Aplicación de algoritmos de clasificación para el análisis de tejido mamario y detección de cáncer de mama

José Rosario Villanueva Morales
Instituto Tecnológico de Celaya
11030385@itcelaya.edu.mx

José Jorge Lugo Rodríguez
Instituto Tecnológico de Celaya

Leonardo Landeros Vázquez
Instituto Tecnológico de Celaya
12030296@itcelaya.edu.mx

Daniel Omar Ramírez Buenrostro
Instituto Tecnológico de Celaya
12030161@itcelaya.edu.mx

Norma Verónica Ramírez Pérez
Instituto Tecnológico de Celaya
norma.ramirez@itcelaya.edu.mx

Resumen

La estrategia más efectiva para reducir la mortalidad por cáncer es la detección precoz, unida a actividades de prevención primaria y promoción de hábitos de vida saludables. El cáncer de mama es una patología difícilmente previsible en sus causas, puesto que la mayoría de los factores de riesgo conocidos son la edad, paridad, historia familiar de cáncer de mama, los cuales no son modificables y la disminución de las tasas de

mortalidad estaría en su *diagnóstico precoz*. El presente artículo describe como con ayuda de algoritmos de clasificación se puede aplicar el método de espectroscopia para que pueda realizar una detección oportuna de cáncer de mama, para realizarlo se utilizaron algoritmos de clasificación supervisada. Los datos fueron obtenidos de la base de datos “Breast Tissue” donada por INEB-Instituto de Engenharia Biomédica alojados en el repositorio UCI, con la finalidad de ver el comportamiento de los algoritmos en cuanto a la tasa de error, tiempo de ejecución y realizar una comparación de los resultados obtenidos del autor original J. Estrela da Silva, J. P. Marques de *Sá*. [7]

Palabras clave: algoritmos de clasificación, cáncer de mama, espectroscopia.

Abstract

The most effective strategy for reducing cancer mortality is early detection, attached to primary prevention and promotion of healthy lifestyles. Breast cancer is difficult to predict disease in their causes, since most of the known risk factors include age, parity , family history of breast cancer, which they are not modifiable and declining mortality rates would be in early diagnosis . This article describes how using classification algorithms can be applied spectroscopy method for you to make an early detection of breast cancer, for to do it we used supervised classification algorithms. Data were obtained from the database "Breast Tissue" donated by INEB-Instituto de Engenharia Biomédica housed in the repository UCI, in order to see the behavior of the algorithms in terms of error rate, runtime and a comparison of the results of the original author J. Estrela da Silva, J. P. Marques de Sá[7].

Keywords: *classification algorithms, breast cancer, spectroscopy*

1. Introducción

El cáncer de mama es uno de los cánceres tumorales que se conoce desde antiguas épocas. Desde hace varias décadas, el cáncer de mama se ha incrementado en grado notable alrededor del mundo, sobre todo en países occidentales y este crecimiento permanece, a pesar de que existen mejores instrumentos de diagnóstico como diversos programas de detección temprana, mejores tratamientos y mayor conocimiento de los factores de riesgo [9].

Actualmente existe mucha información acerca del cáncer de mama y de cómo puede ser detectado. Para el presente artículo se usó una base de datos de tejidos mamarios al que se le aplicó un algoritmo de clasificación que consiste en un procedimiento de agrupación de una serie de vectores con varios criterios, estos son por lo general distancia o similitud. La cercanía se define en términos de una determinada función de distancia, por ejemplo la distancia Euclídea que es la distancia "ordinaria" entre dos puntos de un espacio Euclídeo, la cual se deduce a partir del teorema de Pitágoras [3].

Aunque existen otras más robustas que permiten extenderla a variables discretas. La medida más utilizada para determinar la similitud entre los casos es la matriz de correlación entre los nxn casos, sin embargo, también existen muchos algoritmos que

se basan en la maximización de una propiedad estadística llamada verosimilitud que consiste en la credibilidad o congruencia de un elemento determinado dentro de una obra de creación concreta. Generalmente, los vectores de un mismo grupo (o *clusters*) comparten propiedades comunes. El conocimiento de los grupos puede permitir una descripción sintética de un conjunto de datos multidimensional complejo. De ahí su uso en minería de datos. Esta descripción sintética se consigue sustituyendo la descripción de todos los elementos de un grupo por la de un representante característico del mismo [2].

