مقارنة بعض المقدّرات الحصينة في دوال التمييز بأستخدام المحاكاة

أ. م. خلود يوسف خمو

كلية الادارة والاقتصاد / جامعة بغداد

قسم الاحصاء

م. م. سرى صباح كيتب

كلية المأمون الجامعة

الملخص

ان التطور في صناعة الحواسيب من الناحيتين (المادية والبرمجية) أمكن حساب مقدّرات حصينة معقدة والتي تزودنا بأسلوب جديد في التعامل مع البيانات في حال اخفاق طرائق التمييز التقليدية في تحقيق خواصها المثلى لاسيما مع إحتواء البيانات على نسبة من الشواذ و بالتالي فشل الحصول على أقل احتمال لخطأ التصنيف. يهدف البحث الى مقارنة مقدّرات حصينة مقاومة لتأثير الشواذ كمقدّر H الحصين، مقدّر S الحصين، ومقدّر MCD الحصين و كذلك تحصين احتمال خطأ التصنيف مع بيان تأثير الشواذ على نسب خطأ التصنيف عند استخدام طرائق التمييز التقليدية. ومن ضمن أهداف البحث مقارنة المقدّرات للتعرّف على أفضل مقدّر يعطي أقل احتمال لخطأ التصنيف ولا سيما مع تنوّع نسب التلويث ومع حجوم عينات مختلفة بأتباع اسلوب المحاكاة .

Abstract

The development in manufacturing computers from both (Hardware and Software) sides, make complicated robust estimators became computable and gave us new way of dealing with the data, when classical discriminant methods failed in achieving its optimal properties especially when data contains a percentage of outliers. Thus, the inability to have the minimum probability of misclassification. The research aim to compare robust estimators which are resistant to outlier influence like robust H estimator, robust S estimator and robust MCD estimator, also robustify misclassification probability with showing outlier influence on the percentage of misclassification when using classical methods. ,the other aim of research is to compare estimators to find the best estimator which can gave less probability of misclassification especially with the variety of contamination percentage and different samples sizes and the data contaminated according to a technique that had never been used in other research on the country level.

1. المقدمة وهدف البحث

ان المقدرات الحصينة تزودنا بأسلوب جديد في التعامل مع البيانات ومع تطور البرامجيات والتي كانت عائقاً امام تطبيق العديد من المقدرات الحصينة والتي تحتاج كفاءة برمجية عالية بالاضافة إلى السرعة. وبالنظر لاخفاق طرائق التحليل المميز التقليدية في تحقيق خواصها المثلى عند احتواء البيانات على نسب من الشواذ وبالتالي اخفاق الحصول على أقل احتمال لخطأ التصنيف، لذا كانت فكرة البحث وهي استخدام مقدرات حصينة مقاومة لتأثير الشواذ ومن تلك المقدرات مقدّر H الحصين، مقدّر S الحصين ومقدّر MCD الحصين، ومن ثم تحصين احتمال خطأ التصنيف. ومن ضمن أهداف البحث مقارنة المقدّرات للتعرف على أفضل مقدّر والذي يعطي أقل احتمال لخطأ التصنيف لاسيما مع تنوع نسب التلويث وحجوم عينات مختلفة وذلك بأتباع أسلوب المحاكاة .

2. الجانب النظري

1.2 دوال التمييز

سنتناول في بحثنا دوال التمييز الخطية والتربيعية التقليدية، بالإضافة إلى المقدّرات الحصينة منها مقدّر H الحصين، مقدّر S الحصين ومقدّر MCD الحصين .

1.2.1 دالة التمييز الخطية

لتكن n_j ، $\bar{x}_j$ و S_j تمثل حجم العينة، متجه المتوسطات$(p\times 1)$ و مصفوفة التباينات المشتركة $(p\times p)$ للمجموعة j على التوالي حيث يتم التعويض عن $\underline{\mu}_j$ بمتجه المتوسطات للعينة $\underline{\bar{x}}_j$ و عن Σ بمصفوفة التباينات المشتركة المدمجة S للعينة (Pooled Covariance Matrix) فإن دالة التمييز الخطية تكون[7]:

$$\hat{Z}_j(x)=\hat{\beta}_{oj}+\hat{\underline{\beta}}'_j\underline{X} \qquad j=1,2,\dots,g-1 \quad \dots\dots(1)$$

$$\hat{\beta}_{oj}=\log\frac{\pi_j}{\pi_g}-\frac{1}{2}\hat{\underline{\beta}}'_j(\underline{\bar{x}}_j+\underline{\bar{x}}_g) \quad \dots\dots(2)$$

