Clasificación de tumores mamarios utilizando regresión logística y KNN

Randall Chaves[1], Brandon Guido[1] y Valentín Chavarría[1]

randall.chavesleon@ucr.ac.cr, brandon.guido@ucr.ac.cr, valentin.chavarria@ucr.ac.cr

RESUMEN

Una parte vital para la sobrevivencia de personas con tumores mamarios es el diagnóstico temprano. Parte de los esfuerzos en facilitar este proceso ha sido recurrir a conocimiento estadístico. Por esto se implementó modelos estadísticos para la clasificación de muestras de tejido de tumores mamarios y a partir de esto se elaboró un análisis comparativo entre los modelos. Se utilizaron datos de biopsias por aguja fina (FNA), donde cada observación contiene variables sobre la topología de las células; estas fueron realizadas a pacientes con tumores mamarios en el Centro de Ciencias Clínicas de la Universidad de Wisconsin entre 1984 y 1985. Para la clasificación se implementa KNN y modelos selección de variables sobre regresión logística. Los utilizados son la selección de subconjuntos por algoritmos *backward*, *forward* y la regresión Lasso. Con los modelos se obtuvieron niveles de exactitud superiores al 90%. Además, se realizó un análisis detallado de los niveles de sensibilidad para cada uno de los modelos, se logró reducir la posibilidad de falsos negativos. En promedio se obtiene un nivel de sensibilidad del 97,24% y se alcanza un máximo de exactitud del 98,59%. Aunado a esto, se obtuvo una reducción de las variables utilizadas que permite acceder a valores de significancia adecuados. A partir de estos resultados se obtuvo que el método más efectivo, para la clasificación de las muestras de tejido, fue la regresión logística posterior a la ejecución de una selección de variables. Debido a esto se afirma, que la implementación de métodos estadísticos puede ser relevante en el diagnostico de tumores mamarios. se tiene que la biopsia por FNA puede brindar información útil para el profesional de salud.

PALABRAS CLAVE: cáncer, sensibilidad, curva ROC, diagnóstico.

INTRODUCCIÓN

El cáncer de mama tiene una de las mayores incidencias a nivel mundial. Con el objetivo de reducir las consecuencias negativas por este padecimiento se ha tratado de encontrar maneras efectivas para detectar de forma temprana su presencia, asimismo como para su tratamiento. Cuando existe la presencia de una masa anormal en un seno mamario, por lo general se recurre a una biopsia. Existen diversos métodos para realizar una biopsia, una de esas técnicas es la de aguja fina (FNA). En su trabajo, Willems y col., 2011, explica que dentro de los puntos a favor de la FNA se incluye el menor impacto de la prueba para el paciente. Asimismo, se tiene que es económica en comparación a otros métodos. No obstante, debido a la gran variedad de tumores que pueden presentarse en el tejido mamario, en general, el nivel de confianza que se le puede tener al diagnóstico obtenido queda en desventaja respecto a las demás. Asimismo, Teague y col., 1997,

indican que en ese escenario debe recurrirse a otros procedimientos para poder descartar que sea una lesión maligna, lo que aumenta sus costos.

Por ello, surge la idea de utilizar inferencia estadística para sacar mayor provecho a la información que ya se recopila a través de la biopsia por aguja fina. Justamente este tipo de estrategia es la que se plantea abordar en este trabajo, para sacar más provecho de la información que se obtiene con FNA. Esto se ha realizado de manera previa en otros estudios como el de Alfaro, Arroyo & Hernández (2020), donde se comparó la capacidad de predicción de distintos métodos estadísticos de aprendizaje supervisado para la misma base de datos.

El aprendizaje supervisado es una técnica de inferencia estadística. Esta consiste en transformar entradas específicas en una salida con base en el conjunto de datos de entrenamiento (Paez, 2019). Es decir, lo que hace es observar el comportamiento de un porcentaje de la base de datos de las entradas-salidas como guías y aprende una función que asigna una salida según la entrada que se utilice. (Russell & Norving, 2010, como se citó en Paez, 2019). El otro porcentaje de las observaciones que no se utilizaron en la construcción del modelo se utilizan para hacer las pruebas de este y realizar diagnósticos sobre el modelo calibrado.

Para obtener un modelo de clasificación primero se realiza un análisis exploratorio de la base de datos. Posterior a esto se utiliza el método de regresión logística y el método de KNN. Con esto se espera obtener modelos eficientes que permitan dar diagnósticos confiables con la información de una muestra citológica de una biopsia por FNA.

