتي:

$$E=\begin{cases}\pi_1=\frac{3}{7}\ f_1\sim N_2\left(\begin{bmatrix}-1\\0\end{bmatrix},\begin{bmatrix}\frac{9}{25} & \frac{63}{250}\\ \frac{63}{250} & \frac{49}{100}\end{bmatrix}\right)\\ \pi_2=\frac{3}{7}\ f_2\sim N_2\left(\begin{bmatrix}1\\\frac{2}{\sqrt{3}}\end{bmatrix},\begin{bmatrix}\frac{9}{25} & 0\\ 0 & \frac{49}{100}\end{bmatrix}\right)\end{cases} \quad \ldots. (10)$$

3) دالة الهدف L: والتي تكون كالآتي:

$$L=\begin{cases}\pi_1=\frac{1}{3}\ f_1\sim \frac{1}{2}N_2\left(\begin{bmatrix}-\frac{3}{2}\\0\end{bmatrix},\begin{bmatrix}\frac{3}{10} & \frac{1}{4}\\ \frac{1}{4} & \frac{3}{10}\end{bmatrix}\right)+\frac{1}{2}N_2\left(\begin{bmatrix}\frac{3}{2}\\0\end{bmatrix},\begin{bmatrix}\frac{3}{10} & \frac{1}{4}\\ \frac{1}{4} & \frac{3}{10}\end{bmatrix}\right)\\ \pi_2=\frac{2}{3}\ f_2\sim N_2\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}\frac{4}{5} & \frac{2}{5}\\ \frac{2}{5} & 1\end{bmatrix}\right)\end{cases} \quad \ldots. (11)$$



شكل (2) الرسم المحيطي (Contour Plot) دوال الكثافة الهدف (D, E, K, and L) [22]

تمت المقارنة بين طرائق التحليل التمييزي باستعمال المعدل والانحراف المعياري لخطأ التصنيف عند كل دالة كثافة لبية والذي صيغته كالآتي:

$$\widehat{MR}= 1- m^{-1} \sum_{j=1}^{m} 1 \{ Y j \text{ is correctly classified using KDA} \} \quad \dots (12)$$

ومن ثم تحديد أفضل أسلوب من بين الأساليب المستعملة في التقدير التي تمتلك أقل نسبة الخطأ ($\widehat{MR}$).

وكانت نتائج تحليل تجارب المحاكاة كما في الجداول التالية:

جدول (1) المعدل والانحراف المعياري لخطأ التصنيف وفق أساليب التحليل التمييزي للانموذج الأول

Target Density	n	k	method	KDA	RKDA
D	100	1000	Mean	0.19092	0.22343
			STD	0.00211	0.00115
E			Mean	0.15322	0.21134
			STD	0.00115	0.00132
K			Mean	0.05462	0.00462
			STD	0.00143	0.00111
L			Mean	0.00835	0.00342
			STD	0.00127	0.00045
Target Density	**n**	**k**	**method**	**KDA**	**RKDA**
D	500	1000	Mean	0.17992	0.21243
			STD	0.00689	0.00685
E			Mean	0.14222	0.20022
			STD	0.00785	0.00761
K			Mean	0.00533	0.04362
			STD	0.00152	0.00142
L			Mean	0.00265	0.00258
			STD	0.00973	0.00151
Target Density	**n**	**k**	**method**	**KDA**	**RKDA**
D	1000	1000	Mean	0.01131	0.11845
			STD	0.00343	0.01433
E			Mean	0.05534	0.11324
			STD	0.00114	0.01146
K			Mean	0.00335	0.00226
			STD	0.00131	0.00245
L			Mean	0.00167	0.00118
			STD	0.00253	0.00103
Target Density	**n**	**k**	**method**	**KDA**	**RKDA**
D	5000	1000	Mean	0.01969	0.08745
			STD	0.00757	0.03667
E			Mean	0.00431	0.08224
			STD	0.02922	0.01954
K			Mean	0.00662	0.00271
			STD	0.00961	0.00151
L			Mean	0.02933	0.00024
			STD	0.02847	0.00191

جدول (2) المعدل والانحراف المعياري لخطأ التصنيف وفق أساليب التحليل التمييزي للانموذج الثاني

