استخدام المقدرات الحصينة في الدوال التمييزية

أ.م.د. خلود يوسف خمو
كلية الإدارة والاقتصاد/ جامعة بغداد

م.م. سرى صباح كتيب
جامعة المأمون

Use Robust estimators in discriminate functions

Assis. Proof. Dr. Khlood Yousef Hamo
College of Admin. & Eco. Baghdad Uni.

Assis. Lec. Sura sabah ktiteb
Uni. Of Al. Mamon

تاريخ استلام البحث ٢٠١٥/٦/١٥ تاريخ قبول النشر ٢٠١٥/١١/٢٣

الملخص:

إنَّ دوال التمييز التقليدية قد درست من قبل العديد من الباحثين ووجدوا إنها أقل حصانة عند عدم تحقق بعض الفرضيات المتعلقة بدوال التمييز ولا سيما مع وجود القيم الشاذة، وقام الباحثون بالتركيز على مدى عشر سنوات على مقدرات حصينة كانت صعبة الحساب سابقاً لكن مع التطور في صناعة الحواسيب من الناحيتين المادية والبرمجية أمكن حساب مقدرات حصينة والتي تزودنا بأسلوب جديد في التعامل مع البيانات مقارنة مع المقدرات التقليدية، وكانت فكرة البحث استخدام مقدرات حصينة مقاومة لتأثير الشواذ كمقدر H الحصين، مقدر S الحصين ومقدر MCD الحصين كذلك تحصين احتمال خطا التصنيف مع بيان تأثير الشواذ على نسبة خطأ التصنيف عند استخدام الطرائق التقليدية، ومن ضمن أهداف البحث مقارنة المقدرات للتعرف على أفضل مقدر والذي يعطي أقل احتمال لخطأ التصنيف ولاسيما مع تنوع نسب التلويث وأحجام عينات مختلفة.

الكلمات المفتاحية: مقدرات حصينة ، دوال تمييزية.

Abstract:

The classical discriminate functions have been taught by many researchers, and they found that it is less robustness when some assumptions related with discriminate functions are not achieved specially with the outliers existence, thus the researchers for the last ten years concentrated on robust estimators which was difficult to compute before. But with the development in manufacturing computes from both hardware and software sides, complicated robust estimators became computable and it gives us a new way of dealing with the data compared to the classical estimators . So, the idea of the research is to use robust estimators which are resistant to the outlier influence like robust H estimator, robust S estimator and robust MCD estimator ,and also robust misclassification probability with showing outlier influence on the percentage of misclassification when using classical methods, some of the research aims are to compare estimators to find the best estimator which can give minimum probability of misclassification especially with the variety of contamination percentage and different sample sizes and the data contaminated.

Key words: estimator, MCD estimators ,misclassification ,robustness
Use Robust estimators in discriminate functions.

1 . المقدمة:

إنَّ التحليل المميز يصف المجتمعات من خلال الكشف عن الاختلافات الموجودة بين مفردات مجتمع معين وان مسالة تصنيف مفردة واحدة إلى أحد المجتمعات يدعى بأسلوب التمييز، وهناك استخدامان للتمييز الأول: وصف وتقدير البيانات من خلال تحديد المضامين التشخيصية للبيانات، والثاني التنبؤ والذي يزودنا بتقدير شامل لكفاءة قواعد التصنيف.

قد درست دوال التمييز التقليدية من قبل العديد من الباحثين ووجد انها أقل حصانة عند عدم تحقق بعض الفرضيات المتعلقة بدوال التمييز بوجود القيم الشاذة، لذا قام الباحثون بالتركيز في عشر سنوات الأخيرة على مقدرات حصينة لتحصين دوال التمييز إذْ تزودنا المقدرات الحصينة بأسلوب جديد في التعامل مع البيانات مقارنة مع المقدرات التقليدية المعروفة.

2.1 هدف البحث:

نظراً لإخفاق طرائق التحليل المميز التقليدية في تحقيق خواصها المثلى عند احتواء البيانات على نسب من الشواذ وبالتالي إخفاقها في الوصول على أقل احتمال لخطأ التصنيف، لذا كانت فكرة البحث استخدام مقدرات حصينة مقاومة لتأثير الشواذ ومن تلك المقدرات مقدر H الحصين، مقدر S الحصين ومقدر MCD الحصين.

وتم تحصين احتمال خطأ التصنيف والذي استخدم لأول مرة على مستوى القطر مع بيان تأثير الشواذ على نسب خطأ التصنيف عند استخدام الطرق التقليدية، ومن ضمن أهداف البحث مقارنة المقدرات للتعرف على أفضل مقدر والذي يعطي أقل احتمال لخطأ التصنيف لاسيما مع تنوع نسب التلويث وإحجام العينات المختلفة بإتباع أسلوب المحاكاة إذْ تم تلويث البيانات بأسلوب لم يتطرق إليه أي بحث على مستوى القطر.

2. الجانب النظري:

1.2 مقدر Huber الحصين [5,7,8,12]

يعتمد هذا المقدر على موازنة المشاهدات لتقليل تأثير الشواذ وانَّ عائلة M تعتمد على تقديرات الأوساط الحسابية والتباينات الموزونة ومن هذه الطرائق طريقة Huber ويمكن توضيح الطريقة كما يلي:

1. تحسب مسافة مهلنوبس لكل مشاهدة وكالآتي:

$$D(\underline{x}_i, x) = [(\underline{x}_i - \underline{T}(x))' S^{-1}(x)(\underline{x}_i - \underline{T}(x))]^{1/2} \quad \ldots\ldots (1)$$

وان

$$T(x)=\frac{1}{n}\sum_{i=1}^{n}x_i \qquad \dots (2)$$

$$S(x)=\frac{1}{n}\sum_{i=1}^{n}(x_i-T(x))(x_i-T(x))' \qquad \dots (3)$$

2. لتشذيب البيانات يعاد حساب المتوسطات ومصفوفة التباينات إذْ تبدل $T(x)$ و $S(x)$ بأوزان تعتمد على d_i اي

$$T^*(x)=\frac{\sum w_i x_i}{\sum w_i} \qquad \dots (4)$$

$$S^*(x)=\frac{\sum w_i (x_i-T^*(x))(x_i-T^*(x))'}{\sum w_i^2} \qquad \dots (5)$$

وان

$$w_i=\begin{cases} 2d_i & \text{if } d_i>2 \\ 1 & \text{if } d_i\le 2 \end{cases}$$

ثم يعاد حساب الخطوة (2) بصورة تكرارية بالاعتماد على نتائج التكرار السابق ونتوقف عندما يكون الفرق بين تكرارين متعاقبين قليلاً وفق مستوى معين من الدقة، إذْ يتم الحصول على بيانات مشذبة لكافة المجاميع، ويتم ضرب القيم بالتكرار الأخير بالمتغيرات كافة ولكافة المجاميع عندها يتم تطبيق الدالة التمييزية الخطية والتربيعية.

2.2 مقدر S الحصين:

هناك طرائق مختلفة لبناء قاعدة التمييز أي تحديد المشاهدة الجديدة x هل تنتمي إلى p_1 أو p_2 ؟ والطريقة الشائعة قاعدة التمييز الخطية لفشر Fisher's Linear Discriminant Rule والتي تندرج تحت الفرضية $\Sigma_1=\Sigma_2=\Sigma$ وان [3,11]

$$A(x)=x'\lambda-k \qquad \dots(6)$$

إذْ

$$\lambda=\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2) \qquad \dots(7)$$

$$k=\left(\frac{\underline{\mu}_1+\underline{\mu}_2}{2}\right)\underline{\lambda}$$

وتؤشر $x\in R^p$ التي تأتي من المجتمع الأول إذا كان

$$A(x)>\ln\left(\frac{C_2\pi_2}{C_1\pi_1}\right)=\tau \qquad \dots (8)$$

إذْ C_1 و C_2 كلفة خطأ التصنيف للمجموعة (1) والمجموعة (2).

π_1 و π_2 الاحتمال السابق للمجتمع (1) والمجتمع (2) x تعود للمجتمع الأول p_1.

