Artículo **Revista de Ingeniería Biomédica y Biotecnología**
Diciembre, 2020 Vol.4 No.12 1-12

Evaluación de la regresión logística como clasificador de espectros Raman en el diagnóstico automático de cáncer de mama

Evaluation of logistic regression as Raman spectra classifier for automatic breast cancer diagnosis

DE LA PARRA-GONZÁLEZ, Israel†*[1], LUNA-ROSAS, Francisco Javier[1], RODRÍGUEZ-MARTÍNEZ, Laura Cecilia[1], y FRAUSTO-REYES, Claudio[2]

[1]TecNM/Instituto Tecnológico de Aguascalientes, Departamento de Sistemas y Computación, Av. A. López Mateos 1801 Ote. Fracc. Bona Gens, Aguascalientes, Aguascalientes, México. C.P. 20256
[2]Centro de Investigaciones en óptica, Unidad Aguascalientes, Prol. Constitución 607, Fracc. Reserva Loma Bonita. Aguascalientes, Aguascalientes, México. Código Postal 20200

ID 1er Autor: *Israel De La Parra-González* / **ORC ID:** 0000-0002-7403-0128, **CVU CONACYT ID:** 227008, **Becario-PNPC:** 582938

ID 1er Coautor: *Francisco Javier Luna-Rosas* / **ORC ID:** 0000-0001-6821-4046, **arXiv Author ID:** arXivFco19, **CVU CONACYT ID:** 87098

ID 2do Coautor: *Laura Cecilia Rodríguez-Martínez*

ID 3er Coautor: *Claudio Frausto-Reyes* / **ORC ID:** 0000-0002-5728-6455

DOI: 10.35429/JBEB.2020.12.4.1.12 Recibido 10 de Septiembre, 2020; Aceptado 30 Diciembre, 2020

Resumen

Se evaluó la regresión logística (LR) como clasificador en el diagnóstico de cáncer de mama basado en espectros Raman. El común de los estudios publicados en el tema utilizan reducción dimensional para generar el clasificador. En cambio, en este trabajo se propuso observar el efecto de utilizar todos los valores de intensidad registrados en el espectro como variables de entrada al algoritmo. Se utilizó validación cruzada dejando uno fuera, registrando, eficacia de clasificación, sensibilidad y especificidad. Se emplearon espectros Raman tomados de tejido mamario previamente diagnosticado mediante análisis histopatológico, algunos de tejido sano y otros de tejido con cáncer. Cada espectro se conforma de 605 valores de intensidad en el rango de 687 a 1781 cm^{-1}. El algoritmo de LR tuvo una tasa de clasificación correcta de 100%. Como comparación, se evaluaron de la misma manera: 1) un modelo logístico antecedido por reducción dimensional mediante Análisis de Componentes Principales (PCA+LR), 2) dos clasificadores obtenidos con el algoritmo de K vecinos más cercanos ponderados, y 3) un clasificador mediante el algoritmo de Bayes ingenuo (NB). Se encontró que PCA+LR y NB igualaron el 100% de clasificación correcta. Sin embargo, PCA+LR requiere un mayor tiempo de ejecución.

Diagnóstico de cáncer de mama, Espectroscopía Raman, Regresión logística

Abstract

We evaluated logistic regression as a classifier in the diagnosis of breast cancer based on Raman spectra. Common studies published in the subject use dimensional reduction techniques to generate the classifier. Instead, we proposed to observe the effect of using all intensity values recorded in the spectra as input variables to the algorithm. We used leaving one out cross-validation measuring classification accuracy, sensitivity and specificity. We used Raman spectra taken from breast tissue previously diagnosed by histopathological analysis, some from healthy tissue and some from tissue with cancer. Each spectrum is formed by 605 intensity values in the range of 687 to 1781 cm^{-1}. Logistic regression classifier exhibited 100% classification accuracy. To establish comparative references, we evaluated in the same way: 1) a logistic model preceded by dimensional reduction with Principal Component Analysis (PCA+LR), 2) two classifiers obtained with weighted K nearest neighbors algorithm, and 3) a classifier using the naive Bayes (NB) algorithm. We found that PCA+LR and NB showed the same performance of 100% in classification accuracy. Nevertheless, PCA+LR requires more processing computational time.

Breast cancer diagnosis, Raman spectroscopy, Logistic regression

Citación: DE LA PARRA-GONZÁLEZ, Israel, LUNA-ROSAS, Francisco Javier, RODRÍGUEZ-MARTÍNEZ, Laura Cecilia, y FRAUSTO-REYES, Claudio. Evaluación de la regresión logística como clasificador de espectros Raman en el diagnóstico automático de cáncer de mama. Revista de Ingeniería Biomédica y Biotecnología. 2020. 4-12: 1-12

* Correspondencia del autor (Correo electrónico: israel.delaparra@yahoo.com)
† Investigador contribuyendo como primer autor

© ECORFAN- Taiwan

1. Introducción

Una de las propuestas actuales para automatizar y mejorar los procedimientos de diagnóstico de cáncer, y en particular, cáncer de mama, es mediante espectroscopía Raman. La idea general consiste en obtener el espectro Raman de la muestra de tejido sospechosa para poder analizarlo y clasificarlo según sus características, que están relacionadas con los componentes biomoleculares presentes en la muestra (Sathyavathi et al., 2015). Este tipo de diagnóstico automático de cáncer de mama es un tema que se ha abordado desde diversos puntos de vista y para el que se han encontrado resultados muy alentadores que muestran su potencial para introducirse en ambientes clínicos (Jermyn et al., 2016; Ralbovsky & Lednev, 2019).

La proliferación de los estudios sobre aplicaciones de la espectroscopía Raman en el área médica, entre ellas, el diagnóstico automático de cáncer, se debe en gran parte a que es una técnica que tiene una alta sensibilidad para detectar cambios moleculares como los atribuibles al cáncer. Esto es porque las características del espectro Raman están cercanamente relacionadas con las características moleculares de la muestra que se esté analizando (Choo-Smith et al., 2002). Además, el avance tecnológico en los dispositivos ópticos necesarios para esta técnica ha aumentado su potencial para llevarla a entornos de diagnóstico in-vivo (Jermyn et al., 2016). La investigación al respecto sigue desarrollándose con enfoques diferentes y desde diversas disciplinas.

