استخدام دالة التمييز الخطية لتصنيف مرضى الثلاسيميا في مستشفى كركوك العام

أ.د.صلاح حمزة عبد* م.م.أسيل عبد الرزاق**

المستخلص :

تم في هذا البحث تطبيق دالة التمييز الخطية على 134 مريض بالثلاسيميا في مستشفى كركوك العام لغرض بناء منظومة معادلات لغرض التصنيف .أن منظومة المعادلات قد كانت مقبولة بنسبة تصنيف صحيح تبلغ 76.1% .

Abstract:

In this Re

search we applied linear discriminate function on 134 patients of Thelassemiea in general hospital of Kirkuk to construct a system of equations for classification process.

* استاذ/ الجامعة المستنصرية / كلية الادارة والاقتصاد/ قسم الاحصاء

* * مدرس مساعد/ الجامعة المستنصرية/ كلية الادارة والاقتصاد/ قسم الاحصاء

مقبول للنشر بتاريخ 2009/9/30

Our estimated system of equations is reasonable since it 76.1% classification rate.

1: المقدمة:

يعتبر التصنيف المبني على دالة التمييز الخطية احدى ابرز مواضيع متعدد المتغيرات المستخدمة في الاحصاء الحيوي، ولعل اهميته تنبع من امكانية استخدامة من خلال متغيرات يتحصل عليها بسهولة لغرض تصنيف المرضى حسب شدة مرضهم الى احدى درجات الشدة وبسرعة ، على انه لو اريد تصنيف هؤلاء المرضى بغير ذلك ، اي بفحوصات طبية لكانت اعقد من ناحية الاجراءات واطول مدة واكثر كلفة ، ولربما تؤدي بحياة المرض ثمنا لذلك .

هنالك العديد من الادبيات العلمية التي تتناول موضوع التصنيف المبني على دالة التمييز الخطية ، سنذكر احدثها تجنبا للاطالة ،ففي عام 1993 نشر كل من Taharazako, Chiga, Morita الباحثين في جامعة كاكوشي بحثا عن استخدام دالة التمييز الخطية في تمييز خط اللغة الصينية الشمالية القديمة المستخدمة كلغة في البلاط الامبراطوري الصيني على المستوى البصري . وفي عام 1998 تم بناء خوارزمية على اساس دالة التمييز الخطية لتمييز الانماط في الانفجارات النووية وذلك من قبل ke وزملاءهُ Daizhi ,Jauan , Fei ، وفي عام 2002 نشر الباحث الهولندي Klinken بحثا في مجلة التأمين المالي طبق فيه دالة التمييز الخطية في حقل التأمين على الحوادث الصناعية , وفي سنة 2006 نشر الباحث Joo بحثا تناول فية استخدام دالة التمييز الخطية قي الات الدعم الموجه يدويا لحمل الاشياء الثقيلة ، واما في عام 2008 فقد قام كل من Ma , Ru , Zang , Liu , Ceu , بتقدير نوعية صفحات الويب في الانترنيت مستخدمين دالة التميز الخطية لتصنيف الصفحات الاعتيادية والصفحات ذات الاهمية على وفق عقد مبرم بين هؤلاء الباحثين مع المؤسسة القومية الصينية للبحث والتطوير.

هدف البحث :

يهدف هذا البحث الى استخدام متغيرات الوزن والطول والعمر واكياس الدم لتصنيف مرضى الثلاسيما الى احد ثلاث درجات ، شديدة ومتوسطة وواطئة بأستخدام دالة التمييز الخطية وذلك

استنادا على بيانات متحصل عليها لمرضى مصابين بالثلاسيميا في مستشفى كركوك العام عددهم 134 مريض 93 منهم مصاب بالثلاسيميا ذات الدرجة الشديدة و 19 منهم مصاب بالثلاسيميا ذات الدرجة الواطئة اما البقية اللذين يبلغ 22 مريضا فمصابين بالثلاسيميا متوسطة الدرجة .