Los algoritmos que se aplicaron para hacer la clasificación son: naivebayes, multiclass classifier,random tree,Bftree y naivebayesNet.

Se aplicaron los 5 algoritmos para posteriormente obtener la información de cada uno de ellos, y hacer una comparación para determinar sus resultados entre ellos, y observar que algoritmo nos arroja resultados similares al que obtuvo el autor donante de la base de datos.

2. Métodos

Para el desarrollo del proyecto, se realizó una investigación exploratoria, descriptiva y correlacional, a través de las siguientes fases que se ilustran en el diagrama 1.

Diagrama 1: Fases para el desarrollo del proyecto

A continuación se describe cada una de las fases:

Fase 1: Revisión bibliográfica.

Para empezar con la investigación se accedió a información relativa a los tipos de aprendizaje, tipos de algoritmos de clasificación y métodos estadísticos con la finalidad de tener un panorama más amplio del tema. Por lo que a continuación se describen algunos conceptos fundamentales de dos tipos de aprendizaje:

No supervisado: el aprendizaje no supervisado es muy importante cuando se dispone de muestras sin etiquetas de clase, cuando el costo de etiquetarlas por un experto es alto o cuando los patrones pueden variar con el tiempo, por lo que es necesario primero procesar los datos para luego clasificarlos. La principal ventaja que presenta el aprendizaje no supervisado es que se puede obtener un conjunto de entrenamiento empleando muestras no etiquetadas valiéndose de algoritmos de agrupamiento [4].

Supervisado: se utilizan en problemas en los cuales se conoce a priori el número de clases y los reconocimientos de patrones representantes de cada clase. Básicamente consiste en que, para clasificar automáticamente una nueva muestra, se tiene en cuenta la información que se pueda extraer de un conjunto de objetos disponibles divididos en clases y la decisión de una regla de clasificación o clasificador.

Estos algoritmos tienen como objetivo determinar cuál es la clase, de las que ya se tiene conocimiento, a la que debe pertenecer una nueva muestra, teniendo en cuenta la información que se puede extraer del conjunto de entrenamiento[5].

Tipos de Algoritmos de Clasificación.

Naivebayes: Es un clasificador probabilístico fundamentado en el teorema de Bayes y algunas hipótesis simplificadoras adicionales. Es a causa de estas simplificaciones, que se suelen resumir en la hipótesis de independencia entre las variables predictoras, que recibe el apelativo de ingenuo [10].

Asume que la presencia o ausencia de una característica particular no está relacionada con la presencia o ausencia de cualquier otra característica, dada la clase de la variable.

Multiclass Classifier: El objetivo es construir una función que dado un nuevo punto de datos, se pueda predecir correctamente la clase a la que pertenece.

En el aprendizaje multiclase resuelve problemas de clasificación de más dos clases.

Mientras que algunos algoritmos de clasificación permiten, naturalmente el uso de más de dos clases, otros son por naturaleza algoritmos binarios; pero pueden convertirse en clasificadores multinomiales con una variedad de estrategias.

Random Tree: Es una colección (conjunto) de predictores es llamado forest further (término introducido por L. Breiman). La clasificación del método es la siguiente: los árboles clasificadores al azar toman el vector de características de entrada, lo clasifican con todos los árboles en el forest, y emite la etiqueta de clase que recibió la mayoría de los "votos"[8].

En caso de una regresión, la respuesta del clasificador es el promedio de las respuestas sobre todos los árboles.

Bftree: Es un algoritmo que sirve para la construcción de árbol de decisión. Esta método utiliza la división binaria para ambos atributos nominales y numéricos. Para los valores perdidos, se utiliza el método de casos fraccionarios [1].

Naive Bayes Net: Toda la dependencia en red bayesiana tiene que ser modelada. La red bayesiana puede ser aprendida por la propia máquina, o puede ser diseñada antes si se tiene suficiente conocimiento de las dependencias.