Parte importante en las técnicas para diagnóstico de cáncer mediante espectros Raman son los algoritmos necesarios para analizar dichos espectros e identificar las características relevantes que permitan el diagnóstico acertado. Este problema se origina porque las características moleculares que se observan en el espectro Raman tienen relación con muchos factores propios del tejido de muestra y no sólo con la presencia o ausencia de cáncer, haciendo difícil su identificación y generalización.

Existen muchas propuestas de algoritmos diversos para este análisis, siendo los más populares aquéllos que involucran aprendizaje automático del tipo supervisado, es decir, que realizan un entrenamiento usando espectros que corresponden a tejido ya diagnosticado por medio de análisis histopatológico (Jermyn et al., 2016).

Otro algoritmo muy comúnmente utilizado para realizar el análisis y la clasificación de espectros Raman, es el análisis de componentes principales (PCA). Este método no es clasificador, ni supone un entrenamiento de aprendizaje automático supervisado, sino que se utiliza para extraer información significativa en el análisis de un conjunto de espectros, generalmente para hacer reducción dimensional de las variables predictoras que pueden diferenciar las clases de tejido antes de introducirse a otro algoritmo que sí sea clasificador (Jermyn et al., 2016).

Entre los algoritmos clasificadores de aprendizaje supervisado usados en este tipo de aplicación es frecuente encontrar clasificadores que utilizan probabilidad condicional (Bayes) (Martínez Romo et al., 2015) o basados en el vecino más cercano (“Nearest Neighbor”) (Q. B. Li, Wang, Liu, & Zhang, 2015; Q. Li, Gao, & Zhang, 2014; Q. Li, Hao, & Xu, 2017). Muy comúnmente, los algoritmos clasificadores empleados hasta el momento usan un conjunto reducido de variables predictoras obtenido mediante algoritmos como PCA, descomposición de componentes moleculares, o evidencia previa reportada en otros estudios (Fallahzadeh, Dehghani-Bidgoli, & Assarian, 2018; Kim, Lee, Min, Byun, & Lee, n.d.; Krishnamoorthy, Prakasarao, Srinivasan, Sivarama, & Singaravelu, 2019; Martínez Romo et al., 2015; Sathyavathi et al., 2015; Vanna et al., 2020). En este tipo de aplicaciones, el conjunto de espectros se procesa para identificar características relevantes distintivas, o bien, se toman solamente ciertos puntos o mediciones de los espectros antes de realizar el modelo de clasificación.

La regresión logística (LR) es un algoritmo de aprendizaje supervisado de amplio uso en el área biomédica, incluso considerado el más popular en el análisis de datos epidemiológicos cuando la medida de enfermedad es binaria (Kleinbaum & Klein, 2010).

ISSN 2523-6857
ECORFAN® Todos los derechos reservados

Este algoritmo ya se ha empleado en otras ocasiones para diagnóstico de cáncer de mama mediante espectros Raman (Bi, Rexer, Arteaga, Guo, & Mahadevan-Jansen, 2014; Dingari et al., 2013; Haka et al., 2002, 2005; Kong et al., 2014; Sathyavathi et al., 2015), pero generalmente acompañado de alguna reducción dimensional para crear el modelo logístico con un número reducido de variables predictoras.

En el presente trabajo se propuso aplicar un clasificador por LR para diagnóstico automatizado de cáncer de mama mediante espectros Raman, utilizando la intensidad en cada punto del espectro como variables predictoras; es decir, sin realizar reducción previa en el número de variables de entrada para el clasificador. Las razones que justifican la elección de este algoritmo se relacionan con su amplio uso en temas biomédicos, lo cual se esperaría que facilite su aceptación; pero también con su simplicidad y facilidad de implementación que podría permitir migrarse a diferentes plataformas tecnológicas.

Además, el desempeño de dicho clasificador se comparó contra un clasificador también basado en LR, pero que sí utiliza reducción previa de variables predictoras mediante PCA, y contra dos algoritmos clasificadores comunes (Bayes ingenuo y K-vecinos cercanos ponderados) en los cuales tampoco se realizó reducción previa de variables. Al compararse el presente trabajo con los ejemplos representativos que se citan, puede apreciarse la similitud en las técnicas propuestas para el diagnóstico y también la novedad que presenta al proponer la creación del modelo logístico sin realizar reducción dimensional previa.

2. Materiales y métodos

El objetivo principal de este estudio es evaluar la efectividad de clasificación que presenta el algoritmo de regresión logística (LR) al ser aplicado con espectros Raman de muestras de tejido mamario con la intención de apreciar la factibilidad de utilizarlo en herramientas de diagnóstico automático de cáncer de mama. Se evaluó su efectividad en dos escenarios: en el primero, las variables predictoras fueron todos los valores de intensidad registrados en el espectro; y en el segundo, las variables predictoras fueron los coeficientes correspondientes a los componentes principales obtenidos mediante PCA.

De esta manera se pudo juzgar sobre la pertinencia de aplicar o no el PCA. También se evaluaron dos algoritmos clasificadores similares a la LR, el Bayes ingenuo y el K-vecinos más cercanos ponderados, con la intención de observar cuál logra una mejor clasificación con este tipo de datos y utilizando todos los valores de intensidad del espectro como variables predictoras.

2.1. Espectros Raman

Los espectros que se utilizaron para este trabajo fueron obtenidos por investigadores del Centro Universitario de los Lagos de la Universidad de Guadalajara, Jalisco, México. Fueron tomados de cortes histológicos realizados a tejidos mamarios sanos y con cáncer obtenidos de biopsias de mama y fijados en formalina. Los tejidos con cáncer fueron diagnosticados como carcinoma ductal infiltrante. La obtención de espectros Raman se realizó con un sistema Raman Renishaw modelo 1000-B que usa un diodo láser de λ= 830 nm y una rejilla de 600 líneas mm^{-1}. El láser fue centrado sobre las muestras con un microscopio Leica modelo DMLM (objetivo de 50x) con una potencia de 35 mW aproximadamente.