2 : التصنيف باسخدام دالة التمييز الخطية :

اذا افترضنا ان احتمال كون مشاهدة معينة تعود إلى المجتمع πi هو qi $(i=1,2)$ حيث أن $q1+q2=1$ وأن التوزيع الاحتمالي للمجتمع π_i هو توزيع مستمر (بدون فقدان صفة العمومية لأن المعالجة الواردة فيما يلي هي نفسها في حالة التوزيع المتقطع) ، بدالة كثافة احتماية هي $P_i(x)$ $(i=1,2)$, فأن احتمال صحة تصنيف مشاهدة عائدة في الحقيقة الى المجتمع π_i ستكون

$$p(1/1, \quad R) = \int_{R_1} p_1(x)\,dx$$

حيث أن R تمثل كل المنطقة التي يمكن التصنيف اليها ، وان R_1 تمثل منطقة التصنيف العائد للمجتمع π_1 وان

$$d\ \underline{x} = dx_1 \quad dx_2 \ldots\ldots\ldots dx_p$$

كما أن احتمال سوء تصنيف مشاهدة عائدة في الحقيقة الئ المجتمع π_i هو :

$$p(2/1, \quad R) = \int_{R_2} p_1(x)\,dx$$

حيث أن R_2 تمثل منطقة التصنيف العائد للمجتمع π_2.

وبالمثل , فأن احتمال صحة تصنيف مشاهدة عائدة في الحقيقة الئ المجتمع π_2 ستكون :

$$P(2/2,R)= \int_{R_2} P2(\underline{x})d\,\underline{x}$$

وان احتمال سوء تصنيف هذه المشاهدة سيكون:

$$P(1/2,R)= \int_{R_1} P2(\underline{x})d\,\underline{x}$$

ومما ورد أعلاه فأنه يمكن القول بأن احتمال سحب مشاهده من المجتمع π_i (i=1,2) ومن ثم صحة تصنيفها إليه هو عبارة عن $q_i\ p(i/i\ ,R)$ (i=1,2) وبالمثل فأن احتمال سحب مشاهده من المجتمع π_i ومن ثم تصنيفها بالخطأ الـ π_2 هو $q_1\ p(1/2\ ,R)$, وان احتمال سحب مشاهدة من المجتمع π_2 ومن ثم تصنيفها بالخطأ الـ π_i هو $q_2\ p(1/2,R)$

ولغرض الحصول علئ معيار تصنيف جيد فأن الأسلوب الرياضي المتبع يتمثل بتقليص متوسط الخسائر الناجمة عن الكلف المترتبة عن سوء التصنيف والمتمثل بالشكل:

$$C(2/1)p(2/1,R)q_1+C(1/2)p(1/2,R)q_2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

وذلك من خلال تقسيم الحيز R إلى حيزين مثل R_1 و R_2 بشكل يجعل الدالة (1) أعلاه اقل ما يمكن .

وفي الحقيقة ؛ فأنه إذا كانت q_1 و q_2 معلومتان فأن أسلوب تقليص الداله (1) سيدعى بأسلوب بيز Bayes procedure

وهذا ومن الجدير بالذكر بان الوجهه الرياضيه الاحتماليه الاخرى في النظر الى المسألة هي عند عدم وجود احتمالات اوليه ؛وفي هذه الحاله فأن الخساره المتوقعة في تصنيف مشاهده متأنية اصلا من المجتمع π_1 ستكون :

$$r(1,R)=C(2/1)P(2/1R)$$

وأن الخسارة المتوقعة في تصنيف مشاهدة متأنية اصلا من المجتمع π_2 ستكون :

$$r(2,R)=C(2/1)P(2/1R)$$

فيقال بأن اسلوب تقسيم منطقة التصنيف *R هو على الاقل بجودة اسلوب تقسيم منطقة التصنيف R اذا كان $r(2,R^*) \le r(2,R)$ و $r(1,R^*) \le r(1,R)$ كما ويقال بأن اسلوب تقسيم منطقة التصنيف R هو افضل من اسلوب تقسيم منطقة التصنيف R^* اذا كانت على الاقل واحدة من المتراجحتين اعلاه هي متراجحه تامه .