وان $\pi_2 = 1-\pi_1$ وفي حالة التطبيق فإنَّ π_1,π_2,C_1,C_2 غير معروفة ولهذا توضع $\tau = 0$ وبإعادة كتابة (6).

$$A(x) = \frac{1}{2}[d_{\mu_2,\Sigma}{}^2(x) - d_{\mu_1,\Sigma}{}^2(x)] \quad \dots (9)$$

إذْ

$$d^2{}_{\mu_j,\Sigma}(x) = (\underline{x} - \underline{\mu}_j)'\Sigma^{-1}(\underline{x} - \underline{\mu}_j) \quad \dots(10)$$

هي مسافة مهلنوبس بين x و μ_j .

واضح ان التقدير التقليدي (متجه المتوسط ومصفوفة التباين المشترك المدمجة) تتاثر بقوة بالشواذ فيكون التقدير A(x) غير حصين وبقوة وتم تقليل تأثير الشواذ بواسطة مقدرات M للمتوسط والتباين المشترك لكل مجتمع، ولكن ضعف مقدر M هو في نقطة الانهيار Breakdown point والتي هي كسر عالي من الشواذ إذْ المقدر يمكن أنْ يقاوم بضعف الأبعاد العالية، لذا تم استخدام مقدر ذو انهيار عالٍ يعتمد على مقدر أصغر قطع بيضوي فالعينة $x_1,\dots,x_n$ تعرف كثنائي (t_n, C_n) والتي تقلل محددة C تحت القيد الاتي:

$$\frac{1}{n}\sum_{i=1}^{n}\rho[\sqrt{(\underline{x}_i - \underline{t})'C^{-1}(\underline{x}_i - \underline{t})}] = b \quad \dots (11)$$

وتحت كل الأزواج الممكنة (t,C) إذْ $t \in R^p$ و C هي مصفوفة متماثلة موجبة التحديد positive definite $p \times p$ ، والدالة ρ يجب أنْ تكون متماثلة مستقرة وغير متناقصة في الفترة $[0,\infty)$، ولها مشتقة مستمرة في عدد محدود من النقاط وان $\rho(0) = 0$ والثابت b يوضح بالشكل $E_{f0}[\rho(\|Z\|)]$ حيث تمثل F_0 توزيع الأنموذج المركزي وللحصول على الاتساق عند التوزيع الطبيعي ناخذ $F_0 = N(0, I_p)$ان الاختيار الشائع إلى ρ هو دالة Biweight:

$$\rho(u) = \min\left(\frac{u^2}{2} - \frac{u^4}{2c^2} + \frac{u^6}{6c^4}, \frac{c^2}{6}\right) \quad \dots(12)$$

إذْ c ثابت يحقق قيمة مركزية لنقطة الانهيار ويسمى مقدر (Biweight estimator).

3.2 مقدر MCD الحصين: [6]

يعرف بأنَّه أصغر محددة تباين مشترك Minimum Covariance Determinant كمتوسط $\hat{\mu}_{j,0}$ ومصفوفة تباين مشترك $\hat{\Sigma}_{j,0}$ لـ h_j من المشاهدات من n_j والتي تكون مصفوفة التباين المشترك اقل مايمكن إذْ إنَّ:

$$h_j > [(n_j + P + 1)/2] \quad \dots (13)$$

وان $n_j - h$ يجب أنْ تَكُونَ أقل من عدد الشواذ في j من المجتمعات وبسبب كون العدد غير معروف في كل مجموعة فتؤخذ:

$$h_j = [(n_j + P + 1)/2] \approx 50\% \quad \text{... (14)}$$

ويوصى بأخذ $h_j \approx 0.75 n_j$.

وبالاعتماد على التقديرات الأولية $\hat{\mu}_{j,0}$ و $S_{j,0}$ لكل مشاهدة x_{ij} وللمجموعة j تحسب المسافة الحصينة.

$$RD_{ij}^0 = \sqrt{(\underline{x}_{ij} - \underline{\hat{\mu}}_{j,0})' S_{j,0}^{-1} (\underline{x}_{ij} - \underline{\hat{\mu}}_{j,0})} \quad \text{... (15)}$$

ويعطي الوزن لـ x_i كالآتي:

$$RD_{ij}^0 = \begin{cases} 1 & if \quad RD_{ij}^0 \le \sqrt{\chi^2_{p,0.975}} \\ 0 & otherwise \end{cases} \quad \text{... (16)}$$

ومن ثم نحصل على مقدر MCD للمجموعة j كمتوسط $\hat{\mu}_{j,MCD}$ ومصفوفة تباين مشترك $\hat{\Sigma}_{j,MCD}$ لمشاهدات المجموعة j والتقديرات الحصينة للموقع والتشتت تسمح لنا برصف الشواذ في البيانات، وتعطي تقديرات أكثر حصانة لاحتمال خطأ التصنيف ثم تحسب لكل مشاهدة x_{ij} بالمجموعة j المسافة الحصينة النهائية كالآتي:

$$RD_{ij} = \sqrt{(x_{ij} - \hat{\mu}_{j,MCD})' \hat{\Sigma}_{j,MCD}^{-1} (\underline{x}_{ij} - \underline{\hat{\mu}}_{j,MCD})} \quad \text{... (17)}$$

وان x_{ij} قيمة شاذة إذا وفقط إذا $RD_{ij} > \sqrt{\chi_{p,0.975}^2}$ وبفرض ان $\tilde{n}_j$ تشير الى عدد غير الشواذ في المجموعة j وان $\tilde{n} = \sum_{j=1}^{g} \tilde{n}_j$ ويمكن تقدير احتمال خطأ التصنيف بحصانة كالآتي:

$$\hat{P}_j^R = \frac{\tilde{n}_j}{\tilde{n}} \quad \text{... (18)}$$

4.2 قواعد التمييز الخطية الحصينة: [6]

بإعطاء التقدير الحصين لمراكز المجموعة $\hat{\mu}_j$ ومصفوفة التباين المشترك $\hat{\Sigma}_j$ فإنَّ قاعدة التمييز الخطية الحصينة RLDR(Robust Linear Discriminant Rule) كالآتي:

تعيين x إلى π_k إذا:

$$\hat{d}_k^{RL}(x) > \hat{d}_j^{RL}(x) \quad ,\forall j = 1,...,g, \quad j \neq k \quad \text{... (19)}$$

وإنَّ:

$$\hat{d}_j^{RL}(x) = \hat{d}_j^{RL}(\underline{x}, \underline{\hat{\mu}}_j, \hat{\Sigma}) = \underline{\hat{\mu}}'_j \hat{\Sigma}^{-1} x - \frac{1}{2} \underline{\hat{\mu}}'_j \hat{\Sigma}^{-1} \underline{\hat{\mu}}_j + \ln(\hat{P}_j^R) \quad \text{... (20)}$$

وإنَّ احتمال $\hat{P}_j^R$ يمكن أنْ يقدر كما في (18) وإلى g=2 فإنَّ قاعدة التصنيف للصيغة (20) هي قاعدة التمييز لفشر للعينة يمكن أنْ توصف كالآتي:

$$\begin{cases} x \in \pi_1 & if \ (\underline{\hat{\mu}}_1 - \underline{\hat{\mu}}_2)' \hat{\Sigma}^{-1} (x - (\underline{\hat{\mu}}_1 + \underline{\hat{\mu}}_2)/2) > 0 \\ x \in \pi_2 & otherwise \end{cases}$$

ولبناء قاعدة التمييز الخطية الحصينة RLDR بعد أنْ ننظر إلى المقدرات الأولية لمتوسط المجاميع ومصفوفة التباين المشترك والتي يرمز لها بـ $\hat{\mu}_{j,0}$ و $\hat{\Sigma}_0$ وبالاعتماد على $\hat{d}_j^{RL}(x,\hat{\mu}_{j,0},\hat{\Sigma}_0)$ فإنَّ المسافة الحصينة.