Cada espectro fue registrado en la región de 680 a 1780 cm^{-1} con una exposición de tiempo de 10 s y resolución menor a 2 cm^{-1}. El sistema Raman fue calibrado con un semiconductor de silicio en el pico Raman de 520 cm^{-1}. Con este entorno experimental, se colectaron espectros Raman en zonas de tejido sano y zonas de tejido dañado. Cada espectro consta de 605 puntos de intensidad.

2.2. Pre-procesamiento de los espectros Raman

Cuando se mide el espectro Raman, particularmente de una muestra biológica, se agregan también contribuciones no deseadas debidas a diferentes factores. De esta manera, la señal recibida se conforma con el espectro Raman de la muestra y también con algunos artefactos añadidos. Las contribuciones no deseadas más comunes son: a) ruido debido a la fluorescencia de la muestra o su sustrato, y b) ruidos de alta frecuencia que pueden ser picos debidos a rayos cósmicos, o ruido de disparo que es un ruido aleatorio presente a lo largo de todo el espectro.

ISSN 2523-6857
ECORFAN® Todos los derechos reservados

Para poder realizar un mejor análisis de los espectros reales de las muestras biológicas, se necesita eliminar o minimizar estos artefactos no deseados presentes en las señales obtenidas. Este proceso se denomina comúnmente pre-procesamiento de los espectros y se conforma de diferentes algoritmos según la aplicación específica en la que se esté trabajando (Bocklitz, Guo, Ryabchykov, Vogler, & Popp, 2016).

En el caso particular de los espectros utilizados para el presente trabajo, el pre-procesamiento que se realizó para eliminar los ruidos de alta frecuencia fue implementado con un filtro Savitzky-Golay de tercer orden y ancho de ventana de 11 puntos (Savitzky & Golay, 1964). Y para eliminar la componente de fluorescencia en los espectros se utilizó el Algoritmo Raman Vancouver (Zhao, Lui, I., & Zeng, 2010) con un polinomio de sexto orden y un umbral de 2.5%.

Además, para comparar y clasificar espectros Raman es necesario tener puntos de referencia comunes dado que las intensidades registradas pueden variar según la calidad de enfoque que se hace con el microscopio o algunos factores de la preparación de la muestra. La manera común de obtener estos puntos de referencia es normalizando todos los espectros. En este caso se utilizó una normalización con respecto a la intensidad del pico ubicado en 1444 cm^{-1} dado que se revisaron varias alternativas de normalización y esta opción fue la que presentó mejores resultados al analizar el potencial discriminatorio que se obtiene (Martínez Romo et al., 2015).

2.3. Regresión logística (LR – Logistic Regression)

La LR es una técnica de modelado matemático que se ha usado para desarrollar sistemas clasificadores por su capacidad de expresar una variable binaria en función de un conjunto de variables predictoras. El modelo logístico tiene la forma de la Ec. (1), donde **x** es una nueva observación caracterizada por las variables predictoras $x_1, x_2, \dots, x_m$ y $P(\mathbf{x})$ define su probabilidad de pertenencia a una determinada clase y puede redondearse a los valores 0 y 1 para efectos clasificatorios (Kleinbaum & Klein, 2010; Ng, n.d.; Sayad, n.d.).

$$P(\mathbf{x}) = \frac{1}{1+ e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2+\cdots+\theta_m x_m)}} \qquad (1)$$

Para generar el modelo logístico que pueda usarse como clasificador, se definen los coeficientes $\theta_0, \theta_1, \theta_2, \dots, \theta_m$ mediante procedimientos de aprendizaje supervisado buscando los valores que minimicen el error de clasificación para los datos de entrenamiento.

En esta implementación, los espectros Raman que se utilizaron fueron capturados en 605 puntos, y en el algoritmo se tomaron los valores de intensidad en cada uno de esos puntos como variables de entrada para el modelo, siendo así, 605 variables predictoras.

Los algoritmos y las pruebas realizadas fueron programados en Octave 4.0.3 (Eaton, Bateman, Hauberg, & Wehbring, 2015), con implementación propia de la LR basada en una plantilla propuesta para fines didácticos (Ng, n.d.) y utilizando la función de optimización fminunc( ). Los mismos algoritmos y pruebas también fueron programados en R versión 3.4.2 (R Core Team, 2017) realizando la traducción del código generado en Octave y utilizando la función de optimización equivalente ucminf( ) (Nielsen & Mortensen, 2016). Aunque los resultados obtenidos en ambas implementaciones no fueron idénticos, debido a las diferencias entre los algoritmos de optimización utilizados, las métricas que se reportan en este trabajo (eficacia, sensibilidad y especificidad) sí lo fueron.

2.4. Análisis de Componentes Principales (PCA) + Regresión Logística (LR)

Aunque existen propuestas diversas para extraer información relevante de los espectros antes de realizar un clasificador, se eligió evaluar el algoritmo de Análisis de Componentes Principales (PCA) debido a que es el más común entre las propuestas revisadas en donde se ha aplicado LR para clasificar espectros Raman de tejido con cáncer y sano (Bi et al., 2014; Dingari et al., 2013; Haka et al., 2002, 2005; Kong et al., 2014; Sathyavathi et al., 2015). De esta manera, se compararon resultados obtenidos con este esquema común, contra los obtenidos con el esquema de clasificación con LR sin reducción dimensional ni selección de características espectrales distintivas.

ISSN 2523-6857
ECORFAN® Todos los derechos reservados

En el esquema PCA+LR, el análisis inicial descompone los espectros en varios Componentes Principales (PC), ordenados según su aporte a la varianza de los datos, y asocia un peso a cada uno de ellos para cada espectro. Así, dichos pesos pueden ser ahora las variables predictoras para el modelo logístico. Esta elección de las variables predictoras conlleva la reducción dimensional de que se ha hablado pues, en lugar de utilizar 605 intensidades que conforman cada espectro, es posible utilizar un número muy pequeño de variables, generalmente entre 1 y 5. El número máximo de PC que se generan es uno menos que el número de espectros analizados. Sin embargo, el número real de pesos que se utiliza suele ser mucho menor debido a que la gran mayoría de la varianza de los datos está explicada con los primeros PC (Jermyn et al., 2016; Y. Z. Li et al., 2013). Por ejemplo, en los estudios revisados que aplicaron PCA+LR (Haka et al., 2002) se reportó clasificación exitosa empleando 2, 3 o 4 PC dentro de los 10 primeros como máximo.