2.1 احتمالات سوء التصنيف : Misclassification Probabilities

يمكن احتساب احتمالات سوء التصنيف من خلال متغير التمييز الخطي :

$$y = (\underline{\mu}_1 - \underline{\mu}_2)' \sum^{-1} \underline{x}$$

وذلك اذا افترضنا ان $\underline{x}$ تعود الى المجتمع (i=1,2)i فيكون :

$$Y \approx \{(\underline{\mu} N_1 - \sum^{-1} \underline{\mu_1})'_2 \underline{\mu}$$

$$, \quad \delta^2 = (\underline{\mu_1} - \underline{\mu_2})' \sum^{-1} (\underline{\mu_1} - \underline{\mu_2}) \}$$

فيكون احتمال سوء التصنيف هو :

$P_{21}=p$ { $\underline{\chi}$ يعود للمجتمع الأول تصنيف $\underline{\chi}$ إلى المجتمع الثاني }

$$P(Y \le (\underline{\mu_1} - \underline{\mu_2})' \sum^{-1} (\underline{\mu_1} + \underline{\mu_2})) =$$

$$= p\left[Z \le \frac{\frac{1}{2}(\underline{\mu_1} - \underline{\mu_2})' \sum^{-1} (\underline{\mu_1} + \underline{\mu_2}) - (\underline{\mu_1} - \underline{\mu_2})' \sum^{-1} \underline{\mu_1}}{\sqrt{(\underline{\mu_1} - \underline{\mu_2})' \sum^{-1} (\underline{\mu_1} - \underline{\mu_2})}} \right]$$

$$= p\left[Z \le \frac{(\underline{\mu_1} - \underline{\mu_2})' \sum^{-1} \left[\frac{1}{2}\underline{\mu_1} + \frac{1}{2}\underline{\mu_2} - \underline{\mu_1} \right]}{\partial} \right]$$

$$=p\left[z \le -\frac{1}{2}\partial \right] = \Phi(-\partial/2)$$

{ $\underline{\chi}$ يعود للمجتمع الثاني تصنيف $\underline{\chi}$ إلى المجتمع الأول} $= P_{12}=P$

2.2 التصنيف في حالة عدة مجاميع :

سنناقش في هذه الفقرة مسألة تصنيف مشاهدة معينه الى واحد من k من المجتمعات التي تتوزع توزيعا طبيعيا متعدد المتغيرات بمتجهات المتوسطات $\underline{\mu_k}$ ..... ، $\underline{\mu_2}$ ، $\underline{\mu_1}$ على التوالي ومصفوفة تباين مشتركة لكل المجتمعات $\sum$ وفي حالة كون معلمات هذه المجتمعات مجهولة فأنه يمكن استخدام تقديراتها

$$\underset{-j}{\hat{\mu}} = \underset{-j}{\bar{\chi}}$$

$$S=\frac{1}{N-K}\sum_{J=1}^{K} Aj$$

وان قاعدة التصنيف ستكون بأن نصنف المشاهدة $\underline{\chi}$ الى المجتمع i اذا كان :

$$D_i^2 = Min\{D_1^2, D_2^2, ..., D_K^2\}$$

$$D_i^2 = (\underline{\chi} - \underset{-ji}{\bar{\chi}})' S^{-1} (\underset{-}{\chi} - \underset{-Ji}{\bar{\chi}})$$

3: الجانب التطبيقي :

الثلاسيميا (تسمى ايضا انيميا البحر الابيض المتوسط) من اهم امراض الدم الوراثية الانحلالية التي تسبب تكسر كريات الدم الحمراء الشائعة على مستوى العالم بشكل عام وعلى

مستوى منطقة البحر الابيض المتوسط بشكل خاص والثلاسيميا نوعان الالفا والبيتا وحين يأتي الحديث فيما يلي عن هذا المرض فأننا نقصد به (البيتا ثلاسيميا).