$$RD_{ij}^0 = \sqrt{(\underline{x}_{ij} - \underline{\hat{\mu}}_{j,0})' \hat{\Sigma}_0^{-1} (\underline{x}_{ij} - \underline{\hat{\mu}}_{j,0})} \quad \dots (21)$$

ولكل مشاهدة في المجموعة j نفرض

$$w_{ij} = \begin{cases} 1 & if \quad RD_{ij}^0 \le \sqrt{\chi_{p,0.975}^2} \\ 0 & otherwise \end{cases}$$

وللحصول على تقديرات التباين المشترك الأولية $\hat{\Sigma}_0$ هناك ثلاث طرق مختلفة وهي:

الطريقة الأولى [3,6]

طبقت لأول مرة من قبل Chork & Rousseeuw (1992) باستخدام مقدر MVE و Croux & Dehon (2001) باستخدام مقدر S فبعد الحصول على تقديرات MCD وهي $\hat{\mu}_{j,MCD}$ و $\hat{\Sigma}_{j,MCD}$ لكل مجموعة فإنَّ مصفوفة التباين المشترك المدمجة PCOV هي:

$$\Sigma_{PCOV} = \frac{\sum_{j=1}^{g} n_j \hat{\Sigma}_{j,MCD}}{\sum_{j=1}^{g} n_j} \quad \dots (22)$$

الطريقة الثانية: [5,7,8,12]

فكرتها تعتمد على دمج المشاهدات POBS من دون مصفوفة التباين المشترك للمجموعة وهي تحديث للطريقة المعطاة من قبل He & Fung (2000) والذين استخدموا مقدر S لدالة التمييز الخطية الحصينة ففي حالة العينتين $x_{11}, x_{21}, \dots, x_{n1,1}, x_{12}, x_{22}, \dots, x_{n2,2}$ ومن مجتمعين متوسطهما μ_1 و μ_2 ومصفوفة التباين المشترك Σ.

أولاً: نقدر μ_1 و μ_2 كتقدير الموقع للمجموعتين باستخدام مقدر MCD ذات الأوزان المعادة.

ثانياً: تدمج مراكز المشاهدات لتصبح بالشكل Z's كالآتي:

$$(z_1, \dots, z_{n1+n2}) = (x_{11} - \hat{\mu}_1, \dots, x_{n1,1} - \hat{\mu}_1, x_{12} - \hat{\mu}_2, \dots, x_{n2,2} - \hat{\mu}_2)$$

ثالثاً: نقدر مصفوفة التباين Σ باستخدام مقدر MCD ذات الأوزان المعادة لـ z's التقدير الجديد لـ μ_1 يصبح $\hat{\mu}_1 + \hat{\delta}$ وكذلك لـ μ_2 يصبح $\hat{\mu}_2 + \hat{\delta}$ تنجز هذه الخطوات التكرارية عدة مرات، لكن في حالة مقدر MCD يتطلب خطوات إضافية لتحسين النتائج.

الطريقة الثالثة: [4]

فكرتها هي دمج مقدرات الطريقتين السابقتين وتهدف إلى إيجاد تقريب سريع لأصغر محددة تباين مشترك داخل المجموعة MWCD والمقاسة من قبل McLachlan & Hawkins (1997) بدلا من تطبيق النسب التجريبية لكل مجموعة، بفرض إيجاد h من المشاهدات خارج مجموعة البيانات ذات حجم n بحيث إنَّ عينة مصفوفة التباين المشترك داخل المجاميع المدمجة $\hat{\Sigma}_H$ لها أقل محددة وإن:

$$\delta_{ij} = \begin{cases} 1 & if \quad x_{ij} \in H \\ 0 & if \quad x_{ij} \notin H \end{cases} \qquad \dots (23)$$

5.2 قواعد التمييز التربيعية الحصينة:[6]

قاعدة التمييز التربيعية الحصينة RQDR Robust Quadratic Discriminant Rule تكون بوضع x إلى π_k إذا:

$$\hat{d}_k^{RQ}(x) > \hat{d}_j^{RQ}(x) \qquad \forall j = 1,\dots,g \neq k \qquad \dots \quad (24)$$

وإن

$$\hat{d}_j^{RQ} = -\frac{1}{2}\ln\left|\hat{\Sigma}_{j,MCD}\right| - \frac{1}{2}(\underline{x} - \underline{\hat{\mu}}_{j,MCD})'\hat{\Sigma}_{j,MCD}^{-1}(\underline{x} - \underline{\hat{\mu}}_{j,MCD}) + lN(\hat{P}_j^R) \qquad \dots (25)$$

6.2 احتمال خطأ التصنيف:

6.2.1 طريقة التعويض: [9]

افرض إن التمييز يعتمد على مصفوفة البيانات X لـ n_j والتي تأتي من المجتمع j وإذا كانت قاعدة التمييز معرفة بواسطة مناطق التوزيع Ri وبفرض ان n_{ij} عدد المفردات في $\prod_i$ والتي تقع في R_i بحيث إنَّ $\sum_i n_{ij}$ فإنَّ تقدير احتمال خطأ التصنيف يكون:

$$\hat{P}_{ij} = \frac{n_{ij}}{n_j} \qquad \dots (26)$$

6.2.2 طريقة U لـ Jack-Knifing:

المشكلة في طريقة التعويض هي نفس المشاهدات تستخدم لتعريف قاعدة التمييز Discriminant Rule والحكم على دقتها وإحدى الطرائق لتجنب هذه المشكلة هي طريقة U لـ Jack-Knifing وتعرف كالآتي:

افرض ان r=1,…,n و x_r هو أول n_1 من الأسطر للمصفوفة X والتي تمثل المجتمع الأول $\prod_1$ ولكل r افرض ان $R_1^{(r)},\dots,R_g^{(r)}$ تعرف كمناطق لبعض قواعد التمييز والتي تعتمد

على مصفوفة $(n-1)\times p$ والتي تحصل عليها بحذف السطر r^{th} من X فان x_r يمكن أنْ نحكم عليها على أساس القاعدة وهي غير مشتقة من x_r.

إذا رمزنا n^*_{i1} عدد مفردات n_1 لـ $x_r \in R_i^{(r)}$ ، فإنَّ:

$$P^*_{i1} = n_{i1} / n_i \quad \dots (27)$$

وهو تقدير لاحتمال خطأ التصنيف وبواسطة تكرار الإجراء لبقية المجتمعات j=1,…,g نحصل على تقدير p^*_{ij} ، وهذه الطريقة تقود لتقدير معول عليه لاحتمال خطأ التصنيف مقارنة مع الطريقتين السابقتين.

7.2 اختبارات التمييز: [2,10]

7.2.1لاختبار المتغيرات ذات القوة التمييزية المعنوية والتي تتبع أقل خطأ تصنيف:

هناك عدة أساليب منها الاختيار المتدرج Stepwise Selection وفيها يكون اختيار دالة التمييز القياسي لمعنوية الفرق بين متوسطات المجاميع تعميم متعدد المتغيرات لمقياس Wilk's Lambda أُحادي المتغير وبأخذ التباينات المشتركة بنظر الاعتبار إذْ يوفر مقياس Wilk's Lambda متعدد المتغيرات ولـ P من المتغيرات.

$$\hat{P} = |W_{pp}| / |T_{PP}| \quad \dots (28)$$

وإنَّ:

W: مصفوفة مجموع مربعات الانحرافات داخل المجاميع.

T: مصفوفة مجموع مربعات الانحرافات الكلية وتكون إحصائية F المناظرة المستندة على اساس هذا المقياس كالآتي:

$$F(\hat{P}) = [1-\hat{P}][n-P-1] / \hat{P}P \quad \dots (29)$$

والتي تتوزع $F_{1,n-p-1}$. وعلى أساس هذه الاحصائية هناك اساليب متدرجة تستخدم لاختيار المجموعة الجزئية j من المتغيرات منها الاختيار بالحذف الشرطي، الاختيار المتقدم Forward Selection، الاختيار التراجعي Backward Selection وطريقة الاختيار الكامل All possible selection.