Para observar el comportamiento del algoritmo PCA+LR según cuántas variables predictoras se utilicen, el modelo clasificador se realizó varias veces. En cada una, se tomó un conjunto diferente de pesos asociados a los primeros 10 PC como variables predictoras. Se comenzó con las 10 posibilidades de tomar sólo un peso como variable predictora, luego con las 45 posibilidades de tomar dos pesos como variables predictoras, y así se continuó hasta encontrar la mínima cantidad de pesos que se requieren para obtener el máximo porcentaje de clasificación correcta.

Los algoritmos y las pruebas realizadas fueron programados con implementaciones ya existentes del PCA, princomp( ) en Octave 4.0.3 (Eaton et al., 2015) y también prcomp( ) en R versión 3.4.2 (R Core Team, 2017) y los algoritmos de LR programados según se explicó en la sección anterior.

2.5. K vecinos más cercanos ponderados (WKNN – Weighted K Nearest Neighbors)

El algoritmo WKNN es un método en el que una nueva observación es clasificada según su similitud con observaciones previas tomadas de un conjunto de aprendizaje.

Esta técnica requiere el establecimiento de tres parámetros: 1) la función de distancia que establece la similitud o cercanía entre observaciones con base en sus variables predictoras; entre menor distancia, mayor similitud. 2) El valor *"k"*, que establece la cantidad de observaciones previas cercanas que se usarán para la clasificación. Por ejemplo, si *k=1*, significa que solamente se utiliza la observación más cercana a la nueva observación para decidir su clase. Conforme *k* aumenta, se van utilizando también las siguientes observaciones más cercanas a la nueva observación para decidir su clase. Y 3) la función "kernel" que le otorga un peso o valor de similitud a cada distancia calculada entre las observaciones cercanas utilizadas y la nueva observación. Al final, el algoritmo asigna, a la nueva observación, la clase que obtuvo la mayor sumatoria de pesos para las observaciones cercanas utilizadas pertenecientes a ella (Hechenbichler & Schliep, 2004).

Se eligió utilizar el algoritmo WKNN para comparar sus resultados con los obtenidos al usar regresión logística (LR) ya que es un algoritmo simple e intuitivo para realizar clasificación y ya se han reportado algunos resultados de la utilización de variaciones del mismo en el diagnóstico de cáncer de mama sobre espectros Raman (Dingari et al., 2013; González-Solís et al., 2016; Q. B. Li et al., 2015; Q. Li et al., 2014, 2017). De esta manera, se pretende valorar los resultados obtenidos con LR al compararse con otros algoritmos usando los mismos datos, las mismas variables predictoras y el mismo esquema de validación dejando uno fuera.

Los experimentos se realizaron con una implementación del algoritmo WKNN (Hechenbichler & Schliep, 2004; Samworth, 2012; Schliep & Hechenbichler, 2016) en R versión 3.4.2 (R Core Team, 2017). Dicha implementación utiliza la función de distancia Minkowski, permitiendo establecer el parámetro *p* de la misma, siendo *p=2* el valor por omisión. En este caso se utilizó el valor por omisión, lo cual es equivalente a usar la distancia Euclidiana. También se usó la función "kernel" por omisión que corresponde a la propuesta de asignación óptima de pesos (Samworth, 2012). Finalmente, la implementación permite buscar el mejor valor de *k* desde 1 hasta un valor máximo especificado.

ISSN 2523-6857
ECORFAN® Todos los derechos reservados

Para ello, toma el conjunto de entrenamiento recibido y genera el clasificador para cada valor de *k* verificándolo nuevamente con validación cruzada dejando uno fuera (Schliep & Hechenbichler, 2016). Después de todas las validaciones, el algoritmo elige el valor de *k* más pequeño que haya logrado la mayor eficacia de clasificación y con este valor de *k* genera el clasificador que es, ahora sí, entrenado con los espectros del conjunto de entrenamiento.

2.6. Bayes ingenuo (NB – Naïve Bayes)

El algoritmo de clasificación NB es un método probabilístico en el que una nueva observación es clasificada según estimaciones de las probabilidades condicionales asociadas a que pertenezca a una clase o a otra. El cálculo de estas estimaciones se fundamenta en asumir que las variables predictoras son independientes entre sí y que siguen una distribución normal para cada clase. Con el proceso de entrenamiento se establecen los parámetros de media y desviación estándar que caracterizan las funciones de densidad de probabilidad de pertenecer a cada clase para cada variable. Cuando se tiene un nueva observación para clasificarse, se obtienen las estimaciones de sus probabilidades según las funciones de densidad del entrenamiento y finalmente se determina cuál clase tiene mayor probabilidad de ser aquélla a que corresponde (Martínez Romo et al., 2015).

Se eligió utilizar el algoritmo NB para comparar sus resultados con los obtenidos al usar regresión logística y WKNN ya que es un algoritmo simple que tiene un costo computacional reducido, no requiere ajuste de parámetros para implementarse y ya se han reportado buenos resultados de su utilización en el diagnóstico de cáncer de mama sobre espectros Raman (Martínez Romo et al., 2015), así como también en otros tipos de cáncer o con variaciones del algoritmo (Luo, Chen, Mao, & Jin, 2013; Pence, Patil, Lieber, & Mahadevan-Jansen, 2015).

Los experimentos se realizaron con una implementación del NB (Meyer, Dimitriadou, Hornik, Weingessel, & Leisch, 2017) en R versión 3.4.2 (R Core Team, 2017).

