International Science and Technology Journal
المجلة الدولية للعلوم والتقنية
Volume 34 العدد
Part 2 المجلد
July 2024 يوليو
المجلة الدولية للعلوم والتقنية
ISTJ
تم استلام الورقة بتاريخ:5/ 6/2024م وتم نشرها على الموقع بتاريخ: 2/ 7/2024م

Using linear discriminant analysis to classify renal failure
(Applied study in Tobruk Medical Center)

www.doi.org/10.62341/ehms5627

Enas H Abduallah **Mahmoud Saed Asbeetah**

enas.hamed@tu.edu.ly mahmoud.asbeetah@tu.edu.ly
Mathematics department, Tobruk University, Tobruk, Libya.

Abstract
This research aimed at classifying renal failure diseases whether acute renal failure (ARF) or chronic renal failure (CRF) based on linear discriminant analysis (LDA). The research was carried out by collecting data from patients with renal failure at Tobruk Medical Center (TMC). The sample was composed of 461 cases of renal failure, 238 cases of them were acute renal failure and 223 cases were chronic renal failure taking into account a set of crucial factors, including age, fasting blood sugar, urine, etc. The discriminant function analysis was applied using the R programming language. The study revealed that the linear discriminant analysis classified the two types and indicated that the variables with the highest impact on renal failure were creatinine and fat blood sugar. Also, the LDA was highly precise in classification, where 88.5% of the sample was classified successfully, which means a small error rate of 11.5%.

Keywords: linear Discriminant Analysis, renal failure, classification.

International Science and Technology Journal
المجلة الدولية للعلوم والتقنية
Volume 34 العدد
Part 2 المجلد
July 2024 يوليو
المجلة الدولية للعلوم والتقنية
ISTJ
تم استلام الورقة بتاريخ:5/ 2024/6م وتم نشرها على الموقع بتاريخ: 2/ 2024/7م

استخدام التحليل التمييزي الخطي لتصنيف مرضى الفشل الكلوي

(دراسة تطبيقيه في مركز طبرق الطبي)

إناس حمد عبدالله[1]، محمود سعيد اسبيته[2]
قسم الرياضيات / جامعة طبرق ليبيا
enas.hamed@tu.edu.ly , mahmoud.asbeetah@tu.edu.ly

الملخص

يهدف هذا البحث الى تصنيف الفشل الكلوي بعد تعريفه وتقسيمه الى نوعين وهما القصور الكلوي والفشل الكلوي التام وذلك باستخدام التحليل التمييزي الخطي وجمعت البيانات من ملفات المرضى الإلكترونية بمركز طبرق الطبي، المتغيرات تم تحديدها وفقا لسجلات المرضى وبعد استشارة بعض الأطباء المختصين وهي كسبيل الذكر (العمر، سكر الدم، اليوريا، الكرياتينين، الالبومين، حمض اليوريا، الفوسفات.....)، العينة بلغت 461 مريض بالفشل الكلوي قسمت الي عينتين 238 مريض بالقصور الكلوي و223 مريض بالفشل الكلوي التام.

أوضحت النتائج وذلك باستخدام لغة البرمجة **R** المتغيرات الأكثر تأثيراً على مرض الفشل الكلوي وهي الكرياتينين وسكر الدم وكذلك ملائمة استخدام نموذج التحليل التمييزي الخطي لتصنيف مرضى الفشل الكلوي كذلك كفأه الدالة في التصنيف بنسبة 88.5% وذلك يعني نسبة ضئيلة في خطأ التصنيف بنسبة 11.5%.

الكلمات المفتاحية: التحليل التمييزي الخطي، الفشل الكلوي، التصنيف.

