

Curvas RECEIVER-Operating Characteristic y Matrices de Confusión en la Elaboración de Escalas Diagnósticas

Vielka González¹, Milagros Alegret², Julio Betancourt³

¹Especialista de Bioestadística, Cardiocentro "Ernesto Che Guevara" de Villa Clara (Cuba); ²Profesor Auxiliar CPHE-VC (Cuba); ³Profesor Titular, Hospital Militar Manuel Fajardo (Cuba).

Resumen / Abstract

Introducción. *Las escalas pronósticas o diagnósticas también conocidas como scores de riesgo, son ejemplos típicos en Medicina de criterios de clasificación conjunta, su obtención en muchos casos requiere del empleo de clasificadores multivariantes. Ante una serie de clasificadores, debemos decidir cuál de ellos se desempeña mejor, en este sentido las matrices de confusión y las curvas ROC, juegan un papel preponderante.*

Métodos. *Se utilizó como caso de estudio la búsqueda de un índice predictivo para relaparotomías (IPR). Se comenzó analizando múltiples predictores con el objetivo de evaluar su capacidad de predicción sobre la respuesta binaria: hallazgo positivo en relaparotomía (Si/No), con este fin se emplearon los siguientes métodos de clasificación: análisis discriminante, árboles de decisión y regresión logística. Los resultados, fueron cuestionados a través de los parámetros ofrecidos por las matrices de confusión y de áreas bajo la curva ROC.*

Resultados. *Para la regresión logística el área bajo la curva, fue de 0,999 y para el árbol de decisión y análisis discriminante, esta fue de 0,997; la precisión, la exactitud y el CP+ para el primer clasificador fueron, 98,42%, 99,6% y 200 respectivamente, mientras que para los dos restantes los resultados fueron, 98,03%, 99,5% y 143.*

Conclusiones. *La regresión logística fue la que mejor se desempeñó ante el índice predictivo de relaparotomías. Se implementó una metodología simple y adecuada para ponerla en manos de decisores que trasciende a todos los estudios que requieran el uso de criterios de positividad conjunta para la estimación de riesgo.*

Introduction. *Prognostic and diagnostic scales also called risk scores are typical examples in medicine of joint classification criteria, to obtain in many cases requires the use of multivariate classifiers. Faced with a series of classifiers, we must decide which one performs better, in this sense the confusion matrices and ROC curves play a role.*

Methods. *Used as a case study the search for a predictive index for relaparotomies (IPR). It began its analysis*

with multiple predictors in order to assess their predictive power on the binary response: positive finding in re-laparotomy (Yes / No), to this end we used the following classification methods: discriminant analysis, decision trees and logistic regression. The results were questioned over the parameters offered by the confusion matrices and ROC curve areas.

Logistic. Regression results for the area under the curve was 0.999 and for the decision tree and discriminant analysis, this was of 0.997, the precision, accuracy and CP + for the first classifier were 98.42%, 99, 6% and 200 respectively, while for the other two results were 98.03%, 99.5% and 143.

Conclusions. Logistic regression was the best played with the relaparotomies predictive index. We implemented a simple and appropriate to put in the hands of decision makers that transcends all studies involving the use of joint positivity criteria for estimating risk.

Introducción

Los sistemas predictivos se conforman a partir de los índices de gravedad (IG) o indicandos, que se caracterizan por síntomas, signos clínicos, variables fisiológicas, valores de exámenes de laboratorio, signos imagenológicos y estados de comorbilidad (1). El resultado de la aplicación de un índice predictivo es, en su definición más simple, un mecanismo conducente a un sistema de clasificación que determina una conducta consecuente frente al paciente; la validez de los mismos, es expresada precisamente por la exactitud con la que estos operen.