Target Density	n	k	method	KDA	RKDA
D	100	2000	Mean	0.17941	0.23424
			STD	0.00789	0.04983
E			Mean	0.11222	0.19033
			STD	0.00482	0.00662

K			Mean	0.04362	0.00638
			STD	0.00657	0.00289
L			Mean	0.00265	0.00252
			STD	0.00273	0.00055
Target Density	n	k	method	KDA	RKDA
D	500	2000	Mean	0.15841	0.21324
			STD	0.01311	0.02883
E			Mean	0.09122	0.16933
			STD	0.01618	0.01438
K			Mean	0.02262	0.01462
			STD	0.01443	0.01811
L			Mean	0.01535	0.01482
			STD	0.01827	0.02045
Target Density	n	k	method	KDA	RKDA
D	1000	2000	Mean	0.03538	0.15128
			STD	0.00255	0.03318
E			Mean	0.02923	0.03242
			STD	0.03618	0.00231
K			Mean	0.02738	0.02441
			STD	0.00443	0.00231
L			Mean	0.03537	0.01487
			STD	0.03827	0.01042
Target Density	n	k	method	KDA	RKDA
D	5000	2000	Mean	0.17631	0.49228
			STD	0.34355	0.37418
E			Mean	0.11024	0.37342
			STD	0.21711	0.34331
K			Mean	0.36838	0.12541
			STD	0.34543	0.24331
L			Mean	0.37637	0.05581
			STD	0.37927	0.25142

جدول (3) المعدل والانحراف المعياري لخطأ التصنيف وفق أساليب التحليل التمييزي للانموذج الثالث

Target Density	n	k	method	KDA	RKDA
D	100	3000	Mean	0.13841	0.19324
			STD	0.03311	0.00883
E			Mean	0.07122	0.14933
			STD	0.03618	0.03438
K			Mean	0.00262	0.01461
			STD	0.03443	0.03811
L			Mean	0.03835	0.00148
			STD	0.00273	0.00051
Target Density	n	k	method	KDA	RKDA
D	500	3000	Mean	0.09741	0.15224
			STD	0.00789	0.03217
E			Mean	0.03022	0.10833
			STD	0.00482	0.00662
K			Mean	0.03838	0.02639
			STD	0.00657	0.00289
L			Mean	0.00265	0.00151
			STD	0.00827	0.00049

Target Density	n	k	method	KDA	RKDA
D	1000	3000	Mean	0.05641	0.11124
			STD	0.03311	0.00883
E			Mean	0.01078	0.06733
			STD	0.03618	0.03438
K			Mean	0.00262	0.00069
			STD	0.03443	0.00935
L			Mean	0.02394	0.00018
			STD	0.00466	0.00327
Target Density	n	k	method	KDA	RKDA
D	5000	3000	Mean	0.00541	0.07024
			STD	0.00189	0.03217
E			Mean	0.00022	0.02633
			STD	0.00112	0.00662
K			Mean	0.03838	0.00031
			STD	0.00657	0.00165
L			Mean	0.01706	0.00012
			STD	0.03634	0.00103

جدول (4) المعدل والانحراف المعياري لخطأ التصنيف وفق أساليب التحليل التمييزي للانموذج الرابع

Target Density	n	k	method	KDA	RKDA
D	100	5000	Mean	0.16992	0.20243
			STD	0.01889	0.01985
E			Mean	0.13222	0.19034
			STD	0.01985	0.01968
K			Mean	0.03362	0.01638
			STD	0.01957	0.01989
L			Mean	0.01265	0.01058
			STD	0.01973	0.02055
Target Density	n	k	method	KDA	RKDA
D	500	5000	Mean	0.14892	0.18143
			STD	0.00211	0.00115
E			Mean	0.11122	0.16934
			STD	0.00115	0.00132
K			Mean	0.01262	0.00462
			STD	0.00143	0.00111
L			Mean	0.00835	0.00042
			STD	0.00127	0.00045
Target Density	n	k	method	KDA	RKDA
D	1000	5000	Mean	0.00207	0.04924
			STD	0.01771	0.01117
E			Mean	0.00124	0.00533
			STD	0.00588	0.01438
K			Mean	0.01738	0.00069
			STD	0.01443	0.01935
L			Mean	0.00394	0.00018
			STD	0.01534	0.00394
Target Density	n	k	method	KDA	RKDA
D	5000	5000	Mean	0.00541	0.02824
			STD	0.00189	0.00983
E			Mean	0.00552	0.01567
			STD	0.00467	0.00662