Introduction

Discriminant analysis is used to describe or elucidate the variables as well as differences between two or more groups. The goals of descriptive discriminant analysis include identifying the relative contribution of the p-values to the separation of the groups and

تم استلام الورقة بتاريخ:5/ 6/2024م وتم نشرها على الموقع بتاريخ: 2/ 7/2024م

finding the optimal plane on which the points can be projected to illustrate the configuration of the groups in the best way [1].
The nature of the discriminant analysis is exploratory. In addition, linear discriminant analysis as an instrument for separating groups is used to discover cases of differences when the causation is not defined [2].
This study, in turn, focuses on using linear discriminant analysis to classify acute renal failure (ARF) and chronic renal failure (CRF), which would help doctors and others who are concerned about renal failure diseases.

Literature review

1-Pohar and Blas (2004): The study included a comparison between linear discriminant analysis (LDA) and linear logistic discrimination analysis (LLD) by simulation. The study shows that linear discriminant analysis is used when the variable is normally distributed. However, the linear logistic discrimination analysis is applied when the sample space is small and the normality of the variable is not conditional. In addition, the results show that the results for both methods are nearly equal [3].

2- Roush and Kelly (2009):This study aimed to compare linear discriminant analysis (LDA), linear logistic discrimination analysis (LLD), LDA based on rank, and mixture discriminant analysis (MDA), depending on studying Monte Carlo simulation. The study found that linear discriminant analysis and linear logistic discrimination analysis have the same accuracy of classification. In addition, the result shows that mixture discrimination analysis is more accurate in classification, especially if the data is not normally distributed. Furthermore, LDA based on rank is the most accurate classification over linear discriminant analysis and linear logistic discrimination analysis [4].

3-Abdul Hussein (2019):The study was designed to discriminate between two groups (infected and non-infected) by heart diseases using Mahalanobi's formula D^2.In addition, the researcher derived a role from Mahalanobi's formula for discrimination named with R_{D^2}.The research concluded that Mahalanobi's formula is important for discriminating between two groups[5].

تم استلام الورقة بتاريخ:5/ 2024/6م وتم نشرها على الموقع بتاريخ: 2/ 2024/7م

Medical aspect

Kidney is an important vital principle in the human body that abstracts products from the blood, such as nitrogenous waste, and exogenous molecules, such as drugs, in addition to regulation levels of electrolytes, participated in the synthesis of erythropoietin hormone and Metabolism of proteins that are low molecular weight, such as insulin.

Renal failure is defined as the feebleness of the kidney to achieve its excretion function, which causes the preservation of nitrogenous waste products from the circulation.

The renal failure was divided into two parts

1- Acute Renal Failure (ARF): This type is reversible and is intimate when the kidneys occur suddenly in the blood supply or in cases of toxins overload, and this will lead to loss of kidney function unexpectedly. Several factors may cause Acute Renal Failure, 60% of cases are caused by Hypotension, sepsis, haemorrhage, failure oftheheart or liver, and several drugs. These factors known as Prerenal. Approximately 35% of cases produce from extended prerenal failure which brings about an episode of acute tubule necrosis.Another causes acute interstitial nephritis, vasculitis, rhabdomyolysis and arteriolar insults. All these conditions are known as Intrarenal. The rest 5% approximately represented as Post renal may result from several factors such as hypertrophy of prostate, tumour, calculus, carcinoma, neurogenic bladder, clot, and stricture.

2- Chronic Renal Failure (CRF)

In this type, the advancing defeat of kidney functions occurs when the creatinine levels increased for function of a minimum of 3 months or the analysed glomerular filtration rate (GFR) is below 60 ml per minute / 1.73m2 and this will lead to using dialysis or transplantation and this condition called end-stage renal disease(ESRD). The main causes of chronic renal failure are diabetes mellitus type2 and Hypertension.

The second common causes are Glomerulonephritis, Polycystic kidney diseases and renal vascular diseases. Other identified causes, such as nephrolithiasis, persistent obstruction of the urinary tract,

تم استلام الورقة بتاريخ:5/ 6/2024م وتم نشرها على الموقع بتاريخ: 2/ 7/2024م

Vesicoureteral reflux, pyelonephritis, and recurrent kidney infections [6].