La clasificación y predicción de enfermedades basadas en marcadores medidos, han sido por mucho tiempo objeto de investigaciones médicas y bioestadísticas. En muchos estudios hay múltiples marcadores que pueden estar asociados con una respuesta clínica binaria. Pero un solo marcador no puede ser usado para la clasificación con niveles satisfactorios de sensibilidad y especificidad. Por lo tanto, es importante combinar múltiples marcadores para alcanzar altos valores de sensibilidad y especificada (2). Esto nos lleva a un análisis multivariante del problema.

Para clasificar a pacientes en grupos determinados según una serie de características, se puede cuantificar el grado de incertidumbre de pertenecer a uno u

otro grupo a través de clasificadores multivariantes.

Existen varios métodos multivariantes que permiten obtener clasificadores, por lo que pudiera surgir la interrogante: ¿Cómo elegir el mejor de estos métodos existiendo varios?

En este sentido, las curvas ROC y los parámetros de las matrices de clasificación obtenidas para cada clasificador, juegan un papel preponderante.

Las matrices de confusión contienen información acerca de los valores reales y las clasificaciones predichas hechas por cualquier sistema de clasificación. El desempeño de un tal sistema es usualmente evaluado usando los datos en dicha matriz (3).

Los gráficos ROC son útiles para organizar clasificadores y visualizar su desempeño (4).

Actualmente el bioestadístico juega un papel fundamental en la toma de decisiones médicas, teniendo en cuenta el uso creciente de criterios de clasificación conjunta, el desarrollo de técnicas multivariantes más sofisticadas y el crecimiento de medios diagnósticos que ayudan todos, con su buen manejo a que estas decisiones médicas tengan la rapidez y calidad requeridas.


Por tales razones, el objetivo de este trabajo es emplear las curvas ROC y las matrices de confusión como complemento del análisis que se realiza cuando se tra-

baja con clasificadores, ya que constituyen una manera de examinar el desempeño de los mismos y por tanto contribuyen de manera importante al necesario mejoramiento de la práctica médica.

Materiales

Para el desarrollo de este trabajo se utilizó como caso de estudio la búsqueda de un índice predictivo para relaparotomías (IPR), que fue propuesto y defendido en el marco de una tesis doctoral (1).

Se contó con una muestra de 1000 pacientes elegidos aleatoriamente de un universo de 17 614, de estos, 751 no se relaparotomizaron por presentar una buena evolución post laparotomía, considerándose como "negativos", pues en ellos se

 Se utilizó como caso de estudio la búsqueda de un índice predictivo para relapararotomías

puede asegurar que los hallazgos post quirúrgicos hubieran sido negativos de haberse hecho una reintervención (proceder confirmatorio obviamente imposible) y

249 pacientes necesitaron de una relaparotomía urgente por haber tenido una evolución desfavorable postoperatoria.

Desde el punto de vista de la recolección de la información primaria, el caso de estudio se clasificó como una investigación descriptiva transversal. La variable dependiente o de respuesta, fue la variable dicotómica: hallazgo positivo en relaparotomía, los dos posibles resultados de la misma fueron: Sí o No. Las variables independientes o explicativas, fueron precisamente las que conformaron el índice predictivo (Síndrome de respuesta inflamatoria sistémica (SIRS), dolor abdominal difuso, nuevos síntomas pasadas las primeras 48 horas, hipoxemia e imagenología) las cuales también fueron medidas en escala dicotómica con valores: Sí o No.

Métodos

Se comenzó analizando múltiples predictores con el objetivo de evaluar su capacidad de predicción sobre la respuesta binaria: hallazgo positivo en relaparotomía (Si/No), con este fin se emplearon los siguientes métodos de clasificación: análisis discriminante, árboles de decisión y regresión logística. Cada uno de los métodos de clasificación se aplicó sobre una secuencia conjunta de casos en los que cada nuevo caso debió ser asignado a una de entre las dos clases predefinidas, basándose en las características observadas.