و إن:

$$\hat{\underline{\beta}}_j=S^{-1}(\underline{\bar{x}}_j+\underline{\bar{x}}_g) \quad \dots\dots(3)$$

1.2.2 دالة التمييز التربيعية

وتستخدم في حالة عدم تساوي مصفوفة التباين والتباين المشترك للمجموعات وإن دالة التمييز التربيعية تكون:

$$\hat{Z}_j(x)=\hat{\beta}_{oj}+\hat{\underline{\beta}}'_j\underline{X} \quad \dots\dots(4)$$

$$\hat{\beta}_{oj}=\frac{1}{2}\log\frac{|S_g|}{|S_j|}-\frac{1}{2}(\underline{\bar{x}}'_jS_j^{-1}\underline{\bar{x}}_j-\underline{\bar{x}}'_gS_g^{-1}\underline{\bar{x}}_g) \quad \dots\dots(5)$$

والتي تمثل الحد الثابت، أما المعاملات x_j وx_i

$$\hat{\underline{\beta}}_j=(S_j^{-1}\underline{\bar{x}}_j-S_g^{-1}\underline{\bar{x}}_g),\ j=1,2,\dots,g-1 \quad \dots\dots(6)$$

2.2 مقدّر Huber الحصين

إن عائلة **M** تعتمد على تقديرات الأوساط الحسابية والتباينات الموزونة ومن هذه الطرائق طريقة **Huber** ويمكن توضيح هذه الطريقة كما يلي [6]:

أولاً. تحسب مسافة مهلنوبس لكل مشاهدة و كالآتي:

$$D(\underline{x}_i, x) = \left[(\underline{x}_i - \underline{T}(x))' S^{-1}(x)(\underline{x}_i - \underline{T}(x))\right]^{\frac{1}{2}} \qquad ......(7)$$

وإن:

$$T(x) = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad ......(8)$$

$$S(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - T(x))(x_i - T(x))' \qquad ......(9)$$

ثانيا. لتشذيب البيانات يعاد حساب المتوسطات ومصفوفة التباينات حيث تبدل **T(*x*)** و **S(*x*)** بأوزان تعتمد على d_i أي:

$$T^*(x) = \frac{\sum w_i x_i}{\sum w_i} \qquad ......(10)$$

$$S^*(x) = \frac{\sum w_i (x_i - T^*(x))(x_i - T^*(x))'}{\sum w_i^2} \qquad ......(11)$$

وإن:

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

ثم يعاد حساب الخطوة الثانية بصورة تكرارية بالاعتماد على نتائج التكرار السابق ونتوقف عندما يكون الفرق بين تكرارين متعاقبين قليل وفق مستوى معين من الدقة، إذ يتم الحصول على بيانات مشذبة لكافة المجاميع، إذ يتم ضرب القيم بالتكرار الأخير بالمتغيرات كافة ولكافة المجاميع وعندها يتم تطبيق الدالة التمييزية الخطية والتربيعية.

3.2 مقدّر S الحصين

ويمكن تلخيص خوارزمية مقدّر S بالخطوات*:

1. إن مقدّرات S معرفة كحلول للموقع والتشتت (t_n, C_n) لمسألة تقليل محددة التباين المشترك C وفقاً إلى[2, 8] :

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left[d\left(\underline{x}_i, \underline{t}, C\right)\right] = b_o \qquad ......(12)$$

---

*لمزيد من التفاصيل راجع المصدر (1).