The Linear Discriminant Function – Two Groups[1],[7]
1-Calculate the mean for the two groups by:

For the first group:

$$X_k^{-(1)} = \frac{\sum_{i=1}^{n1} X_{ki}}{n1} \qquad (1)$$
$$j = 1,2 \dots \dots \dots k$$

For the second group:

$$X_k^{-(2)} = \frac{\sum_{i=1}^{n2} X_{ki}}{n2} \qquad (2)$$
$$j = 1,2 \dots \dots \dots k$$

2-Determine the difference between means for each variable for two groups by:

$$d_j = X_j^{-(1)} - X_j^{-(2)} \qquad (3)$$
$$j = 1,2 \dots \dots \dots k$$

Then applying the differences in vertical vector with the symbol d Where:

$$d = \begin{bmatrix} d_1 \\ d_2 \\ : \\ : \\ d_k \end{bmatrix}_{k\times 1}$$

3-Determine the sum of squares for each variable for both groups by:
For the first group:

$$S_{jj}^{(1)} = \sum_{i=1}^{n_1} x_{ji}^2 - \frac{(\sum_{i=1}^{n_1} x_{1i})^2}{n_1} \qquad (4)$$
$$j = 1,2 \dots \dots \dots n_1$$

For the second group:

تم استلام الورقة بتاريخ:5/ 2024/6م وتم نشرها على الموقع بتاريخ: 2/ 2024/7م

$$S_{jj}^{(2)} = \sum_{i=1}^{n_2} x_{ji}^2 - \frac{(\sum_{i=1}^{n_1} x_{1i})^2}{n_2} \quad (5)$$
$$j = 1,2 \dots \dots \dots n_2$$

In addition, finding the product for each two variables in every group by using the next formula:
For the first group:

$$S_{ij}^{(1)} = \sum_{i=1}^{n_1} (x_{ij} - x_i^-) - (x_{ji} - x_j^-)$$
$$= \sum_{i=1}^{n_1} x_{ij}\, \boldsymbol{x_{ji}} - \frac{(\sum_{i=1}^{n_1} x_{ij})(\sum_{i=1}^{n_1} x_{ji})}{n_1} \quad (6)$$

For the second group:

$$S_{ij}^{(2)} = \sum_{i=1}^{n_2} (x_{ij} - x_i^-) - (\boldsymbol{x_{ji}} - \boldsymbol{x_j^-})$$
$$= \sum_{i=1}^{n_2} x_{ij}\, \boldsymbol{x_{ji}} - \frac{(\sum_{i=1}^{n_2} x_{ij})(\sum_{i=1}^{n_2} x_{ji})}{n_2} \quad (7)$$

4-finding the combined variance between the two groups by:

$$V_{JJ} = \frac{S_{jj}^{(1)} + S_{jj}^{(2)}}{n_1 + n_2 - 2} \quad (8)$$
$$j = 1,2 \dots \dots \dots n$$

5- applying the combined covariance in the groups by:

$$V_{iJ} = \frac{S_{jj}^{(1)} + S_{jj}^{(2)}}{n_1 + n_2 - 2} \quad (9)$$

Where $i \neq j \quad and \quad i,j = 1,2 \dots \dots \dots k$
By using the equations 8 and 9 the variance and combined covariance is:

$$V = \begin{bmatrix} v_{11} & v_{12} \dots \dots & v_{1k} \\ v_{21} & v_{22} \dots \dots & v_{2k} \\ & \vdots & \\ & \vdots & \\ v_{k1} & v_{k22} \dots \dots & v_{kk} \end{bmatrix} \quad (10)$$

تم استلام الورقة بتاريخ:5/ 2024/6م وتم نشرها على الموقع بتاريخ: 2/ 2024/7م

6-setting the linear discriminant equation as following:

$$L = \alpha_1 x_{1i} + \alpha_2 x_{2i} + \dots\dots\dots + \alpha_k x_{ki} \quad (11)$$

Where:

$\alpha_1, \alpha_2, \dots\dots\dots \alpha_k$Factors of linear discriminant equation

$x_{1i}, x_{2i} \dots\dots\dots x_{ki}$Variablesof linear discriminant equation

It is calculated by:

$$V\alpha = d$$

Where:

$$\alpha = V^{-1}d \quad (12)$$

To determine the relative importance of each variable it is possible to use the following formula:

$$\alpha_i^* = \alpha_i\sqrt{V_{ii}} (13)$$

By comparing the absolute values of α_i^*, the largest value means that $x_k$thecontrast is the most important variable, which is able to discriminate between the two groups, the second largest value means that the contrast variable is the second important variable which is able to discriminate between the two groups.