Para la validación de los algoritmos, hubo necesidad de utilizar métodos estadísticos, para determinar su eficiencia, por lo que se describe la siguiente definición:

Coeficiente Kappa: Es un coeficiente estadístico que se emplea para cuantificar el grado de acuerdo entre los observadores, corrige el factor azar. Es el estudio de fiabilidad por equivalencia o concordancia entre observadores. Cuando el valor obtenido es menor que -1 se dice que las variables tienen poca relación mientras si el valor es cercano a 1, se dice que existe una fuerte relación entre las variables [6].

Fase 2: Recopilación de datos.

En esta fase se analizaron los datos obtenidos para la clasificación, para esto se utilizaron datos de la base “Breast Tissue” (Tejidos mamarios) extraída del repositorio

UCI, la cual cuenta con: 110 instancias, 10 atributos y 6 diferentes clases, que se describen en la tabla 1.

Tabla 1. Atributos de la base de datos

I0	Impedancia (ohm) a la frecuencia 0
PA500	Angulo de fase a 500 KHz
HFS	inclinación de alta frecuencia del ángulo de fase
DA	distancia de impedancia entre los limites espectrales
AREA	espectro bajo el área
A/DA	área normalizada por DA
MAX IP	máximo de espectro
DR	distancia entre I0 y la parte real del máximo de punto de frecuencia
P	longitud de la curva espectral
Clases	Class car(carcinoma), fad (fibro-adenoma), mas (mastopatia), gla (glandular), con (conectiva), adi (adipose).

Cabe mencionar que la base de datos, ya venía normalizada, con respecto a ruido o valores nulos y perdidos, por lo que no hubo necesidad de realizar ningún procesado de los datos.

Fase 3: Aplicación de métodos algorítmicos.

Para la clasificación de las diferentes clases mencionadas en el fase 2, se utilizaron los algoritmos: naivebayes, multiclass classifier, random tree,Bftree y naivebayesNet, los cuales fueron analizados con el Sofware Weka, desarrollado en la Universidad de Waikato (Nueva Zelanda) bajo licencia GPL, lo cual ha impulsado que sea una de las suites más utilizadas en el área en los últimos años. Aunque existen otras herramientas que se pudieron utilizar, se decidió utilizar WEKA, debido a que contiene una gran

colección de algoritmos de clasificación, así como métodos estadísticos para la validación de la clasificación y métodos para la reducción de datos, eliminación de datos, transformaciones sobre los datos, tareas de clasificación, regresión, clustering, asociación y visualización, etc., para ilustrar este punto, podemos ver la caratula principal de dicho software en la figura 1.