2. يكون اختيار أفضل عينة جزئية من العينات المستخرجة وفقاً لمحددة التباين المشترك إذ يتم اختيار العينة الجزئية التي تكون فيها محددة مصفوفة التباين المشترك أقل ما يمكن.
3. بعد اختيار أفضل عينة جزئية من خلال موجّه الموقع و مصفوفة التباين المشترك يتم استخراج مسافات مهلنوبس العادية بين نقاط المشاهدات x_i و الموقع **t** بالاعتماد على مصفوفة التباين المشترك حيث:

$$d(\underline{x}_i, \underline{t}, C) = \left[(\underline{x}_i - \underline{t})' C^{-1} (\underline{x}_i - \underline{t})\right] \quad \ldots\ldots (13)$$

حيث أن العينة $x_1,\ldots,x_n$ تعرف كثنائي (t_n, C_n) و التي تقلل **C** تحت القيد

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left[\sqrt{(\underline{x}_i - \underline{t})' C^{-1} (\underline{x}_i - \underline{t})}\right] = b \quad \ldots\ldots (14)$$

4. يتم استخراج $\rho(u)$ حيث ρ هو دالة **Biweight** التالية

$$\rho(u) = \min\left(\frac{u^2}{2} - \frac{u^4}{2c^2} + \frac{u^6}{6c^4}, \frac{c^2}{6}\right) \quad \ldots\ldots (15)$$

حيث **c** ثابت يحقق قيمة مرغوبة لنقطة الانهيار ويسمى مقدّر(**Biweight's estimator**).

5. بعد ايجاد قيمة **b** تكرر المعادلة $\rho(u)$ ليتم فيها البحث عن قيمة **u** بطريقة التكرار وذلك لتحقيق شرط المقدّر في المعادلة (12).
6. باستخراج قيم الثوابت يتحقق شرط المعادلة (12) ليتم بعدها اختيار مصفوفة التشتت التي تحقق الشرط مع متجه الموقع ليعتبرا مقدّرات **S** الحصينة للموقع والتشتت وبموجبها يتم استخراج المسافات الحصينة لتحقق نقاط البيانات الشاذة.

ويتم بعدها تطبيق الدالة التمييزية الخطية والتربيعية ومن ثم يصنف الخطأ في حال وجود الشواذ في العينة طبقاً لمقدّر **S**، فهناك احتمال خطأ التصنيف للمشاهدة تأتي من ϕ_1 واحتمال خطأ التصنيف للمشاهدة تأتي من ϕ_2 .

4.2 مقدّر MCD الحصين

يعرف مقدّر أصغر محددة تباين مشترك **(Minimum Covariance Determinant)MCD** كمتوسط $\hat{\mu}_{j,o}$ ومصفوفة تباين مشترك $\hat{\Sigma}_{j,o}$ لـ h_j من المشاهدات من n_j والتي تكون مصفوفة التباين المشترك لها أقل ما يمكن[4].

حيث أن:

$$h_j > \left[(n_j + P + 1)/2\right] \quad \ldots\ldots (16)$$

وإن $n_j - h$ يجب أن تكون أقل من عدد الشواذ في **j** من المجتمعات وبسبب كون العدد غير معروف في كل مجموعة تؤخذ:

$$h_j = \left[(n_j + P + 1)/2\right] \approx 50\% \quad \ldots\ldots (17)$$

ويوصى بأخذ $h_j \approx 0.75\, nj$.

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

$$RD^{o}_{ij}=\sqrt{(\underline{x}_{ij}-\underline{\hat{\mu}}_{j,o})'\,S^{-1}_{j,o}\,(\underline{x}_{ij}-\underline{\hat{\mu}}_{j,o})} \quad ......(18)$$

ويعطى الوزن لـ x_i **كالآتي:**

$$RD^{o}_{ij}=\begin{cases}1 & \text{if} \quad RD^{o}_{ij}\le\sqrt{\chi^{2}_{P,0.975}}\\ 0 & \text{other wise}\end{cases} \quad .....(19)$$

ومن ثم نحصل على مقدّر 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 وكالآتي:**

$$RD_{ij}=\sqrt{(\underline{x}_{ij}-\underline{\hat{\mu}}_{j,MCD})'\,\hat{\Sigma}^{-1}_{j,MCD}\,(\vec{\underline{x}}_{ij}-\underline{\hat{\mu}}_{j,MCD})} \quad ......(20)$$