7- Significant test of discriminant linear function:

The significance of discriminant linear function examined by**:**

First: **F test**

F test supports to examination of the ability of independent variables to impact on linear discriminate function, see table 1.

Table 1. Ftest table

Source	SS	Df	MS	F
Between x's	SSB	k-1	M_{SB}	M_{SB}
Error with x's	SSE	n-k	M_{SE}	M_{SE}
Total	SST	n-1		

Where:

1- Sum of squared errors (SSE)

$$SSE = D^2 = \hat{\alpha_1} d_1 + \hat{\alpha_2} d_2 \dots\dots\dots + \hat{\alpha_k} d_k$$

2- Sum of squares between groups (SSB)

تم استلام الورقة بتاريخ:5/ 2024/6م وتم نشرها على الموقع بتاريخ: 2/ 2024/7م

$$SSB = \frac{n_1 n_2}{(n_1 + n_2)(n_1 + n_2 - 2)} \times (D^2)^2$$

3- Total sum of squares (SST)

$$SST = SSB + SSE$$

To examine the hypothesis

H_0: The function is not able to discriminate

H_1: The function is able to discriminate

Second**: T test**

The examination is arranged up by

$H_0: \mu_{L1} = \mu_{L2}$

$$H_1 = \mu_{L1} \neq \mu_{L2}$$

By the formula:

$$t = \frac{L_1^- - L_2^-}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

8-Separation point:

The separation point is the point that separates the two groups; In addition, it is used to classify any new signaller case to any group it belongs to.

If $L_1^- > L_2^-$

Then the new case belongs to the first group if the discriminants value is

$$L > \frac{1}{2}(L_1^- + L_2^-) \qquad (14)$$

In addition, the new case belongs to the second group if the discriminants value is:

$$L < \frac{1}{2}(L_1^- + L_2^-) \qquad (15)$$

If $L_1^- < L_2^-$

تم استلام الورقة بتاريخ:5/ 2024/6م وتم نشرها على الموقع بتاريخ: 2/ 2024/7م

Then the new case belongs to the first group if the discriminants value is :

$$L < \frac{1}{2}(L_1^- + L_2^-) \quad (16)$$

In addition, the new case belongs to the second group if the discriminants value is:

$$L > \frac{1}{2}(L_1^- + L_2^-) \quad (17)$$

9-Error rate**:**

There are two types of error:

1- Virtual error rate

It is calculated according to table 2.

Table 2.The classification of virtual error rate

Real group	Followed to the first group	Followed to the second group	Total
First	n_{11}	n_{12}	n_1
Second	n_{21}	n_{22}	n_2

Where:

n_{11}: The number of cases in the first group and discriminated in the first group correctly,

n_{12}: The number of cases in the first group and discriminated in the second group incorrectly,

n_{21}: The number of cases in the second group and discriminated in the first group incorrectly,

n_{22}: The number of cases in the second group and discriminated in the second group correctly.

$$E = \frac{n_{12} + n_{21}}{n_{12} + n_{21} + n_{12}n_{22}} \quad (18)$$

E: Is the error rate

Thus, $\frac{n_{12}}{n_1}$ the error to discriminate in the first group, $\frac{n_{21}}{n_2}$ the error to discriminate in the second group

تم استلام الورقة بتاريخ:5/ 2024/6م وتم نشرها على الموقع بتاريخ: 2/ 2024/7م

2- The real error:
It is calculated according to the following formula:

$$P_{12} = p_{21} = F\left[\frac{-\sqrt{D^2}}{2}\right] \qquad (19)$$

Where:
F: Normal distribution function,
D^2:Mahalanobis formula$= \hat{\alpha_1} d_1 + \hat{\alpha_2} d_2 \dots\dots\dots + \hat{\alpha_k} d_k$
Whenever the discriminate error is small indicates the accuracy of the discriminate process.