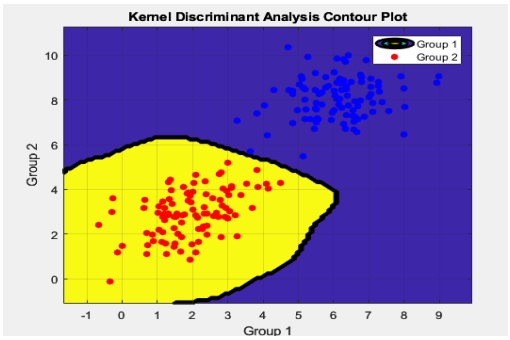
K			Mean	0.02031	0.00411
			STD	0.00165	0.00657
L			Mean	0.00265	0.00112
			STD	0.00827	0.00049

ولبيان افضلية اساليب التحليل التمييزي ندرج ادناه جدول الأفضلية الذي يبين عدد مرات الافضلية ونسب الافضلية لكل اسلوب وعند كل حجم عينة تدريب وعينة اختبار

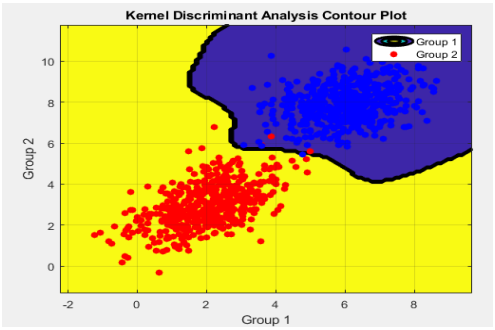
جدول (5) عدد مرات الافضلية ونسب الافضلية لكل اسلوب وعند كل دالة هدف

Target density	No.	Ratio	No.	Ratio
	KDA		RKDA	
D	7	11	0	0
E	6	9	1	2
K	3	5	15	23
L	0	0	16	25

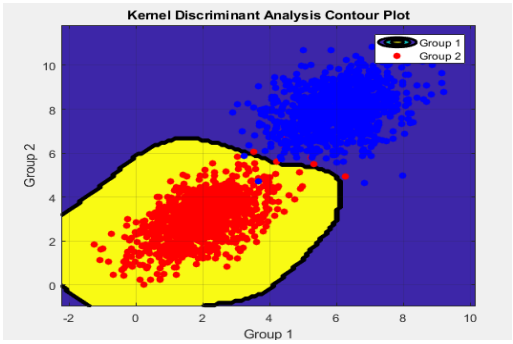
يتبين من جداول نتائج المحاكاة ان اسلوب التحليل التمييزي اللبي حقق افضلية على التحليل التمييزي اللبي الحصين عند دالة الكثافة (K) بنسبة قليلة بلغت (5%). وان اسلوب التحليل التمييزي اللبي الحصين حقق افضلية على باقي الاساليب عند دوال الكثافة المنحرفة عن التوزيع الطبيعي بنسبة افضلية (23) و (25) لكل دالة كثافة على التوالي.



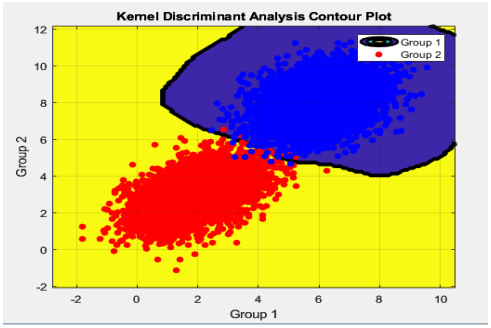
شكل (3) التصنيف وفق التحليل التمييزي اللبي عندما n=100, k=1000



