Estudio comparativo de las técnicas de clasificación sobre el cáncer de mama

Exequiel Fuentes Lettura
Universidad Católica del Norte
Departamento en Ingeniería de Sistemas y Computación
Antofagasta, Chile
Email: efulet@gmail.com

***Resumen*—La detección precisa de células cancerígenas en un paciente es crítica y puede alterar el tratamiento posterior y aumentar las posibilidades de supervivencia. Técnicas de Machine Learning han sido fundamentales para la detección de enfermedades y en la actualidad se están utilizando en varios problemas de clasificación debido a su rendimiento preciso de predicción. Varias técnicas pueden proporcionar diferentes precisiones deseadas por lo que es imprescindible utilizar el método más adecuado que proporcione los mejores resultados deseados. En esta investigación se pretende ofrecer un análisis comparativo entre el modelo de Regresión Lineal y el modelo Rectangular sobre el problema de la clasificación del cáncer de mama de Wisconsin.**

Palabras clave: Machine learning, Regresion Lineal, Modelo Rectangular, Cáncer de Mama

I. Introducción

La utilización de enfoques de Machine Learning en dominios médicos ha aumentado rápida y consistentemente en el tiempo debido a la mejora en la eficacia de estos enfoques en la clasificación, predicción y especialmente en la ayuda en la toma de decisiones. Por otro lado, el aumento en el número de pacientes que requieren atención médica en el marco de evaluaciones periódicas ha llevado al desarrollo técnico de sistemas automatizados, reduciendo el costo de las prestaciones y ayudando en la mejora de los estudios clínicos.

El cambio del perfil epidemiológico de la población junto con el aumento de la esperanza de vida al nacer, ha provocado el aumento de las enfermedades crónicas como el cáncer, constituyendo un importante problema de salud pública en términos de morbi-mortalidad. Es así como, el cáncer ocupa el segundo lugar entre las principales causas de mortalidad en los últimos 30 años.

Las estimaciones a nivel mundial de mortalidad e incidencia por cáncer de mama obtenidas a través de la GLOBOCAN 2008, aparece como primera causa de muerte en la mujer. Esta misma fuente estimó que anualmente fallecen 458.367 mujeres por esta causa, alcanzando una tasa estandarizada por edad de 12,5 por 100.000 mujeres. La tasa de incidencia estandarizada por edad fue estimada en 39 por 100.000 mujeres con un total de 1.383.523 casos nuevos en el mundo [1].

La detección precoz a través de la mamografía incrementa las opciones de tratamiento y tiene un efecto demostrado en disminuir la mortalidad. Sin embargo, esta estrategia puede ser de alto costo en países menos desarrollados por lo que la mayoría de los esfuerzos se han hecho en la prevención primaria. Este tipo de cáncer ha venido en descenso en Norteamérica y Europa en los últimos 25 años como resultado de la detección precoz y mejoría en los tratamientos.

La identificación de las células cancerígenas en un paciente es altamente subjetiva y depende de la experiencia del médico. Esto puede conducir a predicciones inexactas puesto que los experimentos son propensos a error humano y visual, que además pueden verse afectados por una mamografía borrosa. Los desafíos mencionados requieren la necesidad de herramientas precisas para detección y clasificación de células cancerígenas. Técnicas de Machine Learning han sido fundamentales para proporcionar pruebas que apoyen la exactitud de la clasificación de los pacientes con cáncer de mama. Una vez que el diagnóstico de cáncer de mama se ha realizado, el pronóstico se determina posteriormente para predecir el desarrollo futuro y las características de las células cancerígenas [2].

En esta investigación se pretende ofrecer un análisis comparativo entre un modelo de Regresión Lineal y un modelo Rectangular, siendo este último desarrollado para esta investigación. Ambos modelos utilizarán un conjunto de datos públicos denominado “Wisconsin Diagnostic Breast Cancer”, el cual posee 569 instancias de 32 atributos incluyendo su atributo clase correspondiente al diagnóstico (maligno o benigno) [3].

Este artículo esta organizado de la siguiente manera: materiales y métodos utilizados, resultados obtenidos y conclusiones.