2.7. Validación cruzada dejando uno fuera

Cuando se implementan algoritmos de aprendizaje supervisado en alguna aplicación, es muy importante revisar la efectividad de los modelos clasificadores generados y tener una visión de su probable desempeño ante datos nuevos o futuros que podrían presentarse. Esta información se obtiene comúnmente mediante técnicas de validación cruzada (Gautam, Vanga, Ariese, & Umapathy, 2015) que, cuando se aplican en sistemas de diagnóstico, estiman la sensibilidad y la especificidad del clasificador (Dougherty, 2013). La validación cruzada es una técnica para estimar el desempeño de un modelo clasificador que permite tener una visión de su probable desempeño ante datos nuevos o futuros que podrían presentarse.

La forma básica de validación cruzada se llama validación cruzada de k iteraciones, en la que el conjunto de datos de entrada se divide aleatoriamente en k subconjuntos de igual tamaño. Después, se realizan k pruebas independientes tales que en cada una de ellas uno de los k subconjuntos se reserva para validar el modelo mientras que el entrenamiento solamente utiliza los otros k-1 subconjuntos. Finalmente, los resultados de las k pruebas se condensan para obtener una estimación estadística de la calidad del clasificador (Rodriguez, Godoy, Mateos, & Zunino, 2017).

En este trabajo se eligió la evaluación de modelos mediante validación cruzada dejando uno fuera (Gautam et al., 2015), lo que corresponde a tomar el valor de k igual que el tamaño del conjunto de datos. Así, el conjunto de validación siempre es un solo espectro mientras que el entrenamiento se realiza con el resto de los espectros.

Por ejemplo, para el caso en que se tiene un conjunto de 100 espectros, el entrenamiento se realiza con 99 de ellos. Sin embargo, la evaluación se debe repetir 100 veces para cada algoritmo clasificador que se pretenda evaluar, en cada una de las cuales un espectro diferente se utiliza como conjunto de validación. Así que, para un determinado algoritmo de clasificación, cada espectro se utiliza como validación una sola vez y el desempeño se estima con el resultado de las 100 iteraciones.

ISSN 2523-6857
ECORFAN® Todos los derechos reservados

2.8 Efectividad de un algoritmo clasificador

Cuando se estima el desempeño de un algoritmo clasificador, es común realizarlo con tres métricas: eficacia (o exactitud), sensibilidad y especificidad (Dougherty, 2013). La eficacia o exactitud de un clasificador es el porcentaje de predicciones correctas divididas entre el total de predicciones. En esta aplicación particular, la sensibilidad será la probabilidad del clasificador de identificar correctamente los espectros de tejido con cáncer. Y la especificidad será la probabilidad de identificar correctamente los espectros de tejido sano. Para establecer la clasificación "correcta" de los espectros se tomó como referencia el resultado de diagnóstico mediante análisis histopatológico. Los cálculos de las métricas se realizaron con las fórmulas de Ec. (2), Ec. (3) y Ec. (4) (Dougherty, 2013).

$$Eficacia = \frac{TP + TN}{N} \quad (2)$$

$$Sensibilidad = \frac{TP}{p} \quad (3)$$

$$Especificidad = \frac{TN}{n} \quad (4)$$

Donde *TP* es el número de espectros que corresponden a tejido con cáncer y son clasificados correctamente. *TN* es el número de espectros que corresponden a tejido sano y son clasificados correctamente. *N* es el número total de espectros evaluados. *p* es el número de espectros evaluados que corresponden a tejido con cáncer. Y *n* es el número de espectros evaluados que corresponden a tejido sano.

3. Resultados

Espectros Raman

A los espectros obtenidos se les aplicó el pre-procesamiento que se mencionó en la sección 2.2 para eliminación de ruidos de alta frecuencia y componente de fluorescencia y se obtuvieron resultados como los que se muestran (Gráfico 1).

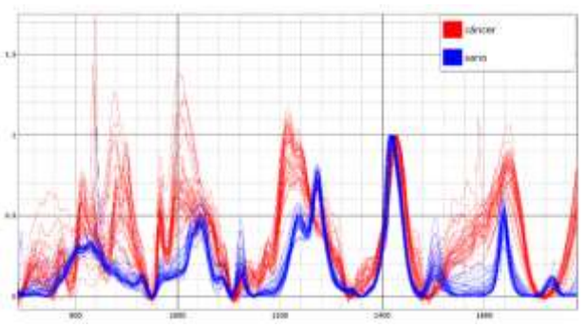

Gráfico 1 Espectros obtenidos después de realizar el pre-procesamiento

3.1 Efectividad de la clasificación

La Tabla 1 muestra las métricas de evaluación obtenidas para los algoritmos con los espectros utilizados.

Algoritmo	Eficacia (%)	Sensibilidad (%)	Especificidad (%)
LR	100	100	100
PCA+LR*	100	100	100
NB	100	100	100
PCA+LR**	98.8	97.4	100
WKNN (k=6)	98.8	97.4	100
WKNN (k=1)	97.6	94.7	100
* al menos 3 PC en Octave y al menos 2 PC en R ** 1 y 2 PC en Octave y 1 PC en R			

Tabla 1 Métricas obtenidas

En tres casos, dos usando LR y el otro usando NB, se obtuvo una eficacia de 100 %. Esto significa que, en cada uno, todas las veces que se generó el modelo dejando un espectro fuera del conjunto de entrenamiento, la clasificación fue correcta al aplicarse el modelo para el espectro que había quedado fuera. Los dos casos que corresponden a LR fueron, uno sin análisis previo de los espectros y el otro basado en el análisis de componentes principales (PCA) utilizando al menos 3 componentes principales (PC) en Octave y al menos 2 PC en R. Sin embargo, no cualquier combinación de tres y dos PC logró la eficacia mencionada; en Octave el único clasificador que lo logró es el que se generó usando los PC números 1, 2 y 5 mientras que en R fue usando los PC números 1 y 2, o bien, estos dos PC en combinación con el PC 5, 6, 8, o 10.

También se observa en la Tabla 1, que en el esquema PCA+LR, la eficacia disminuye a 98.8% cuando se utilizan menos de 3 PC en el clasificador de Octave y menos de 2 PC en el clasificador de R. Además, las únicas opciones de clasificadores que consiguieron esta eficacia son aquellas que utilizan el primer PC. Cualquier clasificador que se puede generar usando uno o dos PC pero que no considera el primero de ellos, obtuvo una eficacia menor. Este mismo valor de eficacia es el máximo obtenido al aplicar el algoritmo de K vecinos más cercanos ponderados (WKNN). Esta eficacia significa que, solamente en una ocasión, de todas las veces que se verificó el clasificador, el resultado fue diferente al diagnóstico histopatológico.