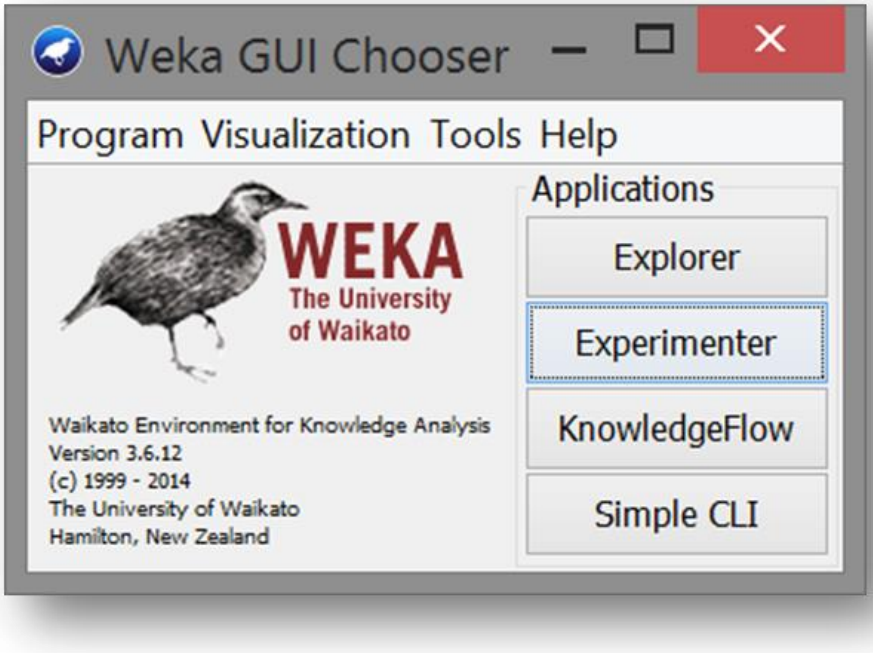

Figura 1. Pantalla principal de Weka

3. Resultados

Al realizar la clasificación de los algoritmos, cada uno de ellos con las características propias de su funcionamiento, arrojaron resultados muy parecidos, en la tabla 2 se desglosan los valores obtenidos por cada uno de ellos, como: las instancias clasificadas correctamente e incorrectamente, incluyendo el porcentaje de acierto y de error, el error absoluto medio, y para la validación de la clasificación también se ilustra los valores del coeficiente de kappa.

Tabla 2. Resultados obtenidos de los algoritmos de clasificación

Algoritmo	Instancias Clasificadas	% de Aciertos	Instancias mal clasificadas	% de error	Error absoluto medio	Coeficiente de kappa
NaiveBayes	75	70,7547	31	29,2453	39,0816	0.6484
Multiclass Classifier	71	66,9811	35	33,0189	96,0921	0.6027
Ramdom Tree	75	70,7547	31	29,2453	35,2397	0.6478
BF Tree	73	68,8679	33	31,1321	42,1754	0.6235
NaiveBayes Net	70	66,0377	36	33,9623	43,221	0.5915

Como se puede observar en la tabla 2, la mayoría de los algoritmos obtienen entre 70 y 75 instancias correctamente clasificadas, sin embargo no se puede confiar o darle credibilidad a estos resultados, puesto que nos está dando un 70% solo de clasificación, lo cual nos indica que no hemos encontrado un algoritmo adecuado para realizar una separación más óptima, debido a que los datos son complejos y medir el orden de cada uno de ellos se hace difícil. Sin embargo, con los resultados que se obtuvieron se evaluó el error cuadrático medio (ecm), el algoritmo NayveBayes nos arrojó un valor de 39.08 y Random Tree un valor de 29.24, debido a la cercanía o a la poca diferencia que existe entre un valor y otro es difícil determinar cuál de ellos resultó más eficiente. Sin embargo para hacer una distinción entre cada uno, se tomó en cuenta la validación del coeficiente de kappa, que nos indica que si se tiene un valor cercano a 1 las variables están fuertemente relacionadas mientras que -1 indica la pobre relación que se tiene con las variables. En la figura 2, se aprecia el valor de coeficiente de kappa de los 5 algoritmos.