2-7-2 اختبار دالة التمييز بعد حذف المتغيرات غير التمييزية:

ومن هذه الاختبارات

أ. اختبار تساوي المجموعات:

وهناك عدة صيغ للاختبار منها مقياس مربع كآي إذْ إنَّ:

$$\chi^2 = -[n-1-1/2(p+g)]\log\hat{e} \quad \dots(30)$$

والذي يتوزع تقريباً مربع كآي بدرجة حرية p(g-1).

ب. اختبار تساوي مصفوفة التباين والتباين المشترك:

إنَّ احصائية الاختبار

$$M = \sum_{i=1}^{g} n_i \ln|S| - \sum_{i=1}^{g} n_i \ln|S_i| \quad \text{... (31)}$$

وقد اثبت Box إذا ضرب M في ثابت C^{-1} والذي يساوي:

$$C^{-1}1 - \frac{2p^2+3p+1}{6(p+1)(g-1)}\left[\sum_{i=1}^{g}\frac{1}{n_i} - \frac{1}{\sum_{i=1}^{g} n_i}\right] \quad \text{... (32)}$$

فإننا نحصل على مقياس Box's M والذي يتوزع تقريباً مربع كآي بدرجة حرية (g-1)(p-1)/2.

3. الجانب التجريبي:

وهنا تم محاكاة حالات مختلفة فالمحاكاة هي تقليد الواقع العملي بحيث نقوم بتوظيف أو تكوين نماذج يظهر فيها عدد كبير من الحالات الافتراضية لتكون نتائج التحليل أكثر شمولية وتعميم وقد كثر الاستعانة بها بسبب التطور الحاصل في صناعة الحواسيب من الناحيتين المادية (Hardware) والبرمجية (Software).

وغالباً مايتم اللجوء لأسلوب المحاكاة للتأكد من صحة جانب تطبيقي موجود أصلاً أو لصعوبة الحصول على بيانات توفر معلومات دقيقة عن ظاهرة معينة أو عندما يصعب إثبات البرهان الرياضي بشكل نظري لبيان أفضلية طرائق تقدير معينة على حساب أخرى.

وللمحاكاة مرونة في اختيار أحجام العينات العشوائية المفترض بها تمثل مجتمع الدراسة وكذلك القدرة على التنوع بالأخطاء العشوائية وتنوع حالات التلويث في البيانات وأيضاً القدرة على إجراء التجربة لعدد كبير من المرات بهدف التوصل على نتائج يعول عليها.

3.1 توليد المتغيرات والمتغيرات المستخدمة:

تم تنفيذ تجارب المحاكاة إذْ تم استخدام أحجام عينات صغيرة، متوسطة، وكبيرة وكالآتي n=40، n=60، n=100 وكل حجم عينة هو المجموع الكلي لمفردات مجتمعين بحيث إنَّ مفردات المجتمع الأول تساوي مفردات المجتمع الثاني وبواقع (27) تجربة وبتكرار (1000) لكل تجربة ولكل حالة تم أخذ ثلاثة أحجام من المتغيرات التوضيحية وهي ثلاثة وستة وتسعة ولحالة التوزيع الطبيعي للخطأ العشوائي للحالة النظيفة clean بالإضافة لحالة تلويث البيانات بقيم شاذة وبنسبتي تلويث 10% و 20% وتم الاعتماد على الإمكانية العالية لبرنامج MATLAB إذْ تم إجراء عدة خطوات للحصول على البيانات وكما يلي:

1. تم توليد البيانات بالشكل النظيف (غير الملوث) للمجتمع الأول والمجتمع الثاني استعانة بالدالة المكتبية Rand والتي تعمل على توليد خلايا من الأرقام العشوائية والتي تتوزع توزيعاً طبيعياً بمتوسط 0 وتباين $\sigma^2 = 1$، ومن ثم تحويل البيانات من التوزيع N(0,1) الى التوزيع $N(\mu,1)$ من خلال العلاقة $X=Z+\mu$ وبهذه الطريقة تم توليد بيانات المجتمع الأول بتوزيع N(-1,1) وللمجتمع الثاني بتوزيع N(1,1) ومن ثم توليد متجه واحد بعد إعطاء ترميز للمجتمع الأول الرقم (1) والمجتمع الثاني (2) لكي يمثل هذا المتجه المتغير y وبالتالي حصلنا على مصفوفة ذات عدة أبعاد إذْ يمثل عدد صفوفها (n_1+n_2) وإن n_1، n_2 تمثل قيم العينة للمجتمع الأول والثاني على التوالي أمَّا أعمدتها فتمثل عدد المتغيرات X's والتي تم أخذها بعدة حالات هي الأُولى: تم أخذ ثلاثة متغيرات والثانية: ستة متغيرات والثالثة: تسعة متغيرات أمَّا العمود الأخير فيمثل المتغيرات المعتمدة y_1، y_2 إذْ إنَّ y_1 يتراوح عددها من $(1,\ldots,n_1)$ و y_2 من $(n_1+1,\ldots,n_1+n_2)$.

2. توليد المتغيرات بالشكل الملوث إذْ تم إتباع أُسلوب توليد متغيرات التوزيع الطبيعي نفسه لتوليد البيانات النظيفة بمتوسط μ وتباين 1 وتم استقطاع جزء من البيانات بنسبة 10% و 20% والتي تمثل نسب التلوث ولكل حجم عينة وللمجتمع الأول والثاني باستخدام نفس دالة التوليد Randn لكن يكون توزيع الشواذ للمجتمع الأول N(9,1) والمجتمع الثاني N(-9,1) وتوجد ثلاث حالات للتلويث التي تمثل تأثير الشواذ Outliers وهي كالاتي:

الحالة الأولى: التلويث باستخدام متوسطات غير متساوية ومصفوفات تباين مشترك متساوية Unequal Means and Equal Covariance's Matrices إذ المجتمعات لها نفس مصفوفة التباين المشترك لكن متوسطات مختلفة بسبب تساوي مصفوفات التباين المشترك مع ملاحظة انه في حالة تساوي مصفوفات التباين المشترك القيم النظرية للاحتمال الكلي لخطأ التصنيف للقاعدة التربيعية هو نفسه للقاعدة الخطية ولحالة التلويث بالشواذ المقدرات الحصينة هي الأفضل بشكل واضح.

الحالة الثانية: التلويث باستخدام مصفوفات التباين المشترك غير المتساوية ومتوسطات Equal Means and Unequal Covariance Matrices بفرض أنَّ المجتمعات لها نفس المتوسط لكن مصفوفات التباين المشترك مختلفة، وفي حالة تساوي المتوسطات وعدم تساوي مصفوفات التباين المشترك يلاحظ بأنَّه هناك فروقات بين التحليل المميز الخطي والتربيعي ويكون الأخير هو الأفضل بسبب الطريقة الخطية تفترض تساوي مصفوفة التباين المشترك.

الحالة الثالثة: التلويث باستخدام متوسطات ومصفوفات التباين المشترك غير متساوية Unequal (Means and Unequal Covarince Matrices) إذْ يفترض المجتمعات لها متوسطات مختلفة ومصفوفات تباين مشترك مختلفة وهذه الحالة هي الأكثر أُنموذجية (مثالية).

3.2 البرامجيات Software:

تم تنفيذ تجارب المحاكاة بكتابة برنامج في نظام Matlab والخاصة بمقدرات H الحصينة، إذْ يعد من البرامج ذات الإمكانية المتقدمة في الجوانب الإحصائية والرياضية والهندسية حيث يوظف الأدوات بشكل مستقيم لبرمجة متقدمة جدا إذْ امتازت طرائق التقدير الحصينة بدرجة من التعقيد كالحاجة إلى لغة برمجة متقدمة Matlab.