METODOLOGÍA

Se utiliza una base de datos que contiene un total de 32 variables distintas, con 569 observaciones. Dichas variables describen la topología de las células extraídas a través de la técnica de biopsia por aguja fina en tumores mamarios. Esta base se recopiló en el lapso comprendido entre 1984 y 1985 en el Centro de Ciencias Clínicas de la Universidad de Wisconsin (Wolberg y col., 1995). Las variables corresponden a:

1. Número de identificación (id): Valores numéricos que identifican cada biopsia registrada.

2. Diagnóstico (diagnosis): Determina si se observa presencia de células malignas o no. Es de tipo categórica, con valores "M" para maligno y "B" para benigno. Esta es la variable respuesta.

3. Radio (radius): Es la media de las distancias desde el centro del núcleo celular hasta los puntos del perímetro.
4. Textura (texture): Desviación estándar de los valores de la escala de grises.

5. Perímetro (perimeter): Perímetro de la imagen digitalizada del núcleo celular.

6. Área (area): Área de la imagen digitalizada del núcleo.
7. Suavidad (smoothness): Variación local en las longitudes de los radios.
8. Compacidad (compactness): Corresponde al cálculo de $\frac{perimetro^2}{area} - 1$
9. Concavidad (concavity): Severidad de las porciones cóncavas del contorno.
10. Puntos de concavidad (concave points): Número de porciones cóncavas del contorno.
11. Simetría (symmetry): Indica el grado de simetría.
12. Dimensión fractal (fractal_dimension): Es la aproximación de un límite.

Cada muestra tomada contenía una gran cantidad de ejemplares de célula, por lo que la información con la que se trabaja viene a resumir el comportamiento general del tejido analizado. Para cada uno de estos aspectos, se tiene tres estadísticos distintos. El primero de ellos es la media (mean), la cual representa la media simple empírica para la variable en cuestión. El segundo, error estándar (se), en representación de la desviación simple con las observaciones del tejido. El ultimo, mayor (worst), este se calcula mediante la media simple de los tres valores de mayor magnitud observados en la muestra para la variable en cuestión.

En la presente investigación se desarrollaron modelos como el de **regresión logística**, es un método de análisis multivariado. Su empleo más útil es cuando se tiene una variable dependiente dicotómica y un conjunto de variables predictoras (cuantitativas o categóricas), y así poder estimar la probabilidad de que cierto evento ocurra teniendo en cuenta las otras variables.

Para modelar esto se utiliza una función *g*(*X*) que nos da resultados en el intervalo de 0 a 1, donde 0 representa que es 0 por ciento probable que ocurra el evento en estudio y 1 representa 100 por ciento probable de que ocurra. La función logística es:

$$g(X) = \frac{e^{\beta X}}{1 + e^{\beta X}}$$

donde $\beta = (\beta_0, \dots, \beta_p)$ y *X* el conjunto de variables predictoras, es decir, $(1, X_1, \dots, X_p)$,luego se puede verificar que

$$\frac{g(X)}{1 - g(X)} = e^{\beta X}$$

A esta expresión se le llama *odds* o chances, y como su nombre en español lo indica, hace referencia a la posibilidad que hay de que ocurra el evento con respecto a que no ocurra (Gareth y col., 2013). Para estimar los coeficientes β, donde estos deben ser estimados utilizando la base de entrenamiento disponible, se utiliza el método de máxima verosimilitud. La función de verosimilitud es:

$$l(\beta) = \prod_{i=1}^{p} p(X_i)^{Y_i} \cdot \left(1 - p(X_i)\right)^{1-Y_i} = X^T\left(Y - p(X)\right)$$

Dentro de las hipótesis necesarias está la independencia de las variables. Por ende, cuando hay presencia de variables con alto nivel de correlación, uno de los posibles enfoques a tomar es reducir su cantidad. Para hacer esto hay diversos métodos, uno es la selección de subconjuntos. La idea surge al comparar distintos grupos de variables y su calibración asociada (Gareth y col., 2013). A través de medidas de error como la devianza se permite este contraste, para seleccionar el que minimice el error presente.

Hay diversos algoritmos para recorrer la cantidad posible de resultados, pues el costo computacional de comparar todos los posibles subconjuntos es alto dependiendo de la cantidad de variables que se tiene al inicio. El método hacia adelante (*forward*) parte de comparar inicialmente las variables por separado, posteriormente selecciona y aumenta de manera gradual el número de variables a considerar. En contraste, el método hacia atrás (*backwards*) compara los modelos obtenidos al eliminar una de las variables como inicio para después reducir progresivamente la dimensión.