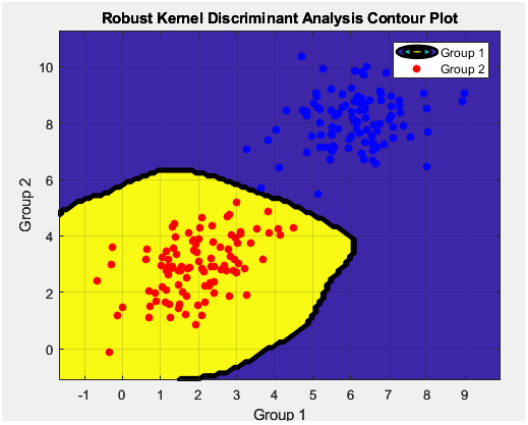
شكل (4) التصنيف وفق التحليل التمييزي اللبي عندما n=500, k=1000



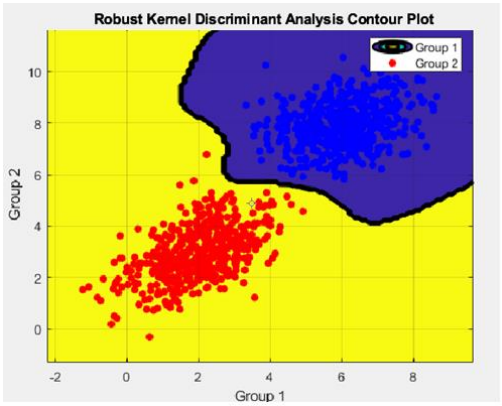
شكل (5) التصنيف وفق التحليل التمييزي اللبي عندما n=1000, k=1000



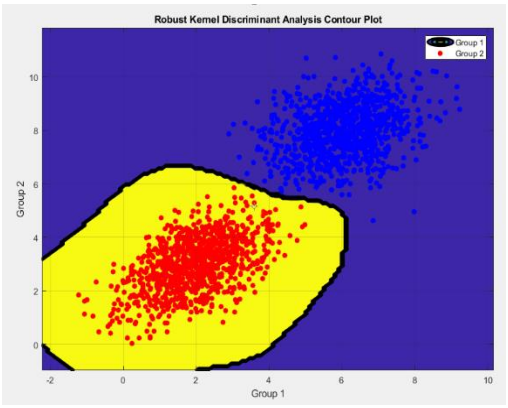
شكل (6) التصنيف وفق التحليل التمييزي اللبي عندما n=5000, k=1000



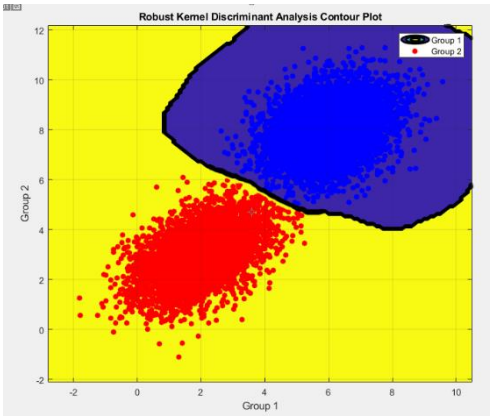
شكل (7) التصنيف وفق التحليل التمييزي اللبي الحصين عندما n=100, k=1000



شكل (8) التصنيف وفق التحليل التمييزي اللبي الحصين عندما n=500, k=1000



شكل (7) التصنيف وفق التحليل التمييزي اللبي الحصين عندما n=1000, k=1000



شكل (8) التصنيف وفق التحليل التمييزي اللبي الحصين عندما n=5000, k=1000

9. الإستنتاجات (Conclusions)

بالاعتماد على ماتوصلنا اليه من نتائج تم ادراج الاستنتاجات الآتية:

1- حقق اسلوب التحليل التمييزي اللبي افضلية عند دول الكثافة (D, E) عند حجم العينة (n=1000, 5000) .
2- حقق اسلوب التحليل التمييزي اللبي افضلية على باقي الاساليب عند دالة الكثافة (K) بنسبة قليلة.
3- حقق اسلوب التحليل التمييزي اللبي الحصين افضلية على باقي الاساليب عند دوال الكثافة المنحرفة عن التوزيع الطبيعي بنسبة افضلية عالية.