ISSN 2523-6857
ECORFAN® Todos los derechos reservados

Al respecto de los parámetros específicos de los algoritmos, vale agregar que, con estos datos, el resultado del PCA indica que el primer PC explica un promedio del 80% de la varianza de los espectros. Los primeros dos PC en conjunto explican un promedio del 85.6% y logran la eficacia de 100% en el clasificador implementado en R. Y la combinación de los dos primeros PC con el quinto PC explica en promedio el 87.8% de la varianza y logra la eficacia de 100% en el clasificador implementado en Octave. Se agregan ejemplos de clasificadores en Octave que utilizan estos PC (Gráfico 2).

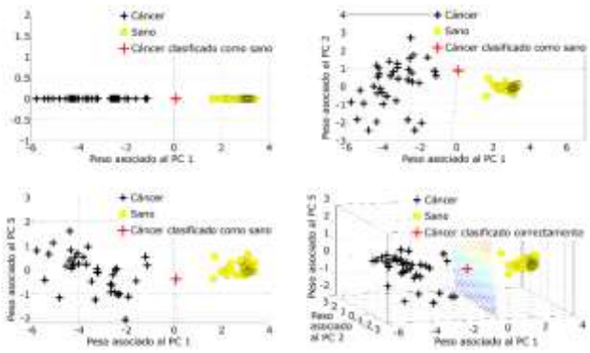

Gráfico 2 Visualización de cuatro clasificadores con el algoritmo PCA+LR dejando fuera el espectro que se clasifica equivocadamente con menos de tres dimensiones y correctamente a partir de tres dimensiones

Recordemos también, que el algoritmo WKNN, que se utilizó, busca la mejor opción para el valor de *k*. Los resultados fueron que en el 98% de los casos verificados, el valor óptimo fue *k*=6 y, en el otro 2%, fue *k*=1. Dada esta diferencia, y para validar la creación de un solo sistema clasificador, se evaluaron ambas opciones con el total de espectros. Se encontró que el mismo espectro que fue clasificado equivocadamente con el algoritmo PCA+LR, también fue clasificado erróneamente en ambos casos del WKNN. Y además, otro espectro fue clasificado equivocadamente en el caso de *k*=1. Por lo tanto, los valores de eficacia fueron 98.8% y 97.6% y de sensibilidad fueron 97.4% y 94.7% ambos para *k*=6 y *k*=1 respectivamente.

3.3 Análisis de los errores de clasificación

Para los clasificadores en que no se obtuvo un 100% de eficacia, los espectros que fueron clasificados equivocadamente corresponden a casos de cáncer.

Es decir, que se obtuvieron uno y dos falsos negativos respectivamente comparados contra el diagnóstico histopatológico. Incluso, se verificó que uno de estos espectros fue el mismo para ambos algoritmos. Entonces, cuando solamente uno de los casos de cáncer fue clasificado como sano, la prueba de diagnóstico con este algoritmo presenta una sensibilidad de 97.4% y cuando fueron dos casos de cáncer los clasificados como sanos, se tiene una sensibilidad de 94.7%. En ambos casos la especificidad es de 100% porque no hay errores en la clasificación de los espectros de tejido sano. Los espectros clasificados erróneamente se muestran comparados contra los demás espectros (Gráfico 3).

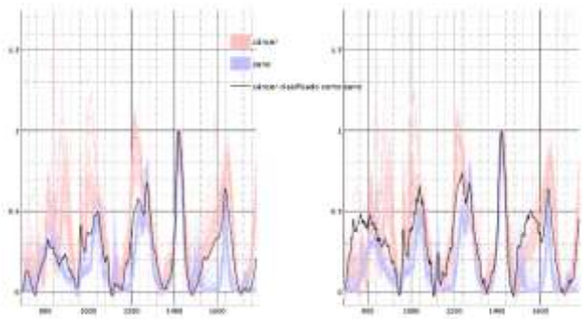

Gráfico 3 Comparación gráfica de espectros individuales contra el resto de espectros. (Izquierda) Espectro clasificado erróneamente por los clasificadores basados en WKNN y en PCA+LR. (Derecha) Espectro clasificado erróneamente por el clasificador basado en WKNN con *k*=1

Una razón por la que, en muchos casos, se prefiere realizar reducción dimensional antes de modelar clasificadores es que el clasificador puede visualizarse gráficamente en una, dos o hasta tres dimensiones. Como ejemplo se muestran los espectros y los umbrales de decisión para cuatro clasificadores generados con algoritmo PCA+LR en Octave (Gráfico 2). Todos ellos corresponden al caso en que se deja fuera el espectro para el que se detectó error en la clasificación usando menos de tres PC. Lo que se varía entre ellos son los PC específicos que se utilizan para generar el clasificador: sólo el PC 1 (arriba-izquierda), PC 1 y 2 (arriba-derecha), PC 1 y 5 (abajo-izquierda) y PC 1, 2 y 5 (abajo-derecha). Por el contrario, los otros modelos de clasificación que se utilizaron en este estudio no usan reducción dimensional y por lo tanto resulta imposible realizar gráficas como éstas porque el número de variables predictoras lo impide.

ISSN 2523-6857
ECORFAN® Todos los derechos reservados

4. Conclusiones

Este trabajo presenta una evaluación del algoritmo de regresión logística (LR) en la implementación de clasificadores basados en espectros Raman de tejido mamario para diagnóstico de cáncer. La característica distintiva de esta evaluación es que el modelo logístico se generó utilizando cada valor de intensidad registrado en los espectros Raman, en lugar de realizar una reducción dimensional o selección de características antes de generar el modelo. No se tiene conocimiento de otro estudio con este tipo de implementación para una aplicación similar.