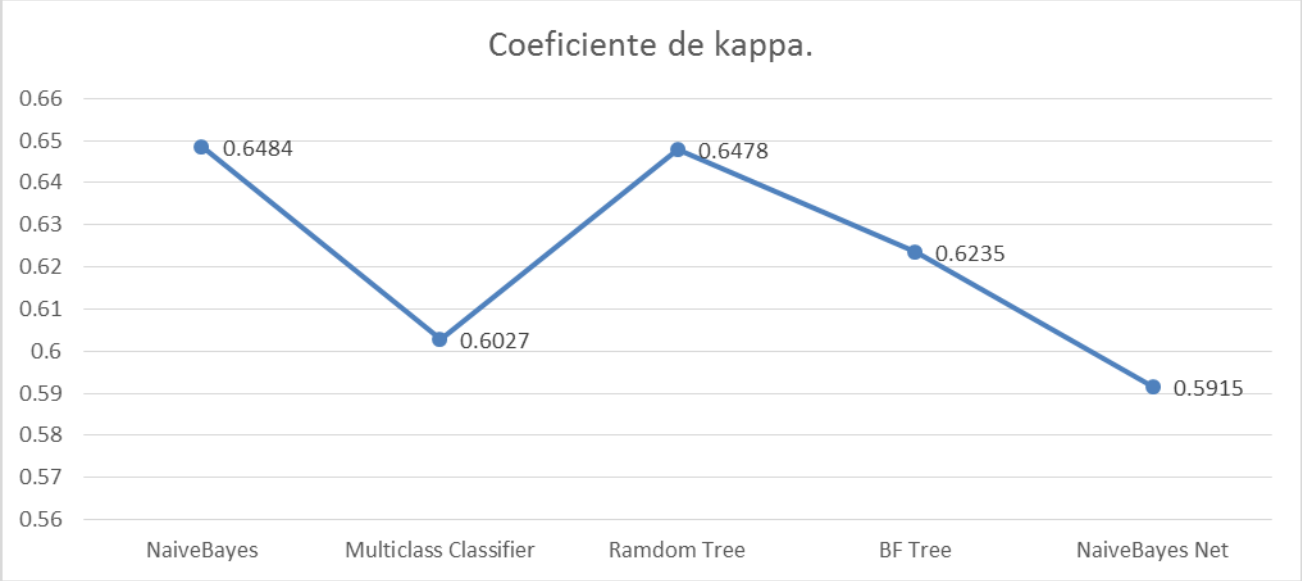

Figura 2. Coeficiente de Kappa

Si tomamos en cuenta el comportamiento de la gráfica se ve claramente que existen dos valores cercanos a 1, en este caso NayveBayes y Ramdom tree, y si verificamos la tabla 2, se ve claramente que son los que clasificaron mejor.

4. Discusión

Las grandes cantidades que se tienen en la actualidad, lleva a investigadores de temas relacionados con la inteligencia artificial, en especial en algoritmos de clasificación a realizar métodos más eficientes, que sean capaces de hacer una separación de clases con más precisión, con la aplicación que se realizó en este estudio motiva a hacer preguntas como ¿cuál es el algoritmo más óptimo?, ¿es importante el tipo de datos?

Estas preguntas solo se podrán contestar haciendo un análisis exhaustivo al comportamiento de los algoritmos, además de considerar también el tipo de máquina que se está utilizando, pues al analizar bases de datos complejas, el coste computacional es muy alto. Por otro lado, en este estudio la separación de clases de la base de datos “Breast Tissue”, los algoritmos aplicados resultaron ser ineficientes en

cuanto a su clasificación, por lo que se concluye que el algoritmo LVQ Network aplicado por el autor original resulto ser más eficiente para este tipo de datos.

Bibliografía:

[1] Manuel Antolín Ayuso, Miguel Ángel Barcenilla Mancha . (SA). Minería de Datos: Intrusiones de Red, de Universidad Carlos III de Madrid Sitio web: http://www.it.uc3m.es/jvillena/irc/practicas/07-08/IntrusionesDeRed.pdf

[2] De la Cruz mantilla, Azucena Sarai - Linares Valdivia, Juan Carlos. (2014). Diseño de un modelo computacional basado en algoritmos de agrupamiento para mejorar el tiempo de respuesta y la correspondencia de resultados de un sistema de búsqueda de información bibliográfica. de Universidad Nacional de Trujillo Sitio web: http://www.inf.unitru.edu.pe/revistas/2014/5.pdf

[3] ARNOLD, S.F. (SA). Distancia Euclidea. 2015, de UV Sitio web: http://www.uv.es/ceaces/multivari/cluster/d_euclidea.htm

[4] Facultad de Matemática y Computación de la Universidad de Oriente (SA). Algoritmos de clasificación supervisada. 2015, de EcuRed Sitio web: http://www.ecured.cu/index.php/Algoritmos_de_clasificaci%C3%B3n_supervisada

[5] Facultad de Matemática y Computación de la Universidad de Oriente. (SA). Algoritmos de clasificación no-supervisada. 2015, de EcuRed Sitio web: http://www.ecured.cu/index.php/Algoritmos_de_clasificaci%C3%B3n_supervisada

[6] Geoffrey R. Norman, John E. De Burgh Norman, David L. Streiner, 1996, Bioestadistica,Harcourt.

[7] JP Marques de Sá, (2000). Breast Tissue Data Set. 2015, de UCI Sitio web: http://archive.ics.uci.edu/ml/datasets/Breast+Tissue

[8] OpenCv. (SA). Random Trees, de OpenCv Sitio web: http://docs.opencv.org/modules/ml/doc/random_trees.html

[9] Miguel Martin Jiménez, Oncología-Cáncer de mama, Aran ediciones sl.

[10] Scikit-learn developers. (2014). Naive Bayes. 2015, de scikit-learn Sitio web: http://scikit-learn.org/stable/modules/naive_bayes.html