II. Desarrollo

II-A. Materiales y métodos utilizados

Los conjuntos de datos “Wisconsin Breast Cancer” localizados en el repositorio “UCI Machine Learning” [3] son utilizados para distinguir entre instancias que pueden ser malignas (cancerígena) o benignas (no cancerígena). Para efectos de esta investigación se ha seleccionado el conjunto

de datos denominado "Wisconsin Diagnostic Breast Cancer" (WDBC), el cual contiene 569 instancias correspondiente a las muestras tomadas a pacientes.

Los detalles de los atributos pueden ser encontrados en el archivo llamado "wdbc.names", los cuales son: "ID number", "Diagnosis" (M = maligno o B = benigno) y 30 valores reales correspondiente a la media, la desviación típica y el "peor" o más largo (la media de los tres valores más largos) por cada característica del núcleo de la célula. Este núcleo, que está presente en una imagen digitalizada, tiene las siguientes características: "Radius", "Texture", "Perimeter", "Area", "Smoothness", "Compactness", "Concavity", "Concave points", "Symmetry" y "Fractal dimension". Por ejemplo, la columna 3 es la media del radio, la columna 13 es la desviación típica del radio y la columna 23 es el "peor" valor del radio.

Para evitar que el atributo "ID number" tuviera alguna incidencia sobre la clasificación fue removido ya que no tiene un aporte real al problema. En el caso del atributo clase, "M" fue reemplazado por "1" y "B" por "0".

Como estrategia de validación se utilizó el método "holdout". Se dividió el conjunto de datos en un 70 % para efecto de entrenamiento y un 30 % para las pruebas de los modelos.

Formalmente, un conjunto de entrenamiento que contiene N instancias puede definirse como $X = \{x^t, r^t\}$, con $t = 1, ..., N$. Donde x es un atributo de entrada definido como $x = [x_1, x_2, ...x_t]$ y r es el rótulo del atributo el cual se define de manera similar a x, entonces $r = [r_1, r_2, ...r_t]$. Lo que se busca es una función G, tal que, $G(x)$ prediga un valor basado en X. Para este caso en particular, X es de dimensión 30.

Se pueden utilizar varios métodos para examinar la clasificación de este problema. En esta investigación se utilizarán el modelo de Regresión Lineal y el modelo Rectangular. El modelo de Regresión Lineal se encuentra implementado en una gran variedad de lenguajes y herramientas. Por otro lado, el modelo Rectangular no esta disponible, por lo tanto se implementó una solución. El desarrollo de los algoritmos fue realizada en Python.

El modelo de Regresión Lineal es un método matemático que modela la relación entre una variable dependiente Y, las variables independientes X_i y un término aleatorio ϵ [4]. Este modelo puede ser expresado como:

$$Y_t = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon \qquad (1)$$

Donde Y_t es denominada la variable dependiente, explicada o regresando. $X_1, X_2, ...X_p$ son conocidas como las variables explicativas, independientes o regresores. $\beta_0, \beta_1, \beta_2, ..., \beta_p$ son parámetros que miden la influencia que las variables explicativas tienen sobre el regresando. β_0 es la intersección o término constante, $\beta_i (i > 0)$ son los parámetros respectivos a cada variable independiente y p es el número de parámetros independientes a tener en cuenta en la regresión. El problema de la regresión consiste en elegir unos valores determinados para los parámetros desconocidos β_i, de modo que la ecuación quede completamente especificada. No se garantiza que estos valores coincidan exactamente con los parámetros reales del proceso generador. Para un problema de dos clases, se puede visualizar la operación como una partición del espacio de alta dimensionalidad de entrada con un hiperplano, esto es, todos los puntos a un lado del hiperplano son clasificados como "malignos", mientras que los demás son clasificados como "benignos".

El modelo Rectangular es un método matemático que modela la relación entre la variable de clase C y las variables independientes X_i [5]. Este modelo genera una ecuación:

$$H = \alpha_1 \leq X_1 \leq \beta_1 \; and \; \alpha_2 \leq X_2 \leq \beta_2 \; and \; ... \; and \; \alpha_i \leq X_i \leq \beta_i \qquad (2)$$

Donde H se denomida clase (espacio) de hipótesis, $X_1, X_2, ...X_i$ son los atributos de entrada, α_i y β_i son los parámetros de la instancia para un atributo X_i. Un algoritmo de aprendizaje en este caso encuentra una hipótesis particular $h \in H$, a fin de aproximar C tanto como sea posible. Aunque el experto define la clase de hipótesis H, los valores de los parámetros no son conocidos. El algoritmo de aprendizaje debe encontrar los valores de los parámetros que definen una hipótesis particular h. Dada una hipótesis h, podemos hacer una predicción sobre una entrada x, tal que:

- $h(x) = 1$ si h clasifica x como un ejemplo positivo
- $h(x) = 0$ si h clasifica x como un ejemplo negativo

Para un problema de dos clases, se puede visualizar la operación como un rectángulo que divide el espacio de alta dimensionalidad de entrada con respecto a los valores de las clases, esto es, todos los puntos dentro del rectángulo son clasificados como "malignos", mientras que los demás que están fuera del rectángulo son clasificados como "benignos".

Ahora, se debe estimar la capacidad del clasificador para la predicción sobre nuevas instancias. Exiten diversas técnicas para estimar esta capacidad. En esta investigación se calculará la matriz de confusión e indicadores como el error real, error aparente, sensibilidad, especificidad, recall y F_β.

La matriz de confusión es una tabla que muestra la distribución de los errores por las distintas categorías. Para el caso de dos clases, su forma es como la que se muestra en el Cuadro I.

Cuadro I
MATRIZ DE CONFUSIÓN

		Clase verdadera	
		Benigno	*Maligno*
Clase predicha	*Benigno*	a	b
	Maligno	c	d

Donde a es el número de predicciones correctas para instancias negativas, b es el número de predicciones incorrectas para instancias negativas, c es el número de predicciones incorrectas para instancias positivas y d es el número de predicciones correctas para instancias positivas.

La tasa de error se define como:

$$erate = \frac{\sum error}{N} \tag{3}$$

Donde N es la cantidad de casos.

El error real o tasa de error verdadera se define como la probabilidad de clasificar incorrectamente nuevos casos. Para ello, se utiliza el conjunto de datos de prueba sobre la ecuación 3.

El error aparente se define como la tasa de error obtenida al clasificar las mismas instancias de entrenamiento. Para ello, se utiliza el conjunto de datos de entrenamiento sobre la ecuación 3.

La sensibilidad indica la capacidad del clasificador para dar como casos positivos los casos que realmente son positivos, esto es, la proporción de positivos correctamente identificados. En el caso de esta investigación, la sensibilidad indicará la capacidad para detectar cáncer de mama en pacientes enfermos. La sensibilidad se calcula como:

$$se = \frac{a}{a+c} \tag{4}$$

La especificidad indica la capacidad del clasificador para dar como casos negativos los casos realmente negativos, esto es, la proporción de negativos correctamente identificados. En el caso de esta investigación, la especificidad indicará la capacidad para detectar la ausencia de cáncer de mama en pacientes sanos. La especificidad se calcula como:

$$es = \frac{d}{b+d} \tag{5}$$

La precisión es la proporción de verdaderos entre los predichos como positivos. Se refiere a la dispersión del conjunto de valores obtenidos de mediciones repetidas de una magnitud. Cuanto menor es la dispersión mayor la precisión. La precisión se calcula como:

$$precision = \frac{a}{a+b} \tag{6}$$

Donde $a + b \neq 0$.

Recall es la proporción de verdaderos positivos predichos de entre todos los positivos, esto es, la fracción de instancias relevantes que han sido clasificadas. Recall se calcula como:

$$recall = \frac{a}{a+c} \tag{7}$$

Donde $a + c \neq 0$. Tanto la precisión como recall son entendidas como medidas de relevancia.

El indicador F_β se considera como una media armónica que combina los valores de la precisión y de recall. De tal forma que:

$$F_\beta = (1+\beta^2)\frac{precision \times recall}{(\beta^2\ precision) + recall} \tag{8}$$

Si β es igual a uno, se está dando la misma ponderación (o importancia) a $precision$ que a $recall$, si β es mayor que uno se está dando más importancia a $recall$, mientras que si es menor que uno se está dando más importancia a la $precision$ que a $recall$.

II-B. Resultados

Esta sección proporciona una discusión de los resultados y analisis del modelo de Regresión Lineal y el modelo Rectangular. Los clasificadores fueron implementados en Python. El modelo de Regresión Lineal utiliza la librería "scikit-learn" la cual proporciona un conjunto de herramientas para análisis de datos. El modelo Rectangular se implementó en Python.