وتعد x_{ij} **قيمة شاذة إذا و فقط إذا:**

$$RD_{ij}>\sqrt{\chi^{2}_{P,0.975}}$$

وبفرض أن $\tilde{n}_j$ **تشير إلى عدد غير الشواذ في المجموعة j وأن:**

$$\tilde{n}=\sum_{j=1}^{g}\tilde{n}_j$$

ويمكن تقدير احتمال خطأ التصنيف بحصانة كالآتي:

$$\hat{P}^{R}_{j}=\frac{\tilde{n}j}{\tilde{n}} \quad ......(21)$$

4.2.1 قواعد التمييز الخطية الحصينة

بإعطاء التقدير الحصين لمراكز المجموعة $\hat{\mu}_j$ **ومصفوفة التباين المشترك** $\hat{\Sigma}_j$ **فإن قاعدة التمييز الخطية الحصينة (Robust Linear Discriminant Rule)RLDR تعطى كالآتي[2]:**

تعيين x **إلى** π_k **إذا:**

$$\hat{d}^{RL}_{k}(x)>\hat{d}^{RL}_{j}(x) \ , \ \forall\, j=1,...,g \ , \ j\ne k \quad ......(22)$$

وإن:

$$\hat{d}^{RL}_{j}(x)=\hat{d}^{RL}_{j}(\underline{x},\underline{\hat{\mu}}_j,\hat{\Sigma})=\underline{\hat{\mu}}'_j\hat{\Sigma}^{-1}\underline{x}-\frac{1}{2}\underline{\hat{\mu}}'_j\hat{\Sigma}^{-1}\underline{\hat{\mu}}_j+\ln(\hat{P}^{R}_{j}) \quad ......(23)$$

وإن احتمال $\hat{P}^{R}_{j}$ **يمكن أن يقدّر كما في (21) وإلى g=2 فإن قاعدة التصنيف (23) هي قاعدة التمييز لفشر للعينة ويمكن أن توصف كالآتي:**

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

ولبناء قاعدة التمييز الخطية الحصينة **RLDR** بعد أن ننظر إلى المقدّرات الأولية لمتوسط المجاميع ومصفوفة التباين المشترك والتي يرمز لها بـ $\hat{\mu}_{j,o}$ و $\hat{\Sigma}_o$ وبالاعتماد على $\hat{d}_j^{RL}(x,\hat{\mu}_{j,o},\hat{\Sigma}_o)$ فإن المسافة الحصينة **(Robust Distance)** :

$$RD_{ij}^{o}=\sqrt{(\underline{x}_{ij}-\underline{\hat{\mu}}_{j,o})'\,\hat{\Sigma}_o^{-1}(\underline{x}_{ij}-\underline{\hat{\mu}}_{j,o})}\qquad ......(24)$$

ولكل مشاهدة في المجموعة **j** نفرض:

$$w_{ij}=\begin{cases}1 & \text{if}\quad RD_{ij}^{o}\le\sqrt{\chi^2_{P,0.975}}\\ 0 & \text{other wise}\end{cases}$$

التقديرات نحصل عليها كمتوسط ومصفوفة التباين المشترك المدمجة للمشاهدات مع وزن **1** وكالآتي:

$$\hat{\mu}_j=\sum_{i=1}^{nj}w_{ij}x_{ij}$$

$$\hat{\Sigma}=\frac{\sum_{j=1}^{g}\sum_{i=1}^{nj}w_{ij}(x_{ij}-\hat{\mu}_j)(x_{ij}-\hat{\mu}_j)'}{\sum_{j=1}^{g}\sum_{i=1}^{nj}w_{ij}}\qquad ......(25)$$

وللحصول على تقديرات التباين المشترك الأولية $\hat{\Sigma}_o$ هناك ثلاثة طرق مختلفة وسنتناول في هذا البحث الطريقة الثانية فقط*

الطريقة الثانية:

فكرتها تعتمد على دمج المشاهدات**POBS(Pooled Observations)** من دون مصفوفة التباين المشترك للمجموعة، وهي تحديث للطريقة المعطاة من قبل **He & Fung (2000)** والذين أستخدموا مقدّر **S** لدالة التمييز الخطية الحصينة، ففي حالة العينتين نفرض[3]:

$$x_{11},x_{21},......,x_{n1,1} \text{ و } x_{12},x_{22},......,x_{n2,2}$$

ومن مجتمعين متوسطهما μ_1 و μ_2 و مصفوفة التباين المشترك Σ .