3.3 سير عملية المحاكاة:

1. تجري عملية تغذية برنامج المحاكاة بالمدخلات التي حددت.
2. يجري توليد البيانات وفقاً للحالات المفترضة في الخطوة (1) ووفقا لبرامج فرعية ثم ترسل البيانات المولدة إلى البرنامج الرئيسي.
3. تم تطبيق الطرائق غير الحصينة المتمثلة بالدالة المميزة الخطية (LDF) والدالة المميزة التربيعية (QDF).
4. تم تطبيق المقدرات الحصينة وهي مقدر H الحصين، مقدر S الحصين ومقدر MCD الحصين على الدالة المميزة الخطية والتربيعية.
5. تم الاعتماد على حساب معيار احتمال خطأ التصنيف في عملية المقارنة بين الطرائق غير الحصينة والحصينة وتم التفضيل بين المقدرات وفق معيار أقل احتمال خطأ تصنيف.

3.4 مناقشة نتائج المحاكاة:

1. الحالة الطبيعية:

تم توليد البيانات وفقاً للحالة الطبيعية (clean) ومن ثم طبقت الطرائق غير الحصينة والحصينة عليها وتمثل الجداول (1) و (2) و (3) نتائج تحليل الطرائق التمييزية للبيانات المولدة النظيفة للدوال المميزة الخطية والتربيعية للمقدر التقليدي، مقدر H الحصين، مقدر S الحصين.

أمَّا الجدول رقم (10) فيمثل احتمال خطأ التصنيف للبيانات المولدة النظيفة وللدوال المميزة الخطية والتربيعية للمقدر التقليدي، مقدر H الحصين، مقدر S الحصين، مقدر MCD الحصين، يلاحظ وعند استخدام المتغيرات p=3، p=6، p=9 ولأحجام العينات n=40، n=60، n=100 ان مقدر MCD الحصين وللدالتين الخطية والتربيعية هو الأفضل يليه من حيث الأفضلية مقدر S الحصين ثم مقدر H الحصين، وبشكل عام عند احتساب معيار احتمال خطأ التصنيف ولأحجام

العينات المختلفة وباختلاف عدد المتغيرات التوضيحية يلاحظ انَّ الدالة المميزة التربيعية باستخدام مقدر MCD الحصين أعطت أقل احتمال لخطأ التصنيف.

2. حالة تلويث البيانات:

تم توليد البيانات وفقا لنسب التلويث المعتمدة 10% و 20% وتم تطبيق الطرائق غير الحصينة والحصينة عليها وتمثل الجداول (4) و(5) و(6) و(7) و (8) و (9) نتائج تحليل الطرائق التمييزية للبيانات الملوثة للدوال المميزة الخطية والتربيعية التقليدية، مقدر H الحصين، مقدر S الحصين.

أمَّا الجداول رقم (11) و (12) فتمثل احتمال خطأ التصنيف للبيانات الملوثة بنسبتي تلويث 10% و 20% وعند استخدام المتغيرات p=6، p=9 ولأحجام العينات n=40 n=60 و n=100 ولحالة نسبة التلويث 10% و 20% يلاحظ بأنَّ مقدر MCD الحصين وللدالتين الخطية والتربيعية هو الأفضل يليه من حيث الأفضلية مقدر S الحصين ثم مقدر H الحصين.

وبشكل عام عند احتساب معيار احتمال خطأ التصنيف ولأحجام العينات المختلفة باختلاف عدد المتغيرات التوضيحية يلاحظ ان الدالة المميزة التربيعية باستخدام مقدر MCD الحصين أعطت أقل احتمال لخطأ التصنيف.

جدول رقم (1)

نتائج تحليل الطرائق التمييزية للبيانات المولدة النظيفة

QS	LS	QH	LH	QDF	LDF	حجم العينة وعدد المتغيرات	المعاملات
0.6118	-0.2743	-0.521	-0.2262	0.4190	-0.2334	(40,3)	constant
-2.1725	-2.0353	-3.715	-3.7155	-1.5854	-1.4402		Var1
-2.9076	-2.8493	-2.7760	-2.5176	-2.6395	-2.5745		Var2
-1.9763	-1.7534	-1.5710	-2.7185	-2.0192	-1.8152		Var3
-1.3836	-1.6623	-1.0810	-1.0051	-1.2415	-1.6842	(40,6)	constant
-2.599	-2.3703	-2.6016	-3.5781	-2.5108	-2.2342		Var1
-1.7018	-1.9156	-1.860	-1.9351	-1.5865	-1.7982		Var2
-0.7335	-0.5853	0.7613	-1.2202	-0.6622	-0.5036		Var3
-2.7827	-2.3934	-2.3827	-3.7106	-2.6588	-2.2861		Var4
-4.3102	-3.7775	-4.5012	-3.6500	-3.989	-3.4652		Var5
-4.3924	-4.5734	-4.2251	-5.3162	-4.0477	-4.21		Var6
-17.795	1.6506	-11.424	-1.7251	-15.617	1.6126	(40,9)	constant
-6.5211	-5.0007	-4.631	-6.521	-6.4061	-4.8584		Var1
-17.604	-10.476	-15.791	-2.317	-16.589	-10.423		Var2
-13.632	-7.5965	-17.8443	-5.617	-12.922	-7.7244		Var3
-19.102	-10.482	-22.2085	-7.3210	-18.192	-10.338		Var4
-3.1198	-1.6738	-5.0137	-3.7151	-3.0638	-1.6677		Var5
-9.6147	-6.2513	-10.5112	-5.5431	-9.6417	-6.4563		Var6
-4.689	-2.457	-4.5391	-3.1774	-3.9932	-2.2634		Var7
-9.7006	-6.6657	-5.3077	-5.1173	-9.3689	-6.6953		Var8
-13.383	-6.3537	-7.8903	-7.3838	-12.488	-6.3971		Var9

جدول رقم (2)

نتائج تحليل الطرائق التمييزية للبيانات المولدة النظيفة

QS	LS	QH	LH	QDF	LDF	حجم العينة وعدد المتغيرات	المعاملات
0.9638	0.5355	0.6471	0.8891	0.7733	0.4729	(60,3)	Constant
-1.0048	-0.7228	-1.7865	-0.0521	-1.1897	-0.8655		Var1
-2.3226	-2.0671	-2.709	-3.8171	-2.1453	-2.0184		Var2
-3.0537	-2.7058	-4.3177	-2.5171	-2.8556	-2.5962		Var3
-1.796	-1.9247	-1.1298	-1.1960	-1.7513	-1.8802	(60,6)	constant
-2.7523	-2.4366	-2.9842	-3.0946	-2.6378	-2.3437		Var1
-3.4852	-3.6971	-4.5866	-2.1844	-3.1795	-3.4026		Var2
-0.4292	-0.7006	-2.1877	-2.3027	-0.4156	-0.6265		Var3
-2.8	-2.307	-1.599	-1.0717	-2.7682	-2.3904		Var4
-2.2722	-2.0307	-3.551	-2.2119	-2.3963	-2.1943		Var5
-4.6176	-4.3633	-6.3470	-3.0324	-4.37	-4.1529		Var6
-0.9404	-0.10912	-0.0691	-1.2801	-0.9987	-0.1354	(60,9)	constant
-4.0746	-3.4741	-3.884	-1.8700	-4.1152	-3.5156		Var1
-0.2018	-0.7815	-0.3261	-0.3086	-0.2865	-0.8376		Var2
-0.9049	-1.5011	-0.0968	-1.4634	-0.9482	-1.532		Var3
-0.7686	-1.5189	-0.6094	-1.9446	-0.7145	-1.4216		Var4
-4.0636	-3.1853	-4.7137	-3.1648	-4.0158	-3.1573		Var5
-0.9169	-0.9019	-0.9987	-0.1652	-0.7995	-0.7635		Var6
-3.0442	-1.9658	-4.2195	-1.550	-3.0309	-1.9513		Var7
-4.925	-3.6621	-3.7363	-1.0637	-4.753	-3.5868		Var8
-3.1233	-2.2832	-2.7581	-1.1159	-3.0106	-2.2232		Var9

جدول رقم (3)