La validación del modelo es la evaluación de la eficacia de un modelo entrenado que va a ser usado en la práctica. Hay múltiples maneras para conducir la validación de un modelo, el método más utilizado es el llamado método "holdout". Esta forma de validación es usada cuando es imposible o impráctico crear un nuevo conjunto de datos. Los datos son divididos en dos partes, donde una parte es usada para estimar los parámetros del modelo y la otra parte es usada para medir la capacidad de predicción del modelo. En esta investigación, los datos fueron divididos en un conjunto de entrenamiento y un conjunto de prueba. El conjunto de entrenamiento corresponde al 70 % (n=398) de los datos, mientras que el porcentaje restante correspondiente al 30 % (n=171) serán destinados al conjunto de prueba. Los datos de ambos conjuntos fueron divididos utilizando una función proporcionada por "scikit-learn". El Cuadro II muestra un resumen de estos valores.

Cuadro II
VALORES CONJUNTO DE DATOS

Nombre	Wisconsin Diagnostic Breast Cancer
Dimensión	30
Instancias totales	569
Instancias en training set	398
Instancias en testing set	171

El Cuadro III muestra los tiempos de ejecución para cada modelo. El modelo de Regresión Lineal tomó alrededor de 0.010694 segundos en el proceso de entrenamiento y luego 0.000068 segundos para clasificar los datos del conjunto de prueba. Por otro lado, el modelo Rectangular tomó alrededor de 0.050399 segundos en el proceso de entrenamiento y luego 0.072698 segundos para clasificar los datos del conjunto de prueba. Como se observa, el modelo Rectangular toma más tiempo en ambos procesos. Nótese que el modelo Rectangular toma más tiempo durante la clasificación de nuevos datos, a diferencia del modelo de Regresión Lineal que toma más tiempo durante el entrenamiento que durante la prueba.

Cuadro III
TIEMPOS DE EJECUCIÓN

Nombre del modelo	Training/Prediction Time (seconds)
Regresión Lineal	0.010694s/0.000068s
Modelo Rectangular	0.050399s/0.072698s

Los Cuadros IV y V muestran las matrices de confusión para el modelo de Regresión Lineal y el modelo Rectangular respectivamente. Como se observa, en el Cuadro IV, de 63 casos benignos el modelo fue capaz de predecir 53 casos acertadamente y 10 casos erróneamente. De los 108 casos malignos, el modelo fue capaz de predecir los 108 casos acertadamente. En el caso del modelo Rectangular, ver Cuadro V, de 63 casos benignos el modelo fue capaz de predecir 52 casos acertadamente y 11 casos erróneamente. De los 108 casos malignos, el modelo fue capaz de predecir 84 casos acertadamente y 24 casos erróneamente.

Cuadro IV
MATRIZ DE CONFUSIÓN REGRESIÓN LINEAL

		Clase verdadera	
		Benigno	*Maligno*
Clase predicha	*Benigno*	53	0
	Maligno	10	108

Cuadro V
MATRIZ DE CONFUSIÓN MODELO RECTANGULAR

		Clase verdadera	
		Benigno	*Maligno*
Clase predicha	*Benigno*	52	24
	Maligno	11	84

El Cuadro VI condensa un análisis comparativo de los dos modelos, es decir, del modelo de Regresión Lineal y del modelo Rectangular. Como se observa, el modelo Rectangular tiene un pobre desempeño comparado con el modelo de Regresión Lineal.

La probabilidad de clasificar incorrectamente nuevos casos en el modelo de Regresión Lineal es 0.0585 y en el modelo Rectangular es 0.2047. Por lo tanto, es más probable clasificar nuevos casos erróneamente si se utiliza el modelo Rectangular.

El modelo Rectangular clasifica con un error aparente de 0.1533 las mismas instancias de entrenamiento, a diferencia del modelo de Regresión Lineal que clasifica con un error de 0.0427. Es decir, el modelo Rectangular clasifica más casos erróneamente durante el entrenamiento.

La sensibilidad es también conocida como recall. Este indicador se refiere a la capacidad de detectar a pacientes enfermos que realmente tienen cáncer de mama. Como se observa, el modelo de Regresión Lineal clasifica mejor los casos positivos, que realmente son positivos, a diferencia del modelo Rectangular.