وان كلمة الثلاسيميا يونانية الأصل تعني فقر دم منطقة البحر الأبيض المتوسط حيث إن هذا المرض عرف وأشتهر في هذه المنطقة بشكل كبير ويعرف أيضا بأسم انيميا البحر المتوسط (Mediterranean anemia) وفي الولايات المتحدة الأمريكية كان يعرف باسم انيميا كوليز (cooleys anemia) وقد تم اخذ 134 مشاهدة من البيانات في عام 2005 من مستشفى كركوك العام لاستخدامها في الدراسة التحليلية للبحث هي الظاهرة بتفاصيل متغيراتها في الجدول رقم (3) في الملحق.

ولبيان الكيفية التي يظهر بها المرض , نقول بأن الهيموغلوبين عبارة عن بروتين موجود في كريات الدم الحمراء وظيفتة نقل الاوكسجين من الرئة الى كافة خلايا الجسم ويتكون من الهيم (مادة الحديد+صبغة) والغلوبين (البروتين) وهناك انواع مختلفة من بروتينات الغلوبين المهمة منها الالفا , البتا , الدلتا و الكاما غلوبين وتحدث المشكلة عندما يحدث نقص في مكونات بروتين الغلوبين بسبب خلل جيني وكل زوج من الجينات الاتية من الام والاب يعمل على تنظيم ناتج بروتينات الغلوبين في الجسم فمثلا اذا كان الجين المسؤول عن بروتينات البيتا غلوبين به خلل فلن يكون ناتج التكوين كاملا وصحيحا مما يؤدي الى خلل في تكوين سلاسل البيتا وبالتالي يؤدي الى مرض البيتا ثلاسيميا ، كما انه اذا كان نفس الخلل موجود في بروتين الالفا غلوبين فأنه ينتج مرض الالفا ثلاسيميا .

تحليل النتائج :

تم استخدام نظام MINITAB بنسخته الـ(13) لتحليل البيانات المتحصل عليها حول مرضى الثلاسيميا في مستشفى كركوك العام والواردة في ملحق هذا البحث , سنقوم بعرض وتحليل النتائج على وفق ما يلي:-

1 : يتبين من الجدول رقم (1) مستخلص للتصنيف المبني على اساس دالة التمييز الخطيه للمرضئ ال(134), أذ نلاحظ منه بأن 89 مريض ممن كانوا مصابين بثلاسيميا ذات الدرجة الشديدة فعلا , صنفوا وفق دالة التمييز الخطية على هذه الدرجة من الاصابة ايضا اي ان نسبة التمييز الصحيح هنا تبلغ 0.957 اما الاربعة المتبقين لاتمام ال 93 المصابين فعلا بالثلاسيميا على وفق هذه الدرجة

فقد صنف واحد منهم بأستحدام دالة التمييز الخطيه للدرجة الواطئه من الاصابة في حين صنف الثلاثة المتبقين للدرجة المتوسطة من الاصابة بهذا المرض .

نفس الشئ بالنسبة للمجموعة الاخرى من المصابين , الآ وهم المصابين بالثلاسيميا بدرجته الواطئه , فعدد هولاء الفعلي هو 19 , سبعة منهم صنفو بشكل صحيح على وفق دالة التمييز الخطيه , اما البقية فقد صنف 6 منهم كمرضئ من الدرجة الشديدة و 6 منهم كمرضئ من الدرجة المتوسطة .

جدول رقم (1) يبين فيه مستخلص للتصنيف

المجموعة	الاصابة الشديدة	الاصابة الواطئة	الاصابة المتوسطة
Highest	89	6	12
Lowest	1	7	4
Middle	3	6	6
Total N	93	19	22
N Correct	89	7	6
Proportion	0.957	0.368	0.273

اما المصابين بمرض الثلاسيميا بدرجته المتوسطة والباغ عددهم 22 فقد صنف 6 منهم بشكل صحيح اما البقية فقد صنف 12 منهم كمرضئ من الدرجة الشديدة و 4 منهم كمرضئ من الدرجة الواطئه .

2 : يبلغ عدد المصنفين الكلي بشكل صحيح 102 من مجموع 134 وبنسبة تصنيف صحيحة تبلغ 76.1%

3 : مصفوفة المسافة بين الدرجات الثلاثة لشدة الاصابة هي كما يلي في ضوء المتغيرات المأخوذة لغرض التصنيف , الطول , الوزن , العمر , واكياس الدم .