Statistical data

Data was collected from Tobruk Medical Center records with sample size of 238 for acute renal failure and 223 for chronic renal failure by comprehensive survey to the electronics patients' files between 2022 and 2024. Furthermore, data were entered and analysed using R programming language.
By using the variables (age, fast blood sugar, urine, creatinine, albumin, uric acid, phosphate, white blood cells, platelets, and hemoglobin),the variables defined according to the patients' files, after consulting the specialists. The data was ordered and codingby table 3:

Table 3. The defination of each variable

Agex_1	Fast blood sugarx_2	Urinex_3	Creatininex_4	Albuminx_5
Uric acid x_6	Phosphate x_7	White blood cells x_8	Platelets x_9	Hemoglobin x_{10}

Result of linear discriminant

1-Determine the difference between means for each variable for two groups as the following:

تم استلام الورقة بتاريخ:5/ 2024/6م وتم نشرها على الموقع بتاريخ: 2/ 2024/7م

$$\boldsymbol{d} = \begin{bmatrix} -5.296 \\ 4.977 \\ 67.236 \\ 6.987 \\ 2.524 \\ 0.063 \\ 1.821 \\ -1.215 \\ -27.258 \\ -1.835 \end{bmatrix}$$

2-Coefficients of linear discriminants:
Calculating combined covariance by using combined variance between the two groups and setting the coefficient of linear discriminant by equation 1:

$$\begin{aligned} \boldsymbol{L} = & -1.482 - 0.0068x_1 + 0.0019x_2 + 0.0002x_3 + 0.3287x_4 \\ & + 0.0056x_5 - 0.0479x_6 + 0.01943x_7 - 0.0065x_8 \\ & - 0.0007x_9 - 0.0244x_{10} \end{aligned}$$

3-The order of the relative importance of each variable
From equation 13, the importance of each variable ordered in descending order after ignoring the negative sign is confirmed in table 4.

TAable4. The order of the importance of each variable

The order of the relative importance	Variable	$\alpha_i^* = \alpha_i\sqrt{V_{ii}}$
1-	x_4	1.501
2-	x_2	0.161
3-	x_{10}	0.151
4-	x_1	0.111
5-	x_5	0.104
6-	x_6	0.091
7-	x_9	0.062
8-	x_7	0.041
9-	x_8	0.036
10-	x_3	0.012

تم استلام الورقة بتاريخ:5/ 2024/6م وتم نشرها على الموقع بتاريخ: 2/ 2024/7م

The table illustrates the relative importance of each variable, and it is clear that the most important variable to determine the type of renal failure is creatinine x_4 next that are fast blood sugar x_2, hemoglobin x_{10}, age x_1, albumin x_5, uric acid x_6, platelets x_9, phosphate x_7 and white blood cells x_8. On the other hand, the less important variable to determine the renal failure is Urine $\boldsymbol{x_3}$

4-significant test of discriminant linear function:
First: **F test**

Table 5. F test result

Source	SS	Df	MS	F
Between x's	1.535	9	0.17	30.99
Error with x's	2.474	451	5.4×10^{-3}	
Total	4.009	460		

$$F_{0.05}(9,451) = 1.9006$$

It is clear that the computing F is more than critical value, the null hypothesis rejected the function is able to discriminate.
Second: **T test**

Table 6.T test result

Confidence interval 95%		Mean differences	P value	Df	T
Lower	Upper				
2.289703	2.658546	2.47412	2.2e-16	409.86	26.372

Since the p value of the T test is about 2.2e-16 and this value is les s than 0.5, that means rejecting the null hypothesis and accepting t he alternative hypothesis. As a result, the linear discriminant is abl e to disseminate between two groups.
5- Separation point:
L_1^-: the average of acute = -1.149
L_2^-: the average of chronic = 1.280
The separation point$L = \frac{-1.149+1.280}{2} = 0.0655$
So, in$L_1^- < L_2^-$

تم استلام الورقة بتاريخ:5/ 2024/6م وتم نشرها على الموقع بتاريخ: 2/ 2024/7م

If the new variable is less than the separation point, it belongs to the first group(ARF),but if it is bigger, it belongs to the second group (CRF).