Además, se implementa **Lasso** para regresión logística, donde la idea es utilizar las mismas expresiones que se utilizaron en regresión logística, pero a la función de máxima verosimilitud log(*l*(β)) se le agrega una restricción sobre los coeficientes beta. Entonces la expresión a maximizar con la restricción queda de la siguiente manera:

$$\sum_{i=1}^{n}\left\{y_i(\beta X_i) - \log\left(1 - e^{\beta X_i}\right)\right\} + \lambda \sum_{j=1}^{p}\left|\beta_j\right|$$

Dada esta restricción, dependiendo del valor de algunos de los coeficientes se pueden anular (Wang y col., 2015). Es por eso que la cantidad de variables seleccionadas por Lasso dependen del parámetro.

De manera alternativa también se plantea el uso del **método KNN**. En este se tiene que dado un entero positivo *K* y una observación X_0 , se va a buscar los K puntos del conjunto de entrenamiento que estén más cerca de X_0 , este conjunto se representa con N_0. Después de esto se calcula una probabilidad condicional de que el punto corresponda a una clase *j* con la fórmula:

$$P(Y = j | X = x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = j)$$

Y gracias a esto se clasifica dependiendo de cuál sea la clase de mayor probabilidad para la observación X_0 (Gareth y col., 2013). Por lo general, el enfoque para su programación se basa en la selección del número K que garantiza mejores niveles de predicción o clasificación.

Por último, se realizó un análisis comparativo de los resultados esto mediante **Matrices de confusión** y **Curvas ROC** (*Receiver Operating Characteristics*). Como en este caso la variable de estudio tiene 2 valores y los resultados pueden estar correctos o incorrectos, se crea una matriz de

confusión. Gracias a esto se puede calcular sensibilidad (tasa de verdadero positivo), especificidad (tasa de verdadero negativo) y exactitud (tasa de valores bien clasificados) (James y col., 2013). Con ello se calcula la curva ROC en la que se grafica la sensibilidad en el eje de las ordenadas, y 1-especificidad en el eje de las abscisas para varias matrices de confusión.

Para esta investigación y para la implementación de las funciones y métodos anteriormente mencionados se utilizó R (R Core Team, 2013) y los paquetes de este, como ggplot2 (Wickham, 2016), tidyverse (Wickham y col., 2019), ROCR (Sing y col., 2005), leaps (based on Fortran code by Alan Miller, 2020) y glmnet (Friedman y col., 2010).

RESULTADOS

Al revisar el comportamiento de la base de datos se encuentra que una gran parte de las variables tienen altos niveles de correlación. Este fenómeno puede afectar la rigurosidad del modelo, pues influye en supuestos de independencia de variables. Además, al analizar la distribución empírica se encuentra que la mayor parte de las observaciones están concentradas en los primeros cuantiles, es decir, no hay simetría. Por otra parte, se detecta que las escalas difieren bastante entre variables. Esto puede provocar que al momento de implementar los modelos se obtenga resultados no óptimos por el efecto de la diferencia de magnitudes. En este punto destaca el caso de KNN, donde la magnitud de las variables va a tener un efecto directo sobre la distancia sobre el espacio que conforma la base de datos.

Para el ajuste de los modelos se designa como base de entrenamiento un 75% aleatorio de la original. El 25% restante se deja como base de prueba. Es importante recalcar que debido a la naturaleza de este trabajo, reducir la presencia de falsos negativos es primordial. Por esta razón se busca obtener el mejorar dentro de lo posible el rubro de sensibilidad que se vaya obteniendo. Se ajustaron 6 distintos modelos de los cuales 5 fueron generados con regresión logística bajos distintos métodos de selección de variables y uno con KNN.

En primera instancia, se desarrolló el método KNN. Al comparar se observan mejores resultados al utilizar variables escaladas. Además, se obtiene un número de vecinos óptimo de 6, en la Tabla 1 se aprecia que con este se obtiene una sensibilidad del 95,83%.

Tabla 1

Comparación de indicadores según el modelo seleccionado.