نتائج تحليل الطرائق التمييزية للبيانات المولدة النظيفة

QS	LS	QH	LH	QDF	LDF	حجم العينة وعدد المتغيرات	المعاملات
0.26663	0.500	0.307	0.1935	0.2013	0.0611	(100,3)	constant
-1.4238	-1.5401	-1.674	-1.1157	-1.3824	-1.4981		Var1
-1.6054	-1.4069	-2.6703	-2.761	-1.6569	-1.4083		Var2
-1.6976	-1.5939	-2.501	-2.1682	-1.7384	-1.6135		Var3
1.6073	-0.7016	0.6229	0.26452	1.43	0.6755	(100,6)	constant
-2.6305	-2.4718	-3.5431	-2.2836	-2.7438	-2.5621		Var1
-1.5073	-1.592	-2.4538	-2.1413	-1.458	-1.5492		Var2
-3.1856	-3.1127	-3.3614	-2.2285	-2.761	-2.6681		Var3
-2.3849	-2.2535	-3.3321	-3.6003	-2.2363	-2.1271		Var4
-2.9859	-2.9505	-2.7733	-2.5497	-2.8861	-2.8763		Var5
-2.6971	-2.7359	-2.66	-2.8573	-2.7505	-2.7395		Var6
-0.8183	0.6788	0.5539	0.4550	-1.1008	0.3731	(100,9)	constant
-3.6913	-2.4312	-3.37	-2.8792	-2.976	-2.4469		Var1
-2.8301	-3.1844	-3.8534	-2.5503	-3.2512	-3.0545		Var2
-3.1125	-2.7716	-3.6753	-3.8206	-2.5210	-2.77		Var3
-2.4866	-1.8889	-2.156	-2.8044	-2.3852	-1.8005		Var4
-2.919	-2.5823	-3.4320	-2.7006	-2.7996	-2.5315		Var5
-2.462	-1.8082	-2.17	-1.4527	-2.5094	-1.8101		Var6
-2.4824	-2.2638	-2.8426	-2.2723	-3.1855	-2.2648		Var7
-3.3905	-2.1476	-2.9826	-3.0642	-2.8705	-2.1714		Var8
-2.9612	-2.6647	-2.1248	-2.3346	-3.7035	-2.5801		Var9

جدول رقم (4)

نتائج تحليل الطرائق التمييزية للبيانات المولدة الملوثة وبنسبة تلويث 10%

QS	LS	QH	LH	QDF	LDF	حجم العينة وعدد المتغيرات	المعاملات
-0.4952	-0.1144	-0.8251	-0.9461	-0.2674	-0.0135	(40,3)	constant
-1.1897	-1.2754	-1.362	-2.6221	0.0183	0.0722		Var1
-1.4317	-0.8740	-2.3017	-1.7743	0.1955	-0.1013		Var2
-1.4409	-0.5543	-3.7425	-2.2122	-0.1520	0.0437		Var3
0.6326	0.1662	1.2986	0.6106	0.6387	0.0013	(40,6)	constant
0.2635	-0.0812	-0.253	-1.7253	0.2841	0.0627		Var1
-0.2156	-0.3853	-0.5796	-2.7972	-0.0392	-0.1513		Var2
-0.5376	-0.2032	-0.8292	-0.8622	-0.5212	-0.3497		Var3
-0.4695	-0.0487	-0.0705	-1.3625	-0.5010	-0.1970		Var4
0.5181	0.1394	0.0961	-2.5959	0.4604	0.2991		Var5
0.3646	0.2347	0.4618	-2.2692	0.344	0.3131		Var6
0.8623	0.2973	0.3631	0.351	0.8977	0.3971	(40,9)	constant
-0.3191	-0.1164	-0.5822	-0.0641	-0.3027	-0.1608		Var1
0.7998	0.3623	-0.4376	-2.7796	0.8161	0.4386		Var2
-1.1527	-0.4454	-0.83729	-3.8421	-1.0423	-0.5599		Var3
-0.1625	0.2254	-0.7007	0.9594	-0.2035	0.3605		Var4
-1.5652	-0.6878	0.6789	-2.6096	-1.5145	-0.8524		Var5
1.3632	0.3516	0.0981	-1.3677	1.2898	0.3135		Var6
2.028	0.6177	0.943	-2.5920	1.9752	0.9251		Var7
1.7676	0.9378	0.6301	0.5809	1.8106	1.2141		Var8
-2.5701	-1.2649	0.2702	0.4355	-2.6086	-1.5605		Var9

جدول رقم (5)

نتائج تحليل الطرائق التمييزية للبيانات المولدة الملوثة وبنسبة تلويث 10%

QS	LS	QH	LH	QDF	LDF	حجم العينة وعدد المتغيرات	المعاملات
0.3225	0.0608	0.241	0.0105	0.2338	-0.0482	(60,3)	constant
-1.4581	-1.4518	-1.9333	-2.539	0.3005	0.3153		Var1
-1.2268	-1.1602	-2.6601	-1.8482	-0.0685	-0.0802		Var2
-1.9124	-1.8665	-1.5482	-2.9825	-0.2246	-0.2184		Var3
-3.3543	-0.2789	-2.2173	-0.2828	-0.2029	-0.1083	(60,6)	constant
-1.4128	0.0567	0.270	-2.5504	0.1182	0.0197		Var1
-1.6151	-0.3282	-0.2821	-0.8271	-0.2366	-0.2281		Var2
-0.7433	-0.0518	-0.3680	-0.0444	0.0331	-0.0994		Var3
-1.0007	-0.0802	-0.01213	-0.6746	0.0451	0.0082		Var4
-0.8691	0.0052	0.2924	-0.095	-0.0885	0.1025		Var5
-1.7295	0.0815	0.0053	-0.7513	0.1450	0.2046		Var6
0.0953	-0.0366	-0.1293	0.0087	0.103	-0.0526	(60,9)	constant
0.1452	0.0156	0.2106	0.2525	0.1493	0.0282		Var1
-0.2235	-0.1237	-0.8152	-0.1808	-0.2171	-0.1233		Var2
-0.3771	-0.3226	-0.636	-0.1765	-0.3906	-0.3434		Var3
0.1920	0.1966	0.677	-0.2513	0.2225	0.2258		Var4
-0.0869	-0.0064	0.0297	-0.0059	-0.0517	0.0042		Var5
-0.1621	-0.1093	-0.7461	-0.1492	-0.1362	-0.0837		Var6
0.2389	0.1011	-0.0261	0.2553	0.2535	0.1065		Var7
0.1233	0.1547	-0.2177	0.4630	0.1433	0.1406		Var8
0.0615	0.0245	-0.2362	0.5465	0.1716	0.0214		Var9

جدول رقم (6)

نتائج تحليل الطرائق التمييزية للبيانات المولدة الملوثة وبنسبة تلويث 10%

QS	LS	QH	LH	QDF	LDF	حجم العينة وعدد المتغيرات	المعاملات
-0.0862	-0.4309	-0.5971	-0.2301	-0.2002	-0.0324	(100,3)	constant
-0.9784	-0.9280	-0.8531	-0.5560	0.0721	0.0996		Var1
-2.105	-2.0557	-1.638	-1.2655	-0.13	-0.1395		Var2
-1.8401	-1.7719	-1.666	-1.1479	0.0331	0.0237		Var3
0.7066	0.1823	0.7956	0.543	-0.1556	-0.0389	(100,6)	constant
-2.1788	-2.0338	-1.3233	-1.5364	-0.0225	-0.0039		Var1
-0.9713	-1.0248	-2.735	-1.97	0.4339	0.3399		Var2
-2.039	-1.9256	-1.815	-1.0721	-0.0058	-0.0960		Var3
-2.1914	-1.8009	-1.766	-1.323	-0.1835	0.1029		Var4
-1.2093	-1.0474	-2.7389	-1.3477	0.2777	0.2977		Var5
-1.4537	-1.3511	-1.4957	-1.3425	-0.5493	-0.4676		Var6
-0.0431	0.0187	0.5909	-0.0709	0.0007	0.0399	(100,9)	constant
0.3194	0.2725	0.5242	0.1353	0.3446	0.2717		Var1
0.2387	0.2401	0.665	0.2693	0.2556	0.2575		Var2
-0.3072	-0.2007	-0.97	-0.1641	-0.2911	-0.1859		Var3
-0.3594	-0.3706	-0.4791	-0.6215	-0.3527	-0.3568		Var4
-0.3512	-0.254	-0.61	-0.3852	-0.3854	-0.2739		Var5
0.1037	0.0787	-0.049	-0.4910	0.1362	0.1137		Var6
-0.1100	-0.0807	-0.0535	0.0105	-0.1153	-0.0899		Var7
0.2648	0.1978	-0.6465	0.001	0.2796	0.2037		Var8
0.1836	0.0985	0.9941	0.4635	0.1583	0.0883		Var9