<u>أولاً</u>: نقدّر μ_1 و μ_2 كتقدير الموقع للمجموعتين باستخدام مقدّر **MCD** ذات الأوزان المعادة.

<u>ثانياً</u>: تدمج مراكز المشاهدات لتصبح بالشكل **Z's** كالآتي:

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

<u>ثالثاً</u>: نقدّر مصفوفة التباين Σ باستخدام مقدّر **MCD** ذات الأوزان المعادة لـ **z's** .

التقدير الجديد لـ μ_1 يصبح $\hat{\mu}_1+\hat{\delta}$

*لمزيد من التفاصيل عن الطريقتين راجع المصدر (1).

وكذلك التقدير الجديد لـ μ_2 يصبح $\hat{\mu}_2+\hat{\delta}$ تنجز هذه الخطوات التكرارية عدة مرات، لكن في حالة مقدّر MCD يتطلب خطوات إضافية لتحسين النتائج.

4.2.2 قواعد التمييز التربيعية الحصينة

قاعدة التمييز التربيعية الحصينة RQDR (Robust Quadratic Discriminant Rule) تكون بوضع x إلى π_k إذا[2] :

$$\hat{d}_k^{RQ}(x)>\hat{d}_j^{RQ}(x)\quad,\quad \forall\ j=1,\ldots g\neq k \qquad \ldots\ldots(26)$$

وإن:

$$\hat{d}_j^{RQ}=-\frac{1}{2}\ln\left|\hat{\Sigma}_{j,MCD}\right|-\frac{1}{2}\left(\underline{x}-\underline{\hat{\mu}}_{j,MCD}\right)'\hat{\Sigma}_{j,MCD}^{-1}\left(\underline{x}-\underline{\hat{\mu}}_{j,MCD}\right)+\ln\left(\hat{P}_j^R\right) \qquad \ldots\ldots(27)$$

3. الجانب التجريبي

1.3 مقدمة

غالباً ما يتم اللجوء لأسلوب المحاكاة للتأكد من تحقق 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 حجوم العينات العشوائية المفترض بها لتمثل مجتمع الدراسة وكذلك القدرة على التنوع بالأخطاء العشوائية وتنوع حالات التلويث في البيانات، بهدف التوصل إلى نتائج يعوّل عليها.

2.3 توليد المتغيرات و المتغيرات المستخدمة

تم تنفيذ تجارب المحاكاة ولثلاثة حجوم للعينات صغيرة، متوسطة وكبيرة وكالآتي n = 100, n = 60, n = 40 وكل حجم عينة هو عبارة عن المجموع الكلي لمفردات مجتمعين بحيث أن عدد مفردات المجتمع الأول تساوي عدد مفردات المجتمع الثاني، وبواقع (27) تجربة وبتكرار (1000) لكل تجربة ولكل حالة تم أخذ ثلاثة حجوم من المتغيرات التوضيحية وهي ثلاثة، ستة وتسعة ولحالة التوزيع الطبيعي للخطأ العشوائي للحالة النظيفة (Clean) بالإضافة لحالة تلويث البيانات (Contaminated) بقيم شاذة وبنسبتي تلويث 10% و20% وتم الاعتماد على الإمكانية العالية لبرنامج Matlab V.7 ، إذ يعد من البرامج ذات الامكانية المتقدمة في الجوانب الإحصائية، الرياضية والهندسية، حيث يوظف الأدوات بشكل مستقيم لبرمجة متقدمة جدا ، وتم إجراء عدة خطوات للحصول على البيانات وكما يلي:

1. **تم توليد البيانات بالشكل النظيف (غير الملوّث) للمجتمع الأول والمجتمع الثاني بالاستعانة بالدالة المكتبية Randn (والتي تعمل على توليد خلايا من الأرقام العشوائية والتي تتوزع توزيعاً طبيعياً بمتوسط 0 و تباين $\sigma^2=1$). ومن ثم تم تحويل البيانات من التوزيع N(0,1) إلى التوزيع $N(\mu,1)$ من خلال العلاقة $X=Z+\mu$ و بهذه الطريقة تم توليد بيانات المجتمع الأول بتوزيع N(-1,1) وللمجتمع الثاني بتوزيع N(1,1) ومن ثم تم توليد متجه واحد بعد إعطاء ترميز للمجتمع الأول الرقم (1) والمجتمع الثاني الرقم (2) لكي يمثل هذا المتجه المتغير y وبالتالي نحصل على مصفوفة ذات عدة أبعاد حيث يمثل عدد صفوفها (n_1+n_2) وان $n_1 n_2$, تمثل قيم العينة للمجتمع الاول والثاني على التوالي أما أعمدتها فتمثل عدد المتغيرات المستقلة X's والتي تم أخذها بعدة حالات، الحالة الأولى تم أخذ ثلاثة متغيرات والثانية ستة متغيرات والثالثة تسعة متغيرات، أما العمود الأخير فيمثل المتغيرات المعتمدة y_1, y_2 حيث ان y_1 يتراوح عددها من $(1,\ldots,n_1)$ وy_2 من $(n_1+1,\ldots,n_1+n_2)$.**

2. توليد المتغيرات بالشكل الملوّث حيث تم اتباع أسلوب توليد متغيرات التوزيع الطبيعي نفسه لتوليد البيانات النظيفة بمتوسط μ وتباين 1 ولكن تم استقطاع جزء من البيانات بنسبة 10% و 20% والتي تمثل نسب التلوث و لكل حجم عينة وللمجتمع الأول والثاني باستخدام نفس دالة التوليد Randn لكن يكون توزيع الشواذ للمجتمع الأول $N(9,1)$ والمجتمع الثاني $N(-9,1)$.

3.3 مناقشة نتائج المحاكاة

1. الحالة الطبيعية:

تم توليد البيانات وفقاً للحالة الطبيعية (Clean) ومن ثم طُبقت الطرائق غير الحصينة والحصينة عليها ويمثل الجداول رقم (1) احتمال خطأ التصنيف للبيانات المولدة النظيفة وللدوال المميزة الخطية والتربيعية للمقدّر التقليدي، مقدّر H الحصين، مقدّر S الحصين، مقدّر MCD الحصين. يلاحظ وعند استخدام المتغيرات P=9, P=6, P=3 ولحجوم العينات n=100,n=60,n=40 وجدنا أن مقدّر MCD الحصين وللدالتين الخطية و التربيعية هو الأفضل يليه من حيث الأفضلية مقدّر S الحصين ثم مقدّر H الحصين، وبشكل عام عند احتساب معيار احتمال خطأ التصنيف ولحجوم العينات المختلفة وباختلاف عدد المتغيرات التوضيحية يلاحظ بأن الدالة المميزة التربيعية باستخدام مقدّر MCD الحصين أعطت أقل احتمال لخطأ التصنيف.

2. حالة تلويث البيانات:

تم توليد البيانات وفقاً لنسب التلويث المعتمدة 10% , 20% وتم تطبيق الطرائق غير الحصينة والحصينة عليها وتمثل الجداول رقم (2) ,(3) احتمال خطأ التصنيف للبيانات الملوّثة بنسبتي تلويث 20%, 10% فعند استخدام المتغيرات P=9,P=6,P=3 ولحجوم العينات n=100,n=60, n=40 ولحالة نسبة التلويث 10% و 20% يلاحظ بأن مقدّر MCD الحصين وللدالتين الخطية والتربيعية هو الأفضل يليه من حيث الأفضلية مقدّر S الحصين ثم مقدّر H الحصين. وبشكل عام عند احتساب معيار احتمال خطأ التصنيف ولحجوم العينات المختلفة باختلاف عدد المتغيرات التوضيحية يلاحظ بأن الدالة المميزة التربيعية باستخدام مقدّر MCD الحصين أعطت أقل احتمال لخطأ التصنيف.

جدول رقم (1)

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	0.1722	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 (2)

احتمال خطأ التصنيف للبيانات الملوثة وبنسبة 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 (3)

احتمال خطأ التصنيف للبيانات الملوثة و بنسبة 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