El análisis discriminante permite estudiar las diferencias entre dos (en el caso del análisis simple) o más (estaríamos ante el análisis discriminante múltiple) grupos de individuos definidos a priori, con respecto a varias variables simultáneamente (5). Esta constituyó nuestra primera alternativa de construcción de score integral de riesgo.

El objetivo de este método será discriminar, estimar o predecir la variable Y en función de los predictores X_1, \dots, X_p , mediante particiones sucesivas del conjunto de individuos, maximizando una medida de contenido de información respecto a la variable respuesta (6).

El principal inconveniente del análisis discriminante radica en que supone que los grupos pertenecen a poblaciones con distribución de probabilidad normal multivariante para las variables explicativas, con igual matriz de varianzas y covarianzas, por lo que en principio no debieran incluirse variables cualitativas; en este caso, el uso de las variables dicotómicas se justifica ya que se trabajó con códigos (0 para los individuos que no poseían la característica y 1 en caso contrario), además, debido al gran volumen muestral, se puede considerar confiable la aproximación de la distribución binomial a la normal por el teorema central del límite.

Existen procedimientos que aportan una serie de re-

glas o condiciones para predecir o clasificar los casos, estas son de muy fácil interpretación y se les denominan: árboles de decisión, constituyendo la segunda propuesta de construcción de un score de riesgo.

Se empleó el algoritmo CHAID para la segmentación o clasificación de los sujetos de la muestra.

CHAID construirá árboles no binarios (es decir, árboles donde más de dos ramas pueden estar unidas a una sola raíz o nodo), basado en un relativamente simple algoritmo que es particularmente más adecuado para el análisis de grandes bases de datos (6).

Como última técnica estadística que proponemos para la obtención de un score integral de riesgo, tiene un carácter probabilístico más refinado y se basa en una regresión logística (8).

En el análisis discriminante y la regresión logística se seleccionó el método de introducir todas las variables independientes juntas.

Los resultados de cada una de estas propuestas de score integral de riesgo fueron cuestionados a través de los parámetros ofrecidos por las matrices de confusión y de curvas ROC bajo el criterio de establecer diferencias del área bajo la curva.

En la tabla 1, se muestra una matriz de confusión y los parámetros que pueden ser calculados a partir de ella.

Los números a lo largo de la diagonal principal representan las decisiones echas de manera correcta, los números fuera de la misma representan los errores —la confusión—, entre las clases (4).

En el espacio ROC, el verdadero positivo es representado en el eje de las x, y el falso positivo es representado en el eje de las y (9). Mientras las matrices

ofrecen solo un par de sensibilidad y 1-especificidad, las curvas ROC, nos pueden brindar todos los pares según los posibles valores de corte del clasificador cuando este no es discreto, el resultado de representarlos en un eje de coordenadas es la curva ROC (Gráfico 1).

El área que queda bajo la misma es un buen indicador de la precisión de un clasificador.

Esta área es siempre mayor o igual a 0,5. El rango de valores se mueve entre 1 (discriminación perfecta) y 0,5 (no hay diferencias en la distribución de los valores de

la prueba entre los 2 grupos) (10).

El análisis ROC investiga la habilidad de un modelo para separar casos positivos de los negativos (como por ejemplo, predecir la presencia o ausencia de enfermedad), y los resultados son independientes de la prevalencia de casos positivos en la población estudiada (11).

Resultados

El gráfico 1 muestra las áreas bajo la curva de cada clasificador, notar que están prácticamente superpuestas, siendo difícil el análisis.

Véase la tabla 2, donde aparecen calculadas las áreas bajo la curva y las diferencias son mínimas, la regresión logística, con un área de 0,999, superior al árbol de decisión y al análisis discriminante, ambos con áreas de 0,997. El intervalo de confianza al 95% y el error típico para la regresión fue de 0,998 a 1 y de 0,001 respectivamente, para los dos métodos restantes, de 0,993 a 1 y 0,002.