جدول رقم (7)

نتائج تحليل الطرائق التمييزية للبيانات المولدة الملوثة وبنسبة تلويث 20%

QS	LS	QH	LH	QDF	LDF	حجم العينة وعدد المتغيرات	المعاملات
0.3225	0.0608	0.241	0.0105	0.2338	-0.0482	(40,3)	constant
-1.4581	-1.4518	-1.9333	-2.539	0.3005	0.3153		Var1
-1.2268	-1.1602	-2.6601	-1.8482	-0.0685	-0.0802		Var2
-1.9124	-1.8665	-1.5482	-2.9825	-0.2246	-0.2184		Var3
-3.3543	-0.2789	-2.2173	-0.2828	-0.2029	-0.1083	(40,6)	constant
-1.4128	0.0567	0.270	-2.5504	0.1182	0.0197		Var1
-1.6151	-0.3282	-0.2821	-0.8271	-0.2366	-0.2281		Var2
-0.7433	-0.0518	-0.3680	-0.0444	0.0331	-0.0994		Var3
-1.0007	-0.0802	-0.01213	-0.6746	0.0451	0.0082		Var4
-0.8691	0.0052	0.2924	-0.095	-0.0885	0.1025		Var5
-1.7295	0.0815	0.0053	-0.7513	0.1450	0.2046		Var6
0.0953	-0.0366	-0.1293	0.0087	0.103	-0.0526	(40,9)	constant
0.1452	0.0156	0.2106	0.2525	0.1493	0.0282		Var1
-0.2235	-0.1237	-0.8152	-0.1808	-0.2171	-0.1233		Var2
-0.3771	-0.3226	-0.636	-0.1765	-0.3906	-0.3434		Var3
0.1920	0.1966	0.677	-0.2513	0.2225	0.2258		Var4
-0.0869	-0.0064	0.0297	-0.0059	-0.0517	0.0042		Var5
-0.1621	-0.1093	-0.7461	-0.1492	-0.1362	-0.0837		Var6
0.2389	0.1011	-0.0261	0.2553	0.2535	0.1065		Var7
0.1233	0.1547	-0.2177	0.4630	0.1433	0.1406		Var8
0.0615	0.0245	-0.2362	0.5465	0.1716	0.0214		Var9

جدول رقم (8)

نتائج تحليل الطرائق التمييزية للبيانات المولدة الملوثة وبنسبة تلويث 20%

QS	LS	QH	LH	QDF	LDF	حجم العينة وعدد المتغيرات	المعاملات
-0.2169	0.0344	0.0194	0.1029	-0.3145	0.0041	(60,3)	constant
0.2306	0.2063	0.2463	0.4720	0.3049	0.2538		Var1
0.0509	0.0446	0.0127	0.2563	0.0462	0.0432		Var2
-0.2350	-0.1929	-0.1598	-0.2468	-0.2274	-0.1730		Var3
-0.2629	0.0136	0.0173	0.2703	-0.2582	0.0228	(60,6)	constant
0.6959	0.5303	0.5462	0.4705	0.6459	0.4983		Var1
-0.4102	-0.2737	-0.4399	-0.3194	-0.3856	-0.2530		Var2
-0.2588	-0.3062	-0.6396	-0.3855	-0.2614	-0.2902		Var3
0.2322	0.1941	0.2725	0.1928	0.1794	0.1535		Var4
-0.3387	-0.2711	-0.2033	-0.840	-0.2731	-0.2430		Var5
0.1922	0.2131	-0.2589	-0.0579	0.2273	0.237		Var6
0.0464	0.0107	0.0454	0.3274	0.0738	0.0328	(60,9)	constant
-0.3933	-0.3498	-0.4622	-0.1163	-0.4174	-0.3586		Var1
0.0586	0.0073	0.3265	0.0039	0.1221	0.0680		Var2
0.6651	0.5914	0.1175	0.4421	0.6493	0.5776		Var3
0.8636	0.6360	0.4260	0.7753	0.8703	0.6388		Var4
-0.1732	-0.0774	-0.0012	-0.1075	-0.1139	-0.0341		Var5
-0.2848	-0.0562	-0.2005	-0.0967	-0.3518	-0.1025		Var6
-0.2557	-0.2295	-0.273	-0.4722	-0.2704	-0.2317		Var7
-0.3710	-0.2354	-0.4695	-0.4026	-0.3794	-0.2673		Var8
-0.0372	-0.1943	-0.5224	-0.1276	-0.0205	-0.1806		Var9

جدول رقم (9)

نتائج تحليل الطرائق التمييزية للبيانات المولدة الملوثة وبنسبة تلويث 20%

QS	LS	QH	LH	QDF	LDF	حجم العينة وعدد المتغيرات	المعاملات
-0.1994	-0.0543	-0.433	-0.0409	-0.2717	-0.0550	(100,3)	constant
-0.1768	-0.2294	-0.3929	-0.1765	-0.2016	-0.2531		Var1
0.0635	0.1566	0.262	-0.0459	0.0250	0.1297		Var2
0.1873	0.1412	0.5592	-0.5535	0.3027	0.2420		Var3
0.3007	-0.0337	-0.1284	1.3111	0.2889	-0.0607	(100,6)	constant
0.3623	0.2231	0.1465	0.0164	0.3918	0.2620		Var1
0.1573	0.1599	0.7326	0.2536	0.1461	0.1484		Var2
0.1670	0.1607	0.3821	0.4320	0.1257	0.1221		Var3
-0.1462	-0.1604	-0.0278	-0.2467	-0.1650	-0.1772		Var4
0.0111	0.0606	-0.2088	-0.0030	-0.0068	0.0441		Var5
-0.4171	-0.3289	-0.2471	-0.7876	-0.3446	-0.2707		Var6
0.2798	0.1093	0.2059	0.21341	0.2878	0.1216	(100,9)	constant
0.2558	0.3072	0.4477	0.7795	0.2681	0.3305		Var1
-0.0187	-0.0917	-0.828	-0.1226	-0.0573	0.1200		Var2
0.5652	0.4831	-0.4264	-0.370	0.5731	0.4797		Var3
-0.0727	-0.0103	-0.1456	-0.2675	-0.0499	-0.0125		Var4
-0.2103	-0.2973	-0.1729	-0.1500	-0.1854	-0.2628		Var5
-0.5510	-0.4073	-0.553	-0.4763	-0.5282	-0.3845		Var6
-0.2063	-0.2014	0.8207	0.3528	-0.2448	-0.2237		Var7
0.1941	0.1798	0.6134	0.1929	0.209	0.1877		Var8
0.1977	0.1644	0.3519	0.3379	0.1891	0.1477		Var9

جدول رقم (10)

احتمال خطأ التصنيف للبيانات المولدة النظيفة

حجم العينة	LDF	QDF	LH	QH	LS	QS	LMCD	QMCD
(40,3)	0.35	0.1536	0.1155	0.0921	0.05	0.05	0	0
(40,6)	0.3216	0.1733	0.0971	0.06	0	0	0	0
(40,9)	0.382	0.2210	0.0962	0.0513	0	0	0	0
(60,3)	0.3911	0.1503	0.11	0.0982	0.05	0.05	0.05	0.0196
(60,6)	0.2916	0.1325	0.1215	0.0731	0	0	0	0
(60,9)	0.3119	0.0919	0.1425	0.0651	0	0	0.0102	0
(100,3)	0.3713	0.0812	0.0913	0.0522	0.07	0.07	0.0218	0.0121
(100,6)	0.3151	0.1161	0.100	0.0791	0.01	0.01	0	0
(100,9)	0.2860	0.1251	0.1113	0.0971	0.0321	0.0422	0.0175	0.0199