Modelo	Estadístico		
	Sensibilidad	Especificidad	Exactitud
Regresión logística [1]	0.9783	0.9792	0.9789
Regresión logística [2]	0.9783	0.9895	0.9859
Regresión logística [3]	0.9783	0.9479	0.9577
Regresión logística [4]	1.0000	0.9583	0.9718
Regresión logística [5]	0.9411	0.9940	0.9719
KNN	0.9583	1.0000	0.9718

1 *No se descartan variables*
2 Variables obtenidas de SS *forward*
3 Variables obtenidas de SS *backward*
4 Variables obtenidas por análisis de correlaciones
5 Variables obtenidas por Lasso

En el caso de la regresión logística se plantean varios escenarios distintos. En primer lugar, se utiliza la totalidad de variables dentro de la base de datos para calibrar. Se obtiene un nivel de sensibilidad alto, además de que se mejora el rubro de especificidad respecto al caso anterior. No obstante, se observa que los valores *p* asociados a los coeficientes obtenidos no pueden ser considerados significantes, es decir, no se rechaza la hipótesis nula de que estos sean 0. Por esta razón se decide explorar alternativas para la selección de variables.

En el segundo escenario se implementa la selección de subconjuntos con el algoritmo *forward*. En este punto se obtiene una reducción a un conjunto de 15 variables. Se puede ver en la tabla 1 como esta segunda regresión representa una mejora en poder de predicción, pues mantiene su valor se sensibilidad, pero aumenta la especificidad y exactitud.

Como tercer modelo se usa la selección con algoritmo *backward*. Esto se realizó con el objetivo de encontrar una diferencia importante respecto al caso anterior. En este escenario se delimita a 13 variables. Sin embargo, se da un deterioro para el nivel de especificidad y exactitud. Por su parte, el estadístico de sensibilidad no presenta cambios. Además, se sigue presentando un bajo valor de significancia para ambos modelos.

Dado que los resultados obtenidos previamente tienen problemas con el rechazo de la hipótesis nula sobre sus parámetros, se decide tomar en cuenta la correlación presente en la base

de datos. Un primer hallazgo fue que las variables asociadas a los peores valores (*worst*) tienen niveles de correlación con sus variables homólogas asociadas a la media. Por esta razón se decide omitirlas para el modelo.

Asimismo, se destaca el caso de las variables de radio, perímetro y área. Sus correlaciones son sumamente altas. Este resultado es esperable de manera intuitiva, pues en un objeto esférico la medida del radio guarda relación con el perímetro y área de la célula. De esta manera, se plantea el hacer la regresión con solo una de esas variables, área.

Ahora bien, antes de realizar estas simulaciones, se detectó que la variable de media de puntos de concavidad también guardaba bastante relación con las otras, no obstante, en un menor grado. Al realizar una regresión lineal sobre la variable del promedio de puntos de concavidad en términos de las otras, se encuentra que esta puede ser explicada por otras. Por esta razón también se elimina.

A partir del conjunto de variables obtenido, se decide utilizar el algoritmo de selección hacia adelante (*forward*). De esta manera se obtiene un conjunto de 13 variables. Con esta regresión se obtiene un nivel de sensibilidad del 100%, sin embargo, hay una reducción en niveles de especificidad y exactitud respecto a la de mayor sensibilidad previa (regresión logística 1). Por otra parte, los niveles de significancia para este modelo tienen una mejora sustancial que permite rechazar su nulidad.

Como un último acercamiento para la selección de variables y búsqueda de un mejor modelo, se decide utilizar el método Lasso. Este se implementa sobre el conjunto total de las variables y se seleccionan 12 de estas. Con este modelo se alcanza una especificidad más alta que todos los modelos anteriores, pero se pierde el nivel de sensibilidad previo.

De manera alternativa para la comparación de los distintos ajustes realizados bajo regresión logística, se tiene la tabla 2. En este se puede observar que el modelo obtenido con la eliminación previa de variables correlacionadas es el que tiene un mayor porcentaje de área bajo la curva ROC. Aunado a esto, esta área es bastante cercana a la unidad, lo que le brinda mayor validez a ese modelo. Ahora bien, cabe resaltar que los otros modelos de regresión logística también poseen un área bastante considerable.