El modelo de Regresión Lineal es capaz de clasificar casos negativos, casos realmente negativos, con un 100 % de acierto, a diferencia del modelo Rectangular que sólo acierta con un 22 % de acierto. Es decir, pacientes sanos son clasificados como realmente sanos por el modelo de Regresión Lineal.

Por último, F_β nos indica la medida de precisión de un modelo. Por lo tanto, el modelo de Regresión Lineal tiene una precisión de un 91 % y el modelo Rectangular tiene una precisión de un 74 %.

Cuadro VI
INDICADORES PARA LOS MODELOS

	Regresión Lineal	**Modelo Rectangular**
Error real	0.0585	0.2047
Error aparente	0.0427	0.1533
Sensibilidad	0.8413	0.8254
Especificidad	1.0	0.2222
Recall	0.8413	0.8254
F_β	0.9138	0.7482

Se concluye que el modelo de Regresión Lineal tiene un mejor desempeño que el modelo Rectangular propuesto en esta investigación. Una de las razones posibles, es que el modelo de Regresión Lineal es capaz de separar de mejor manera el espacio, donde por un lado existen más casos benignos y por el otro lado de la línea existen más casos malignos. A diferencia, el modelo Rectangular divide el espacio en un rectángulo, entonces muchos de los casos que realmente son benignos quedan dentro del área del rectángulo, la cual indica que pertenecen a casos malignos.

Los resultados muestran como técnicas de Machine Learning pueden proporcionar indicadores que pueden ayudar a la selección de la mejor técnica de clasificación para mejorar el diagnóstico de cáncer de mama.

III. Conclusión

El cáncer de mama es el más frecuente y es la principal causa de muerte en mujeres. Aún, cuando las tasas de incidencia global de cáncer en países menos desarrollados son la mitad de las encontradas en países más desarrollados, las tasas de mortalidad son similares. La sobrevida en cáncer es peor en países menos desarrollados, producto de la combinación de diagnósticos en etapas de enfermedad más avanzada, acceso limitado a tratamientos efectivos y en los tiempos adecuados.

La detección temprana de cáncer de mama puede ser predecida con un alto grado de precisión usando técnicas de Machine Learning. Esto puede resultar en la disminución del costo de prestaciones médicas y puede mejorar el tiempo en el cual un paciente reciba un tratamiento adecuado.

En esta investigación se ofreció un análisis comparativo entre el modelo de Regresión Lineal y el modelo Rectangular para proporcionar una forma de diagnosticar cáncer de mama. Se ha determinado que el modelo de Regresión Lineal es superior en todos los aspectos comparados al modelo Rectangular propuesto. El modelo de Regresión Lineal es capaz de separar de mejor manera el espacio, por lo tanto es capaz de diagnosticar de mejor forma nuevos pacientes.

La precisión de la predicción de los modelos seleccionados en esta investigación hace hincapié en la necesidad de emplear técnicas de Machine Learning no sólo en la predicción de datos de cáncer de mama, sino en otros dominios médicos en las que las predicciones de las condiciones son difíciles de diagnosticar.

Para finalizar, es deseable analizar detenidamente el modelo Rectangular para encontrar posibles mejoras al algoritmo. Siendo una de ellas, el análisis de las reglas encontradas con el objetivo de eliminar las reglas que no aporten al modelo o aquellas que sean redundantes.

Referencias

[1] M. Prieto, *Epidemiología del cáncer de mama en Chile.* Revista Médica Clínica Las Condes. 2011. URL: http://www.clc.cl/clcprod/media/contenidos/pdf/MED_22_4/2_Dra_Marta_Prieto.pdf

[2] H. You and G. Rumbe, *Comparative Study of Classification Techniques on Breast Cancer FNA Biopsy Data.* International Journal of Artificial Intelligence and Interactive Multimedia, Volumén 1, Número 3. 2010. URL: http://www.ijimai.org/journal/sites/default/files/IJIMAI20101_3_1.pdf

[3] W. H. Wolberg, W. N. Street and O. L. Mangasarian, *Wisconsin Diagnostic Breast Cancer.* November, 1995. URL: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

[4] C. M. Bishop, *Pattern Recognition and Machine Learning.* Springer, 2007. ISBN: 0387310738

[5] C. Meneses, *Machine Learning.* Universidad Católica del Norte, 2014.