4. الاستنتاجات

1. عند توليد البيانات للحالتين النظيفة والملوثة ولجميع نسب التلويث ومختلف المتغيرات التوضيحية، لوحظ أن مقدّر MCD الحصين حقق نتائج كفوءة بدرجة عالية إذا أعطى أقل احتمال لخطأ التصنيف، وللدالة التربيعية QMCD تليها الدالة الخطية LMCD، يليه مقدّر S الحصين ثم مقدّر H الحصين، كما أن مقدّر MCD الحصين أعطى ولحالات عديدة احتمال خطأ تصنيف صفر وهذا يعكس خصائص المقدّر 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 أن الشواذ لها تأثير كبير على قاعدة التمييز الخطية LDR وقاعدة التمييز التربيعية QDR.
3. أعطى مقدّر S الحصين نتائج أقل كفاءة بالمقارنة مع مقدّر MCD الحصين بالاضافة إلى أن خوارزمية مقدّر S احتاجت وقت حساب أطول مقارنة مع خوارزمية مقدّر MCD.
4. إن قواعد التمييز التقليدية LDR و QDR في حالة البيانات النظيفة أبرزت بشكل جيد ولجميع حجوم العينات ومختلف المتغيرات التوضيحية، في حين أخفقت وبشكل واضح لحالة البيانات الملوثة.
5. إن قواعد التمييز بالاعتماد على مقدّر MCD الحصين أنجزت بشكل عالٍ الكفاءة ولمجموعة البيانات الصغيرة والمتوسطة والكبيرة.
6. أثبتت الطريقة الثانية لتقدير التباين المشترك الأولي $\hat{\Sigma}_0$ والخاصة بمقدّر MCD الحصين كفاءة عالية علماً بأن هذه الطريقة هي تحديث للطريقة المعطاة من قبل He& Fung (2000).

5. التوصيات

1. نوصي باستخدام مقدّر MCD الحصين بقوة لما يمتاز به من تحقيق نقطة انهيار عالية، حصانة عالية و كفاءة إحصائية عالية. كما أن المسافة الحصينة باستخدام مقدّر MCD هي أكثر دقة من الاعتماد على مقدّرات أخرى مثل MVE.
2. بالنظر للمعوقات في حساب مقدّر MCD الحصين بقوة لا سيما مع زيادة عدد المتغيرات التوضيحية، من تلك المعوقات وقت الحساب لذلك نقترح تطبيق خوارزمية FAST-MCD التي تختزل وقت حساب خوارزمية MCD.
3. نوصي بتطبيق الطريقتين الاولى والثالثة والخاصة لمقدر التباين المشترك الأولي $\hat{\Sigma}_0$ لمقدّر MCD الحصين ومقارنة النتائج مع الطريقة الثانية لبيان إن كان هناك تقارب أو اختلاف في النتائج.
4. لوحظ أن معظم بحوث حالة تمييز الانماط (Pattern Recognition) تقتصر على استخدام الطرائق التقليدية في التمييز مثل دالة التمييز الخطية (LDF) ودالة التمييز التربيعية (QDF)، لكن عند الابتعاد عن الأساس النظري الصرف يلاحظ أن معظم البيانات تحتوي على قيم شاذة بدرجات متفاوتة وبالتالي ستكون النتيجة تحريف خطير في النتائج، لذا يوصى باعتماد مقدّرات حصينة لا تتأثر بالشواذ بالنظر لكون موضوع تمييز القوالب من التقنيات المتقدمة التي تربط علم الإحصاء بعلم الحاسوب.

6. المصادر

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

2. Croux, C., Dehon, C., (2001),"Robust Linear Discriminant Analysis Using S-Estimator", The Candian Jornal of Statistics, vol.29, 1-18.
3. He, X., Fung, W.K., (2000),"High Breakdown Estimations with Application to Discriminant Analysis ", Journal of Multivariate Analysis, vol.72, 151-162.
4. Hubert, M., Driessen, K.V., (2004),"Fast and Robust Discriminant Analysis", Computational statistics and Data Analysis, vol .45, 301-320.
5. Joossens, k., Croux, C., (2005),"Empirical Comparison of the Classification performance of Robust Linear and Quadratic Discriminant Analysis ", http://www.econ .kuleuven.ac.be
6. Launer, R.L & Wilkinson, G.N.,(1979),"Robustness in Statistics", Academic Press ,New York.
7. Mardia, K.V., Kent, J.T & Bibby, J.M., (1979),"Multivariate Analysis", Academic press Inc (London) Ltd.
8. Ruppert, D., (1992),"Computing S-Estimators for Regression and Multivariate Location /Dispersion", American Statistical Association, vol.1, no.3, 253-270.