Tabla 2

Comparación de modelos de regresión logística
(Con la misma numeración de modelos en Tabla 1)

Estadístico	Regresión logística				
	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5
AUC	97.85	99.75	99.32	99.84	96.75

CONCLUSIONES

El cáncer mamario es una amenaza muy presente en la actualidad. Debido a que el diagnóstico temprano es vital para mejorar las expectativas de sobrevivencia, se ha buscado mejorar los métodos de diagnóstico. La técnica de diagnóstico FNA es económica, pero se deja de lado por su poca capacidad predictiva. Este problema ha sido afrontado en otros proyectos a través de métodos estadísticos de clasificación, como la regresión logística y KNN. Para la regresión se decide realizar una selección de variables, pues se detecta un alto grado de correlación. A partir de eso se llega a un modelo óptimo donde se obtiene un nivel de sensibilidad ideal para la pregunta de investigación. Este modelo no es el que alcanza la mayor exactitud dentro de los calibrados, no obstante, es importante resaltar que en el escenario de diagnóstico para tumores, se le da prioridad a disminuir la presencia de falsos negativos. Además, este modelo tiene el porcentaje de área bajo la curva mayor dentro de las 5 regresiones calibradas, lo que le otorga solidez a su uso.

Otro de los resultados obtenidos es relacionado al nivel de significancia de las variables. Este fue uno de los problemas encontrados para las primeras regresiones realizadas, pues a pesar de que se mantenía niveles aceptables de exactitud, sensibilidad y especificidad, al no cumplir con valores adecuados de significancia, se le restaba certidumbre al modelo. Este problema fue resuelto de manera satisfactoria para el modelo óptimo mencionado.

También se implementó el método KNN. En este no se alcanza el nivel de especificidad del modelo óptimo previo, sin embargo, sigue teniendo un nivel bastante bueno de predicción, al igual que en el caso de las otras regresiones logísticas utilizadas. Con estos resultados se concluye que es posible utilizar teoría estadística para obtener más información a partir de la técnica FNA.

Es importante mencionar que en este proceso de selección de variables no se tuvo acceso a un criterio médico profesional. Si bien es cierto se llega a resultados predictivos adecuados, la parte interpretativa del modelo puede verse beneficiada por la inclusión de este tipo de información para futuros trabajos relacionados. Además, se plantea la limitante de que se utilizó solamente dos métodos de clasificación. Existen otros modelos que podrían beneficiar enormemente el desarrollo de programas médicos que faciliten el proceso de diagnóstico para tumores mamarios. Para fututos trabajos se deben tener en cuenta estas recomendaciones para mejorar el alcance de estos análisis.

REFERENCIAS

Alfaro, A., Arroyo, A. and Hernández, D. (2020) "Técnicas de clasificación en datos de cáncer de mama para la confirmación del dictamen médico de especialistas en el área, mediante biopsia por aspiración con aguja fina, Wisconsin breast cancer data set," *SERENGUETI*, 2(1), pp. 105–113. http://www.estadistica.ucr.ac.cr/index.php/es/actividades/revista-serengueti.

Basado en Fortran code by Alan Miller, T. L. (2020). *leaps: Regression Subset Selection* [R package version 3.1]. https://CRAN.R-project.org/package=leaps

Friedman, J., Hastie, T. & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*(1), 1-22. https://doi.org/10.18637/jss.v033.i01

Gareth, J., Daniela, W., Trevor, H. & Robert, T. (2013). *An introduction to statistical learning: with applications in R*. Spinger.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Paez, S. (2019). Analisis comparativo de herramientas open source para data mining sobre datos publicos del Ministerio de Educacion de la Republica del Ecuador.

R Core Team. (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. http://www.R-project.org/

Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics*, *21*(20), 7881. http://rocr.bioinf.mpi-sb.mpg.de

Teague, M. W., Wolberg, W. H., Street, W. N., Mangasarian, O. L., Lambremont, S. & Page, D. L. (1997). Indeterminate fine-needle aspiration of the breast. *Cancer*, *81*, 129-135. https://doi.org/10.1002/(sici)1097-0142(19970425)81:2<129::aid-cncr7>3.0.co;2-n

Wang, H., Xu, Q. & Zhou, L. (2015). Large unbalanced credit scoring using lasso-logistic regression ensemble. *PloS one*, *10*(2), e0117844.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Willems, S. M., van Deurzen, C. H. M. & van Diest, P. J. (2011). Diagnosis of breast lesions: fine-needle aspiration cytology or core needle biopsy? A review. *Journal of Clinical Pathology*, *65*, 287-292. https://doi.org/10.1136/jclinpath-2011-200410

Wolberg, W., Street, W. & Heisey, D. (1995). UCI Machine Learning Repository: Data Set. *archive.ics.uci.edu*.