Teniendo en cuenta que ninguno de los intervalos de confianza calculados para cada clasificador contiene el valor 0,5, se puede decir que existen evidencias estadísticas de que ellos pueden distinguir entre los dos



En el análisis discriminante desarrollado, se seleccionó el método de introducir todas las variables independientes juntas



grupos de la variable respuesta en estudio.

En la tabla 3, se observan los parámetros calculados a partir de las matrices de confusión; la regresión logística presentó mayor precisión que el resto de los métodos analizados (98,42%), ya que fue la que menor porcentaje de falsos positivos arrojó (0,5), para el árbol de decisión y el análisis discriminante este último fue de 0,7, por tanto la precisión fue de 98,03% en ambos casos.

Semejante análisis se realiza con la exactitud, esta fue de 99,6% para la regresión logística debido a que fue la que presentó mayor porcentaje de verdaderos negativos (Sensibilidad) (99,5), para el árbol y el discriminante, este fue de 99,3, siendo la exactitud de 99,5%.

Estas diferencias se ilustran mejor al calcular el cociente de probabilidades positivo (CP+), como indica la lógica, el mayor valor y por tanto el de mejor capacidad para diagnosticar o predecir en este caso la presencia de hallazgos positivos en la relaparotomía es el score confeccionado según la regresión logística, su valor indica que si el paciente es positivo en la relaparotomía es 200 veces más probable según este clasificador, que sea clasificado como tal. Sin embargo el valor de este parámetro para los otros dos clasificadores es de 143.

Discusión



La combinación de clasificadores es una técnica establecida para mejorar el desempeño de una clasificación. Las posibles reglas de combinación propuestas hasta ahora generalmente tratan de disminuir la tasa de errores en la clasificación, lo cual es una medida de desempeño poco apropiada en muchas situaciones reales y particularmente cuando se trata de problemas que presentan dos clases. En este caso, una buena alternativa está dada por el área bajo la curva ROC, cuya

efectividad en medir la calidad de la clasificación ha sido probada en muchas y recientes publicaciones (12-13) y que es además útil para medir la calidad de la jerarquización de un clasificador como es requerido en muchas de las demandas reales (13).

Por último, es posible que las comparaciones empíricas realizadas no demuestren de forma tangible en este caso la supremacía de un método sobre el resto, sino que establecen un tipo de problema para el que un método se comporta mejor que los otros. Se deja abierta la elección del mejor clasificador ante cada nuevo caso, teniendo en cuenta la preparación del investigador, los recursos computacionales y tiempo disponibles, aunque se aconseja elegir el que mejor se comporte para el problema en particular; si se obtu-

viera una discreta mejora por métodos más complejos como los es la regresión logística, debería aplicarse un criterio de parsimonia y elegir un método que a la vez sea sencillo y eficaz. Se ofrece de esta forma una metodología

simple y adecuada para ponerla en manos de decisores, ejemplificada en este caso de estudio, pero que es trascendente a todos los estudios que requieran el uso de criterios de positividad conjunta o clasificadores multivariantes para la estimación de riesgo, categorización y consecuente toma de decisiones sobre sujetos evaluados en diferentes circunstancias médicas.

 Se ofrece esta metodología simple y adecuada para ponerla en manos de decisores, pero que es trascendente a otros estudios similares 