Según los resultados obtenidos, la implementación del clasificador mediante LR sobre todos los valores de intensidad presentó el mismo nivel de desempeño al compararse con una implementación muy comúnmente empleada que es utilizando reducción dimensional mediante análisis de componentes principales (PCA) y con una implementación del algoritmo Bayes ingenuo que ya había sido recomendada (Martínez Romo et al., 2015) para este tipo de aplicaciones, pero en este caso realizada sin selección de características distintivas. Esto sugiere que, hablando del desempeño del clasificador, la reducción dimensional o selección de características previo al proceso de entrenamiento, no presenta beneficio cuando se utiliza un modelo clasificador por LR ni basado en el algoritmo NB. Además, se debe considerar, que en el caso específico del uso de PCA, la implementación del clasificador perdería simplicidad por requerir el cálculo de pesos de los componentes principales antes de poder realizar un diagnóstico y, por consiguiente, requeriría mayor tiempo de procesamiento computacional.

También se aprecia en los resultados obtenidos, que el clasificador basado en LR fue ligeramente mejor en eficacia comparado con los clasificadores basados en el algoritmo de k-vecinos cercanos ponderados (WKNN) al evaluarse utilizando todos los valores de intensidad de los espectros. Esto sugiere que el algoritmo de LR podría tener mayor facilidad para establecer límites certeros entre los espectros que provienen de tejido con cáncer y aquéllos que provienen de tejido sano comparado contra estos algoritmos clasificadores.

El hecho de que los clasificadores basados en LR hayan podido lograr el 100% de eficacia, resalta este algoritmo como una opción viable y muy prometedora para ser implementado en sistemas para diagnóstico de cáncer de mama sobre espectros Raman, mostrando un mejor desempeño que otros clasificadores.

Agradecimientos

Los autores agradecen a CONACYT, al TecNM / Instituto Tecnológico de Aguascalientes, a la Universidad Autónoma de Aguascalientes, al Centro de Investigaciones en Óptica, y al Centro Universitario de los Lagos de la Universidad de Guadalajara por su apoyo en esta investigación. Israel De la Parra González fue financiado por una beca del Programa Nacional de Posgrados de Calidad (PNPC) del Consejo Nacional de Ciencia y Tecnología (CONACYT) y gozó de licencia académica para estudios de posgrado por parte de la Universidad Autónoma de Aguascalientes.

Referencias

Bi, X. H., Rexer, B., Arteaga, C. L., Guo, M. S., & Mahadevan-Jansen, A. (2014). Evaluating HER2 amplification status and acquired drug resistance in breast cancer cells using Raman spectroscopy. *Journal of Biomedical Optics*, *19*(2), 6. https://doi.org/10.1117/1.jbo.19.2.025001

Bocklitz, T. W., Guo, S., Ryabchykov, O., Vogler, N., & Popp, J. (2016). Raman Based Molecular Imaging and Analytics: A Magic Bullet for Biomedical Applications!? *Analytical Chemistry*, Vol. 88. https://doi.org/10.1021/acs.analchem.5b04665

Choo-Smith, L. P., Edwards, H. G. M., Endtz, H. P., Kros, J. M., Heule, F., Barr, H., … Puppels, G. J. (2002). Medical applications of Raman spectroscopy: From proof of principle to clinical implementation. *Biopolymers - Biospectroscopy Section*. https://doi.org/10.1002/bip.10064

ECORFAN® Todos los derechos reservados

Dingari, N. C., Barman, I., Saha, A., Mcgee, S., Galindo, L. H., Liu, W., … Fitzmaurice, M. (2013). Development and comparative assessment of Raman spectroscopic classification algorithms for lesion discrimination in stereotactic breast biopsies with microcalcifications. *Journal of Biophotonics*, *6*(4), 371–381. https://doi.org/10.1002/jbio.201200098

Dougherty, G. (2013). Pattern recognition and classification: An introduction. In *Pattern Recognition and Classification: An Introduction*. https://doi.org/10.1007/978-1-4614-5323-9

Eaton, J. W., Bateman, D., Hauberg, S., & Wehbring, R. (2015). *GNU Octave version 4.0.0 manual: a high-level interactive language for numerical computations*. Retrieved from http://www.gnu.org/software/octave/doc/interpreter/

Fallahzadeh, O., Dehghani-Bidgoli, Z., & Assarian, M. (2018). Raman spectral feature selection using ant colony optimization for breast cancer diagnosis. *Lasers in Medical Science*, *33*(8), 1799–1806. https://doi.org/10.1007/s10103-018-2544-3

Gautam, R., Vanga, S., Ariese, F., & Umapathy, S. (2015). Review of multidimensional data processing approaches for Raman and infrared spectroscopy. *EPJ Techniques and Instrumentation*. https://doi.org/10.1140/epjti/s40485-015-0018-6

González-Solís, J. L., Guizar-Ruiz, J. I., Martínez-Espinosa, J. C., Martínez-Zerega, B. E., Juárez-López, H. A., Vargas-Rodríguez, H., … Palomares-Anda, P. (2016). Cancer detection based on Raman spectra super-paramagnetic clustering. *Physica A*, *455*, 52–64. https://doi.org/10.1016/j.physa.2016.02.060

Haka, A. S., Shafer-Peltier, K. E., Fitzmaurice, M., Crowe, J., Dasari, R. R., & Feld, M. S. (2002). Identifying Microcalcifications in Benign and Malignant Breast Lesions by Probing Differences in Their Chemical Composition Using Raman Spectroscopy. *Cancer Research*, *62*(18), 5375. Retrieved from http://cancerres.aacrjournals.org/content/62/18/5375.abstract

Haka, A. S., Shafer-Peltier, K. E., Fitzmaurice, M., Crowe, J., Dasari, R. R., & Feld, M. S. (2005). Diagnosing breast cancer by using Raman spectroscopy. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(35), 12371–12376. https://doi.org/10.1073/pnas.0501390102

Hechenbichler, K., & Schliep, K. P. (2004). Weighted k-Nearest-Neighbor Techniques and Ordinal Classification. *SFB 386*, *399*.