جدول رقم (11)

احتمال خطأ التصنيف للبيانات الملوثة ولنسبة تلويث 10%

حجم العينة	LDF	QDF	LH	QH	LS	QS	LMCD	QMCD
(40,3)	0.525	0.3963	0.3651	0.2616	0.15	0.1	0.0571	0.0313
(40,6)	0.4770	0.3910	0.3866	0.2711	0.25	0.25	0.0007	0
(40,9)	0.376	0.2271	0.215	0.1351	0.15	0.0761	0.001	0
(60,3)	0.4166	0.3839	0.3768	0.2533	0.1167	0.1167	0.0130	0
(60,6)	0.4	0.3461	0.2715	0.2278	0.1571	0.05	0.005	0
(60,9)	0.5	0.3833	0.3305	0.2715	0.4836	0.25	0.0171	0
(100,3)	0.45	0.51	0.2915	0.2209	0.12	0.09	0.0153	0
(100,6)	0.4421	0.327	0.2511	0.1355	0.1	0.1	0	0.0121
(100,9)	0.48	0.3617	0.2151	0.1173	0.22	0.0931	0.0151	0

جدول رقم (9)

احتمال خطأ التصنيف للبيانات الملوثة ولنسبة تلويث 20%

حجم العينة	LDF	QDF	LH	QH	LS	QS	LMCD	QMCD
(40,3)	0.6	0.395	0.2916	0.2662	0.375	0.225	0.0938	0.0625
(40,6)	0.525	0.325	0.3351	0.2152	0.225	0.1352	0	0
(40,9)	0.5547	0.2851	0.2793	0.1601	0.275	0.1	0.0563	0.0294
(60,3)	0.4667	0.45	0.3357	0.2273	0.2333	0.1620	0.0417	0.0417
(60,6)	0.425	0.35	0.3323	0.3062	0.2572	0.2166	0	0
(60,9)	0.4468	0.3785	0.319	0.2263	0.3333	0.13333	0.01553	0.008
(100,3)	0.4937	0.44	0.3321	0.2415	0.4113	0.2261	0.0389	0.0506
(100,6)	0.43	0.3651	0.3718	0.2153	0.2811	0.1475	0.003	0
(100,9)	0.49	0.43	0.3210	0.2759	0.3365	0.24	0.009	0

الاستنتاجات والتوصيات

الاستنتاجات:

في ضوء تحليل تجارب المحاكاة تم التوصل للاستنتاجات التالية:

1. عند توليد البيانات النظيفة والملوثة ولجميع نسب التلويث ومختلف المتغيرات التوضيحية، وقد لوحظ انَّ مقدر MCD حقق نتائج كفوءة بدرجة عالية إذ أعطى أقل احتمال لخطأ التصنيف وللدالة التربيعية QMCD تليها الدالة الخطية LMCD، يليه مقدر S الحصين ثم مقدر H الحصين، كما ان مقدر MCD الحصين أعطى ولحالات عديدة احتمال خطا تصنيف صفر وهذا يعكس خصائص المقدر الجيد وهي نقطة الانهيار العالية، الحصانة العالية، الكفاءة الإحصائية العالية.
2. أعطت دالة التمييز الخطية والتربيعية التقليديتين أعلى نسبة لخطأ التصنيف مما يتطلب عدم استخدامهم في حال احتواء البيانات على قيم شاذة أي انَّ الشواذ لها تأثير كبير على قاعدة التمييز الخطية LDR وقاعدة التمييز التربيعية QDR.
3. أعطى مقدر S الحصين نتائج أقل كفاءة بالمقارنة مع مقدر MCD الحصين فضلاً عن أنَّ خوارزمية مقدر S احتاجت وقت حساب أطول مقارنة مع خوارزمية مقدر MCD.
4. إنَّ قواعد التمييز التقليدية LDR و QDR في حالة البيانات النظيفة ابرزت كفاءتها لجميع حجوم العينات في حين اخفقت وبشكل واضح لحالة البيانات الملوثة.
5. ان قواعد التمييز بالاعتماد على مقدر MCD الحصين أنجزت بكفاءة عالية ولمجموعة البيانات الصغيرة والمتوسطة والكبيرة.
6. أثبتت الطريقة الثانية لتقدير التباين المشترك الأولي $\hat{\Sigma}_0$ والخاصة بمقدر MCD الحصين كفاءة عالية علماً بأنَّ هذه الطريقة هي تحديث لطريقة He& Fung (2000).

التوصيات:

بناءاً على ما تم التوصل إليه البحث من استنتاجات نقترح بالتوصيات الآتية:

1. نوصي باستخدام مقدر MCD الحصين بقوة لما يمتاز به من تحقيق نقطة انهيار عالية، حصانة عالية وكفاءة إحصائية عالية كما إنَّ المسافة الحصينة باستخدام مقدر MCD هي أكثر دقة من الاعتماد على مقدرات أُخرى مثل MVE.
2. نظراً للمعوقات في حساب مقدر MCD الحصين بقوة لاسيما مع زيادة عدد المتغيرات التوضيحية من تلك المعوقات وقت الحساب لذلك نقترح تطبيق خوارزمية FAST-MCD التي تختزل وقت حساب خوارزمية MCD.

3. نوصي باستخدام مقدر MCD الحصين بقوة في حالة تعدد المجتمعات (أكثر من مجتمعين) ليقارن مع مقدرات حصينة أُخرى وبيان كفاءته.
4. نوصي باستخدام الحالة الثانية والثالثة الخاصة بحالات التلويث في حالة توليد البيانات الملوثة ومقارنتها مع الحالة الأولى عند تطبيق الطرق التقليدية والحصينة لمعرفة الفرق في النتائج.
5. نوصي بتطبيق الطريقتين الأولى والثالثة والخاصة بتقدير التباين المشترك الأولي $\hat{\Sigma}_0$ لمقدر MCD الحصين ومقارنة النتائج مع الطريقة الثانية للبيانات إنْ كان هناك تقارباً او اختلافاً في النتائج.
6. اعتماد طرائق أُخرى للكشف عن الشواذ بشكل أولي ولكل نقطة بيانات كمسافة مهلنوبس الحصينة.

المصادر

اولا: العربية

- النداوي، سرى صباح، (2008) "مقارنة بعض المقدرات الحصينة في الدوال التمييزية مع تطبيق عملي" رسالة ماجستير في الإحصاء، كلية الإدارة والاقتصاد، جامعة بغداد.

ثانيا: الأجنبية

1. Alvin,C. Rencher,(1996),"Methods of Multivariate Analysis", John Wiley & Sons.
2. Croux,C.,Dehon,C.,(2001)."Robust Linear Discriminant Analysis Using S-Estimator", The Canadian Journal of Statistics,vol.29,1-18.
3.Hawkins,D.M.,Mclachlan,G.J.,(1997), "High Breakdown Linear Discriminant Analysis",JASA,Vol.92,136-143.
4.Hubert,P.J.,(1981),"Robust Statistics', John Wily& Sons.
5.Hubert,M.,Driessen,K.V.,(2004),"Fast and Robust Discriminant Analysis", Computational Statistics and Data Analysis,Vol.45,301-320.
6.Lanchenbruch,P.,A.,Sneemger,C.,&Rev,L.T.,(1973),"Robustness of Linear and Quadratic Discriminant Function to certain types of non – normality", Comm.Stat., Vol.1.
7.Launer,R.L.&Wilkinson,G.N.,(1979),'Robustness in Statistics".Academic Press, new York.
8.Mardia,K.V.&Kent,J.T.,Bibby,J.M.,(1979),"Multivariate Analysis" Academic press inc (London)Ltd.
9. Norusis, M., (1986),"SPSS/PC for the IBM PC/XT/AT", Chicago.
11. Ruppert,D., (1992),"Computing S-Estimators for regression and Multivariate Location /Dispersion", American Statistical Association, Vol 1,no.,3,253-270.
12.Wilcox,R.R.,(1997),"introduction to Robust estimation and Hypothesis Testing", Academic Press, New York.