Bibliografía

1. Betancourt Cervantes Julio. Propuesta de un índice predictivo para relaparotomía.[tesis doctoral]. Villa Clara: Universidad médica de Villa Clara; 2008.
2. Ma Shuangge, Huang Jiang. Combining Multiple Markers for Classification Using ROC. *Biometrics*. 2007 Septiembre;63:751-7.
3. Weiss, G. M., Provost, F. Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction. *JAIR*. 2003;19: 315-54.
4. Fawcett Tom. An introduction to ROC analysis. *Pattern Recognition Letters*. 2006, June;27(8):861-74.
5. Mures Quintana M. Jesús, García Gallego Ana, Vallejo Pascual M. Eva. Aplicación del Análisis Discriminante y Regresión Logística en el estudio de la morosidad en las entidades financieras. Comparación de resultados. *Pecvnia*. 2005;1:175-199.
6. Schiattino Lemus Irene, Silva Zamora Claudio. Árboles de Clasificación y Regresión: Modelos Cart. *Rev C & T*. 2008;10(30):161-6.
7. T Hill, P Lewicki. *STATISTICS Methods and Applications*. Tulsa, OK: StatSoft; 2007.
8. Aguilera del Pino A. M. Tablas de contingencia bidimensionales. España: La muralla; 2001.
9. Provost Foster, Fawcett Tom. Robust Classification for Imprecise Environments. *Machine Learning*. 2001;42:203:31.
10. Alonso Domínguez Emma, González Suárez Roberto. Análisis de las curvas Receiver – Operating Characteristic: Un método útil para evaluar procedimientos diagnósticos. *Rev Cubana Endocrinol*. 2002;13(2):169-76.
11. A. Lasko Thomas, G. Bhagwat Jui, H. Zou Kelly, Ohno-Machado Lucila. The use of receiver operating characteristic curves in biomedical informatics. *Journal of biomedical informatics*. 2005, October;38(5):404-15.
12. Marrocco Claudio, Molinara Mario, Tortorella Francesco. Exploiting AUC for optimal linear combinations of dichotomizers. *Pattern recognition*. 2006 Jun;27(8):900-7.
13. Marrocco Claudio, Duin R.P.W., Tortorella Francesco. Maximizing the area under the ROC curve by pairwise feature combination. *Pattern recognition*. 2008 Jun;41(6):1961-74.