Jermyn, M., Desroches, J., Aubertin, K., St-Arnaud, K., Madore, W. J., De Montigny, E., … Leblond, F. (2016). A review of Raman spectroscopy advances with an emphasis on clinical translation challenges in oncology. *Physics in Medicine and Biology*, Vol. 61, pp. R370–R400. https://doi.org/10.1088/0031-9155/61/23/R370

Kim, S., Lee, S. H., Min, S. Y., Byun, K. M., & Lee, S. Y. (n.d.). Dual-modal cancer detection based on optical pH sensing and Raman spectroscopy. *Journal of Biomedical Optics*, *22*(10), 1–6. Retrieved from https://doi.org/10.1117/1.JBO.22.10.105002

Kleinbaum, D. G., & Klein, M. (2010). Logistic regression: A self-learning text. In *Statistics for Biology and Health*. https://doi.org/10.1007/978-1-4419-1742-3

Kong, K., Zaabar, F., Rakha, E., Ellis, I., Koloydenko, A., & Notingher, I. (2014). Towards intra-operative diagnosis of tumours during breast conserving surgery by selective-sampling Raman micro-spectroscopy. *Physics in Medicine and Biology*, *59*(20), 6141–6152. https://doi.org/10.1088/0031-9155/59/20/6141

Krishnamoorthy, C., Prakasarao, A., Srinivasan, V., Sivarama, S. P., & Singaravelu, G. (2019). Monitoring of breast cancer patients under pre and post treated conditions using Raman spectroscopic analysis of blood plasma. *Vibrational Spectroscopy*. https://doi.org/10.1016/j.vibspec.2019.102982

Li, Q. B., Wang, W., Liu, C. H., & Zhang, G. J. (2015). Discrimination of Breast Cancer from Normal Tissue with Raman Spectroscopy and Chemometrics. *Journal of Applied Spectroscopy*, *82*(3), 450–455. https://doi.org/10.1007/s10812-015-0128-6

ISSN 2523-6857
ECORFAN® Todos los derechos reservados

Li, Q., Gao, Q., & Zhang, G. (2014). Classification for breast cancer diagnosis with Raman spectroscopy. *Biomedical Optics Express*, *5*(7), 2435–2445. https://doi.org/10.1364/boe.5.002435

Li, Q., Hao, C., & Xu, Z. (2017). Diagnosis of Breast Cancer Tissues Using 785 nm Miniature Raman Spectrometer and Pattern Regression. *Sensors*, *17*(3), 627. https://doi.org/10.3390/s17030627

Li, Y. Z., Pan, J. J., Chen, G. N., Li, C., Lin, S. J., Shao, Y. H., … Chen, R. (2013). Micro-Raman spectroscopy study of cancerous and normal nasopharyngeal tissues. *Journal of Biomedical Optics*, *18*(2), 6. https://doi.org/10.1117/1.jbo.18.2.027003

Luo, S. W., Chen, C. S., Mao, H., & Jin, S. Q. (2013). Discrimination of premalignant lesions and cancer tissues from normal gastric tissues using Raman spectroscopy. *Journal of Biomedical Optics*, *18*(6), 8. https://doi.org/10.1117/1.jbo.18.6.067004

Martínez Romo, J. C., Luna-Rosas, F. J., Mendoza-González, R., Padilla-Díaz, A., Mora-González, M., & Martínez-Cano, E. (2015). Improving sensitivity and specificity in breast cancer detection using raman spectroscopy and bayesian classification. *Spectroscopy Letters*. https://doi.org/10.1080/00387010.2013.855640

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2017). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. https://doi.org/citeulike-article-id:9958545

Ng, A. Y. (n.d.). Free online machine learning course. Retrieved October 7, 2016, from www.ml-class.org

Nielsen, H. B., & Mortensen, S. B. (2016). *ucminf: General-Purpose Unconstrained Non-Linear Optimization*. Retrieved from https://cran.r-project.org/package=ucminf

Pence, I. J., Patil, C. A., Lieber, C. A., & Mahadevan-Jansen, A. (2015). Discrimination of liver malignancies with 1064 nm dispersive Raman spectroscopy. *Biomedical Optics Express*. https://doi.org/10.1364/boe.6.002724

R Core Team. (2017). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*.

Ralbovsky, N. M., & Lednev, I. K. (2019). Raman spectroscopy and chemometrics: A potential universal method for diagnosing cancer. *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*. https://doi.org/10.1016/j.saa.2019.04.067

Rodriguez, J. M., Godoy, D., Mateos, C., & Zunino, A. (2017). A multi-core computing approach for large-scale multi-label classification. *Intelligent Data Analysis*, *21*(2), 329–352. https://doi.org/10.3233/ida-150375

Samworth, R. J. (2012). Optimal weighted nearest neighbour classifiers. *The Annals of Statistics*, *40*(5), 2733–2763. https://doi.org/10.1214/12-AOS1049

Sathyavathi, R., Saha, A., Soares, J. S., Spegazzini, N., McGee, S., Rao Dasari, R., … Barman, I. (2015). Raman spectroscopic sensing of carbonate intercalation in breast microcalcifications at stereotactic biopsy. *Scientific Reports*, *5*. https://doi.org/10.1038/srep09907

Savitzky, A., & Golay, M. J. E. (1964). Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*. https://doi.org/10.1021/ac60214a047

Sayad, S. (n.d.). *Logistic Regression* (Vol. 2018). Vol. 2018. Retrieved from http://www.saedsayad.com/logistic_regression.htm

Schliep, K., & Hechenbichler, K. (2016). *kknn: Weighted k-Nearest Neighbors*. Retrieved from https://cran.r-project.org/package=kknn

Vanna, R., Morasso, C., Marcinnò, B., Piccotti, F., Torti, E., Altamura, D., … Corsi, F. (2020). Raman Spectroscopy reveals that biochemical composition of breast microcalcifications correlates with histopathological features. *Cancer Research*. https://doi.org/10.1158/0008-5472.CAN-19-3204

ISSN 2523-6857
ECORFAN® Todos los derechos reservados

Zhao, J., Lui, H., I., D., & Zeng, H. (2010). Real-Time Raman Spectroscopy for Noninvasive in vivo Skin Analysis and Diagnosis. In *New Developments in Biomedical Engineering*. https://doi.org/10.5772/7603

ISSN 2523-6857
ECORFAN® Todos los derechos reservados