10. التوصيات (Recommendations)

من خلال ما تم التوصل اليه من استنتاجات ندرج التوصيات الآتية:

1- ضرورة استعمال اسلوب التحليل التمييزي اللبي الحصين في حالة كون البيانات تبتعد عن التوزيع الطبيعي او توجد قيم شاذة ضمنها.
2- استعمال التحليل التمييزي البيزي في حالة كون البيانات تحتوي على قيم شاذة
3- استعمال دوال كثافة غير الكاوسية مثل دالة ايبانكتشوف لتطبيق اسلوب التحليل التمييزي اللبي الحصين.
4- استعمال عرض الحزمة القطرية غير طريقة طريقة العبور الشرعي كان تكون طريقة متجه آلات الدعم (Support Vector Machine)

المصادر:

1. Sakina Shamil Jassim , (2020) , " A comparative study between the linear discriminant loading method and the pulp discriminant loading method " lraqi Journal of Administrative Sciences , Volume (14) , lssue(55) .

2. Alaa lmad Hamid , Basil Khader Daoud ; Basil Younis Thanoun , (2009) , " A statistical study of free evaporation in the mosul region using the core estimator mothed " , Scientific Abstracts Database , Mosul Al-Rafidain Engineering Journal , Vol.17 No. 5 17

3. Abdel Rahim Awad Abdel Khaleq Bassiouni , (2021) , " Using discriminant analysis in classification and commerce and Finance " , Volume 41 , lssue 3 , September 2021 page 297-325

4. Liqaa Ali Muhammad , Amir Ali Abboud , (2020) , " (Comparison of bandwidth estimators (smoothing parameter) using leptospheric functions in analyzing major components) " Heritage University College Journal , issue twentieth 412

5. Gupt , Abhishek M.; Soni , Himanshu H.; Joshi , Raunak M.; Laban, Ronald Melwin, (2022), " DISCRIMINANT ANALYSIS IN CONTRASTING DIMENSIONS FOR POLYCYSTIC OVARY SYNDROME PROGNOSTICATION", A PREPRINT - JANUARY 11,

6. Obudho , Macdonald G., Orwa , George O., Otieno, Romanus O., Were, Festus A. , (2022), "Robust Classification through a Nonparametric Kernel Discriminant Analysis", Open Journal of Statistics, 2022, 12, 443-455 https://www.scirp.org/journal/ojs ISSN Online: 2161-7198 ISSN Print: 2161-718X.2

7. Li , Quanbao, Wei, Fajie, ; Zhou, Shenghan , (2017), " Local kernel nonparametric discriminant analysis for adaptive extraction of complex structures ", Open Phys. 2017; 15:270–279. DOI 10.1515/phys-2017-00303

8. Cai, L., Liu, Y., & Liu, H. (2019). Discriminant Analysis on Riemannian Manifold for Hyperspectral Image Classification. IEEE Transactions on Geoscience and Remote Sensing, 57(7), 4562-4576. (Discusses Discriminant Analysis applied to hyperspectral image classification11

9. Li-Pang Chen, (2022), " Classification and prediction for multi-cancer data with ultrahigh-dimensional gene expressions ", PLOS ONE.

10. Zhifeng Li, Xiaoou Tang, (2009), " Nonparametric Discriminant Analysis for Face Recognition ", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 31, NO. 4, APRIL 2009

11. David Anthony Mercer, Nonparametric Discriminant Analysis in Forensic Ancestry Estimation: An Assessment of Utilized and Alternative Statistical Methods", TRACE: Tennessee Research and Creative Exchange.

12. Obudho, M. , Orwa, G. , Otieno, R. and Were, F. (2022) Robust Classification through a Nonparametric Kernel Discriminant Analysis. Open Journal of Statistics, 12, 443-455. doi: 10.4236/ojs.2022.124028.

13. Seung-Jean Kim Alessandro Magnani Stephen P. Boyd, (2023), (Robust Fisher Discriminant Analysis), Advances in Neural Information Processing Systems 18, p659-666.

14. Wand, M. P., & Marron, J. S. (1991). "Ridge Estimators for the Smoothing of Differenced Data." Journal of Time Series Analysis, 12(2), 97-112.