	Highest	Lowest	Middle
Highest	0.00000	3.67842	2.62245
Lowest	3.67842	0.00000	1.01222
Middle	2.62245	1.01222	0.00000

4 : منظومة معادلات التصنيف الثلاثة هي :

$$H = -18.141 - 0.168w - 0.531o + 0.369h + 0.918b$$

$$L = -17.561 - 0.081w - 0.282o + 0.318h + 0.736b$$

$$M = -18.103 + 0.038w - 0.558o + 0.329h + 0.704b$$

حيث أن M , L , H تمثل على التوالي الدرجات الشديدة والواطئه والمتوسطة للاصابة بمرض الثلاسيميا وأن w عبارة عن متغير الوزن و o عبارة عن متغير العمر و h عبارة عن متغير الطول و b عبارة عن متغير اكياس الدم.

5 : يتبين في الجدول رقم 2 المشاهدات التي تعاني من سوء التصنيف مع درجة المرض الفعلية والدرجة المصنف لها المريض بناءا على دالة التمييز الخطية من خلال المسافة المربعة واحتمال التصنيف .

4: المصادر:

1. **Optical grading of Satsuma mandarin using LDF kazoo morita , shiga , T. and taharazako , S. (1993) Mem. Fac. Agr . kagoshina univ. vol.29 pp.101-111.**

2. **Anew algorithm for linear discriminate function and its application to pattern recognition of nuclear explosion , Zhao ke & wang fei &**

su juan & liu daizhi signal processing proceedings 1998 ICSP , apos 98 Forth internations conference vol. 2 pp.1237-1240

3. **on the use of linear discriminate function in the Realm of industrial accident insurance J.van Klinken Amsterdam Journal of fin an. Insure. 2002 , vol. 69,no.4, pp-202-219.**

4. **SVM support vector machines Linear discriminate function and SVM seong – wook joo Journal of Soc. Of Indus. (2006) vol. 83 ,no. 4 ,pp.1215 – 1221.**

5. **Cen , R. and liu ,y. and Zhang M . and Ru , L. and ma, sh. Web page quality estimation based on LDF State key lab. Of intelligent technologies &system Tsinghua univ. (2008).**

6: الجبوري , شلال وعبد , صلاح (2000) "تحليل متعدد المتغيرات" مطبعة جامعة بغداد, العراق.

5: الملحق:

جدول رقم (3)

ت	عدد اكياس الدم	الثلاسيميا	الطول	العمر	الوزن
1	2	Middle	155	19	42.0
2	4	Middle	161	23	50.0

3	4	Middle	152	18	45.0
4	1	Highest	165	17	40.0
5	4	Lowest	154	19	40.0
6	4	Lowest	157	16	49.0
7	2	Highest	100	8	20.0
8	4	Highest	153	17	56.0
9	4	Highest	123	9	19.0
10	4	Middle	131	14	27.0
11	2	Highest	93	6	15.0
12	1	Middle	165	31	75.0
13	3	Lowest	164	40	58.0
14	4	Lowest	162	42	53.0
15	4	Highest	80	3	12.0
16	2	Highest	91	4	12.0
17	2	Highest	93	6	14.0
18	2	Middle	159	39	69.0
19	3	Lowest	98	5	14.0
20	1	Lowest	171	29	70.0
21	1	Highest	95	5	13.0
22	1	Lowest	95	3	14.0
23	1	Highest	119	9	18.0
24	1	Lowest	89	7	18.0
25	1	Highest	115	4	16.0
26	1	Middle	167	20	57.0
27	1	Highest	104	6	14.0
28	4	Middle	110	9	18.0
29	1	Lowest	119	13	21.0
30	3	Lowest	130	15	28.0
31	2	Highest	106	3	15.0
32	1	Lowest	11	10	20.0
33	1	Middle	174	40	70.0
34	2	Highest	113	11	18.0
35	3	Highest	83	2	12.0
36	4	Middle	134	15	27.0
37	4	Highest	125	14	25.0
38	4	Highest	112	7	20.0
39	3	Middle	132	15	56.0
40	3	Highest	129	14	25.0
41	4	Highest	117	9	22.0
42	4	Highest	115	10	20.0
43	4	Highest	114	10	22.0
44	4	Highest	120	10	19.0
45	4	Middle	127	13	25.0
46	2	Middle	11	7	20.0