6- Error rate

1- Virtual error rate

It is calculated according to table 7.

Table 7. Result of virtual rate

Real group	ARF	CRF
ARF	217(91.1%)	21(8.8%)
CRF	32(14.3%)	191(85.7%)

From the table it noticeable that the percentage of correct classification (88.5%), also the percentage of correct classification of acute (91.1%) and the correct classification of chronic (85.7%) and the incorrect classification of chronic and acute are (14.3%) and (8.8%) respectively. In contrast, the incorrect classification is (11.5%).This is a small incorrect classification depending on the virtual error rate.

2- Real error rate

$$D^2 = 2.474$$

$$P_{12} = p_{21} = F\left[\frac{-\sqrt{2.474^2}}{2}\right]$$

$$F[-0.786] = 0.217$$

Very small real error rate means the error of classification is very small.

Result

1- The linear discriminant analysis can classify the two types of renal failure with an 88.5% correct classification, which means an 11.5% incorrect classification.

2- The linear discriminant analysis has very low incorrect classifications of chronic and acute (14.3%) and (8.8%), respectively.

3- Creatinine is the most important variable affecting renal failure; the following variables are fast blood sugar and hemoglobin.

تم استلام الورقة بتاريخ:5/ 6/2024م وتم نشرها على الموقع بتاريخ: 2/ 7/2024م

4- The result shows a very small real error rate that leads to small error of classification.

Conclusion

The objective or the purpose of this study was toclassify renal failure diseases whether acute renal failure (ARF) or chronic renal failure (CRF) based on linear discriminant analysis (LDA). The result concluded that the linear discriminant analysis can classify the two types of renal failure with small low incorrect classification.

Recommendation

1- Improve the resources for collecting data in the Ministry of Health to enhance medical research.
2- Apply the linear discriminant analysis to other diseases to obtain more factors that impact human health and well-being.
3- Take advantage of the result of the linear discriminant analysis to decrease the reasons for renal failure.

References

[1] Alvin C. Rencher 2002. "Methods of Multivariate Analysis", Second Edition, Brigham Young University.

[2] جونسون ريتشارد، وشرن دين 1998. "التحليل الاحصائي للمتغيرات المتعددة من الوجهة التطبيقية" تعريب دكتور عبد المرضي حامد عزام، دار المريخ للنشر– الرياض.

[3] Pohar, M., Blas, M., 2004, Comparison of logistic regression and linear discriminant analysis: Asimulation study, MetodoloskiZvzki, Vol-1, pp, 143-161.

[4] Rausch, j.R.,Kelley, k. 2009. A Comparison of linear and mixture models for discriminant analysis under Non normality, Behavior research methods, Vol.41, pp. 85- 98.

[5] Abdul Hussein, S. F., 2019, The use of mahalnobis statistic in the linear discriminant analysis between two groups, Al –

تم استلام الورقة بتاريخ:5/ 6/2024م وتم نشرها على الموقع بتاريخ: 2/ 7/2024م

Mustans-yriah university, the journal of administration and economics, vol-119, pp. 59-66.

[6] Mohsen, I.H., Maaroof, R.J., Mohsen, A.H., 2023, "Renal Failure, Types, Causes and Etiology: A Review Article", Vol 03, pp, 1663-1666.

[7] شمس الدين احمد علي احمد2013. "استخدام الدالة الخطية لدراسة مستوي الإصابة بسرطان الغدة الدرقية "رسالة ماجستير، جامعة الجزيرة، كلية العلوم الرياضية والحاسوب – السودان.