Anexo 1

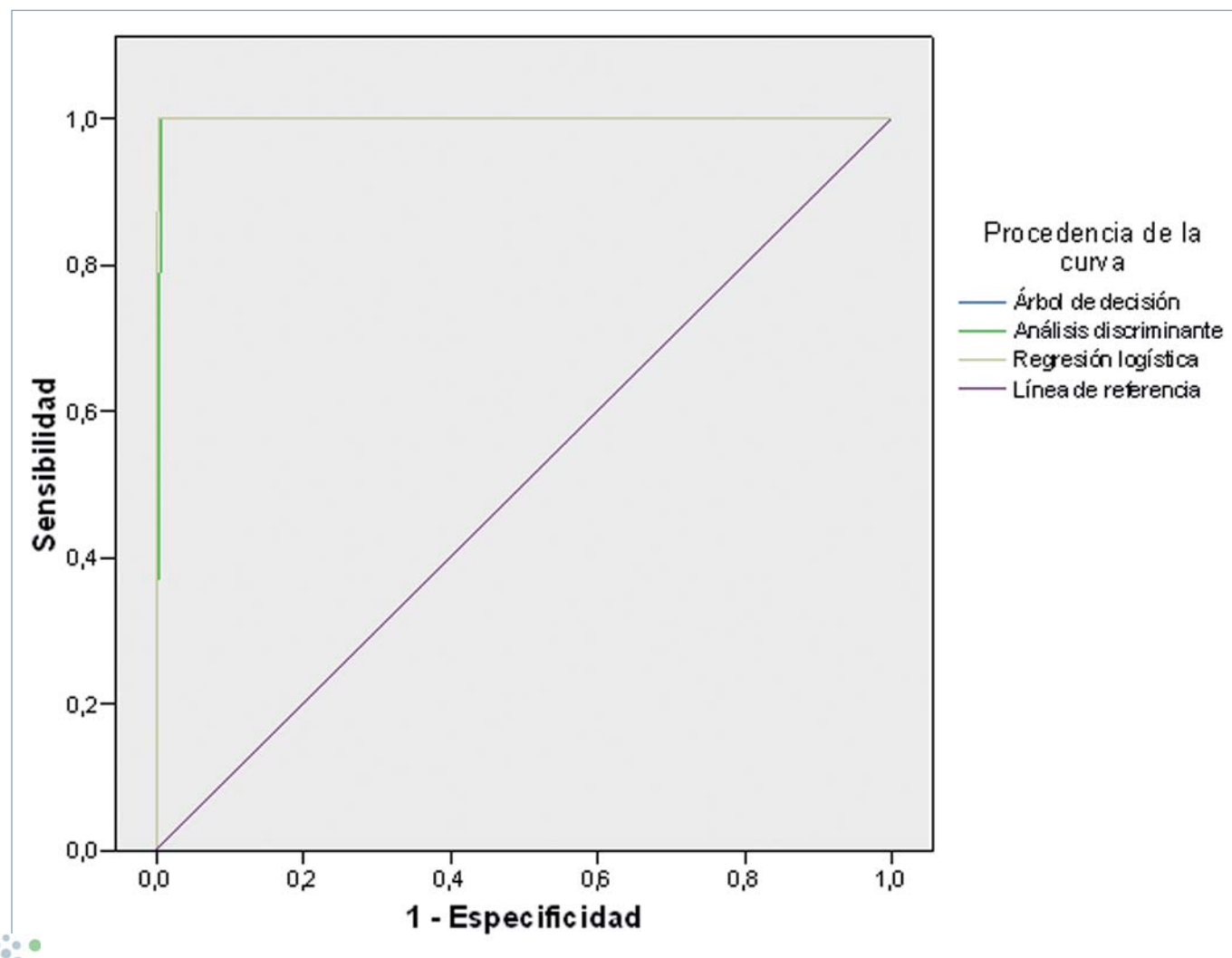


Gráfico 1. Curvas ROC.

Anexo 2

Tabla 1. Matriz de confusión y formulas de parámetros calculados a partir de ella.

| Clase Real | Clase pronosticada por el clasificador | | Total |
|------------|--|------------------------------------|---------------------------------|
| | Sí | No | |
| Sí | Verdaderos positivos (VP) | Falsos negativos (FN) | Total de positivos reales (TPR) |
| No | Falsos positivos (FP) | Verdaderos negativos (VN) | Total de negativos reales (TNR) |
| Total | Total de positivos predichos (TPP) | Total de negativos predichos (TNP) | Total general |

$$\text{Sensibilidad} = VP/TPR$$

$$1\text{-Especificidad} = FP/TNR$$

$$\text{Precisión} = VP/TPP$$

$$\text{Exactitud} = (VP+VN)/\text{Total general}$$

$$CP+ = \text{Sensibilidad}/1\text{-Especificidad}$$

Anexo 3

Tabla 2. Parámetros de las matrices de confusión.

| Métodos | Parámetros de las matrices de confusión | | | | | | |
|------------------------|---|----------------|--------------|--------------|----------------|-----|---------------|
| | Exactitud (%) | FN (1-Sen) (%) | VP (Sen) (%) | VN (Esp) (%) | FP (1-Esp) (%) | CP+ | Precisión (%) |
| Regresión logística | 99,6 | 0 | 100 | 99,5 | 0,5 | 200 | 98,42 |
| Árbol de decisión | 99,5 | 0 | 100 | 99,3 | 0,7 | 143 | 98,03 |
| Análisis discriminante | 99,5 | 0 | 100 | 99,3 | 0,7 | 143 | 98,03 |

Anexo 4

Tabla 3. Áreas bajo la curva ROC.

| Estadísticos | Área bajo la curva | Error típico (a) | Sig. Asintótica (b) | Intervalo de confianza al 95% | |
|------------------------|--------------------|------------------|---------------------|-------------------------------|----------|
| | | | | Inferior | Superior |
| Regresión logística | ,999 | ,001 | ,000 | ,998 | 1 |
| Árbol de decisión | ,997 | ,002 | ,000 | ,993 | 1 |
| Análisis discriminante | ,997 | ,002 | ,000 | ,993 | 1 |