47	4	Lowest	128	15	26.0
48	4	Lowest	140	17	34.0
49	1	Lowest	148	30	50.0
50	1	Middle	120	11	23.0
51	2	Highest	140	17	38.0
52	3	Lowest	139	18	40.0
53	3	Lowest	143	19	33.0
54	1	Middle	129	15	26.0
55	1	Middle	126	13	24.0
56	4	Highest	132	16	28.0
57	4	Highest	128	15	26.9
58	4	Highest	129	15	24.0
59	4	Highest	115	13	20.0
60	4	Highest	116	12	23.0
61	4	Highest	114	12	20.0
62	4	Highest	116	12	23.0
63	4	Highest	115	11	20.0
64	3	Highest	116	11	22.0
65	4	Highest	118	10	20.0
66	4	Highest	119	10	22.0
67	4	Highest	120	10	18.0
68	4	Highest	105	9	19.0
69	4	Highest	114	9	20.0
70	2	Highest	110	8	18.0
71	4	Highest	107	7	15.0
72	4	Highest	114	9	23.0
73	4	Highest	107	6	16.0
74	4	Highest	102	7	15.0
75	4	Highest	130	14	24.0
76	4	Highest	103	8	16.0
77	4	Highest	114	9	20.0
78	4	Highest	100	6	16.0
79	4	Highest	103	3	14.0
80	4	Highest	93	3	14.0
81	2	Highest	91	3	13.0
82	2	Highest	128	15	23.0
83	2	Highest	103	6	15.0
84	2	Highest	96	4	15.0
85	1	Highest	119	9	22.0
86	1	Highest	86	2	11.0
87	1	Highest	87	2	12.0
88	3	Lowest	160	47	54.0
89	1	Lowest	158	34	48.0
90	2	Highest	129	15	26.0

91	4	Highest	122	12	24.0
92	4	Highest	121	11	25.0
93	4	Highest	118	10	21.0
94	4	Highest	117	10	19.0
95	4	Highest	120	9	22.0
96	4	Highest	119	8	23.0
97	4	Highest	119	8	21.0
98	4	Highest	120	8	19.0
99	4	Highest	124	8	20.0
100	1	Highest	131	16	29.0
101	3	Highest	104	7	17.0
102	3	Highest	100	6	14.0
103	4	Highest	122	12	24.0
104	4	Highest	122	11	23.0
105	3	Highest	102	5	13.0
106	3	Highest	126	14	24.0
107	4	Highest	120	10	22.0
108	4	Highest	131	15	26.0
109	2	Highest	100	6	16.0
110	3	Highest	120	12	23.0
112	1	Highest	119	8	21.0
111	1	Highest	130	15	25.0
113	1	Highest	116	9	22.0
114	3	Middle	104	6	15.0
115	3	Middle	97	4	16.0
116	4	Middle	97	4	16.0
117	1	Middle	129	15	25.0
118	4	Highest	112	7	18.0
119	3	Middle	93	3	14.0
120	3	Highest	99	5	15.0
121	4	Highest	135	13	27.0
122	3	Highest	97	4	13.0
123	1	Middle	157	29	55.0
124	3	Highest	124	9	21.0
125	1	Highest	112	2	10.0
126	4	Lowest	154	20	47.0
127	1	Highest	77	3	9.0
128	2	Highest	119	5	18.0
129	1	Highest	98	4	15.0
130	1	Highest	96	2	12.0
131	1	Highest	94	8	15.0
132	1	Highest	83	1	10.0
133	1	Highest	93	2	13.0
134	1	Highest	97	3	17.0

.....

.....

.....