CLASIFICACIÓN AUTOMÁTICA DE DENSIDAD MAMARIA

Guillermina Griffa[†], Micaela Bertero[†] y Valeria Rulloni[†]

[†]*Facultad de Ciencias Exactas, Físicas y Naturales, Universidad Nacional de Córdoba, Av. Vélez Sarsfield 1611, Córdoba, Argentina,vrulloni@unc.edu.ar*

Resumen: La mamografía es actualmente una herramienta de diagnóstico muy útil para la detección de lesiones mamarias. Sin embargo, la sensibilidad de éstas disminuye conforme aumenta la densidad del tejido. Buscando desarrollar una herramienta informática que determine dicha densidad de manera automática, se estudiaron y desarrollaron distintos descriptores basados en análisis de imágenes, tales como histograma, entropía de permutaciones (EP), dimensión fractal (DF), entropía de bloques del gradiente binarizado (EBGB), etc. Éstos constituyen la entrada de los algoritmos de clasificación automática utilizados: Support Vector Machine (SVM), k-vecinos más cercanos (k-NN), Random Forest y Regresión logística, entre otros. El mejor resultado fue obtenido con el clasificador SVM, aunque todos se encuentran dentro del mismo rango. A su vez, mediante un análisis de componentes principales, se llega a la conclusión de que un valor similar puede obtenerse reduciendo la cantidad de características, lo que lleva a una reducción de los tiempos de cómputo.

Palabras clave: *clasificación automática, densidad mamaria, textura, mamografía digital.*
2000 AMS Subject Classification: 68T10 - 62H30 - 62H35

1. INTRODUCCIÓN

Los organismos de salud recomiendan realizarse una mamografía de exploración cada año a las mujeres, comenzando a partir de los 40 años. Este estudio juega un papel central en la detección temprana del cáncer de mama. La densidad de la mama se refiere a la cantidad de tejido parenquimatoso y conectivo que, por ser radiopaco, aparece de color blanco al igual que el tejido canceroso en la mamografía. Es por ello que los tumores pueden ser difíciles de percibir entre los tejidos densos, disminuyendo la sensibilidad de la mamografía a medida que la densidad aumenta. Los métodos iniciales para evaluar la densidad mamográfica fueron completamente subjetivos y cualitativos; sin embargo, en los últimos años se han desarrollado métodos para proporcionar mediciones más objetivas, y con este trabajo se pretende aportar a este desarrollo.

Los radiólogos utilizan el sistema BI-RADS del Colegio Americano de Radiología (ACR por sus siglas en inglés, American College of Radiology)[10] para clasificar la densidad mamaria en cuatro categorías:

- *a:* mamas compuestas por tejido adiposo casi en su totalidad. La mamografía es muy sensible en este contexto, siempre que se incluya en el campo de la imagen el sector que contiene la anormalidad.
- *b:* se observan sectores dispersos de densidad fibroglandular.
- *c:* mamas que presentan densidad heterogénea. En estas resulta de utilidad describir la ubicación del tejido más denso pues puede ocultar algunos nódulos pequeños o pequeñas lesiones no calcificadas.
- *d:* mamas muy densas. Su sensibilidad mamográfica es la más baja.

Cabe destacar que en la práctica dicha clasificación es subjetiva, según la experiencia y criterio personal del profesional. Con el objetivo de desarrollar un algoritmo que sea capaz de realizar esta clasificación de manera automática y que reduzca las subjetividades, el presente trabajo se dividió en cuatro etapas: construcción de una base de datos, segmentación automática de las imágenes para obtener la región de interés, extracción de características de dicha región e implementación de algoritmos inteligentes de clasificación. Éstas se describen a lo largo del trabajo.

2. ESTADO DEL ARTE

Hay diversos antecedentes en clasificación automática de la densidad mamaria. En [1] se buscó la cuantificación de la densidad mamaria en 168 mamografías mediante EP y Red Neuronal Multicapa. Obtuvieron

un error de test de 24,42 % utilizando Validación cruzada Hold-out. En [9] utilizaron 100 mamografías (Oxford Database), 25 por cada clase y "texton spatial dependence matrix (TSDM)", logrando 86 %, 93 %, 80 % y 66 % de presición para clasificar en a, b, c y d respectivamente. En [3] consideraron 60 imágenes revisadas por dos radiólogos, utilizaron análisis fractal, análisis de histograma y un clasificador bayesiano, donde los coeficientes de correlación quedaron 73 para Regional Skewness, 74 para Fractal y 84 para Skewness/Fractal, para clasificación automática versus la del radiólogo (R1). En [11], para análisis de la densidad de mama asistido por ordenador, utilizan las bases DDSM y MIAS, análisis de histograma y k-NN. Obtuvieron errores entre 36.65 % y 21.74 %, el menor del método propuesto. En [6] realizaron un análisis de histograma local y utilizaron k-NN para 615 imágenes. La tasa de errores menores (un solo grado de densidad) estuvo entre 0.36 y 0.45 y la de los mayores entre 0.018 y 0.127. En [7], utilizando la base MIAS compararon a los descriptores: Matriz de co-ocurrencia de los niveles de gris, máscaras de Laws y patrones locales binarios usando k-NN, discriminante de Fisher, análisis discriminante lineal y SVM. Los porcentajes obtenidos más altos de clasificación correcta con y sin selección de características fueron 79 % y 68 % respectivamente.

3. Materiales y Métodos

3.1. Materiales

Para este trabajo se recolectaron 6613 mamografías digitales de alta resolución, correspondientes a 1576 estudios diferentes, debidamente anonimizados, cortesía de la Fundación Carlos Oulton (Córdoba, Argentina). En general cada estudio mamográfico cuenta con 4 mamografías diferenciadas por siglas según la proyección y lateralidad en inglés: Craneo Caudal (CC) y Oblicua Medio Lateral (MLO), agregando al principio de la sigla R o L según sea de la mama derecha o izquierda respectivamente. Las mamografías fueron adquiridas con tres diferentes mamógrafos, todos ellos Hologic Selenia Dimensions. Las mismas estaban clasificadas por el equipo de especialistas médicos del Instituto, contando con 473 tipo a, 2703 tipo b, 2797 tipo c y 423 tipo d, de las cuales 683 poseían implantes mamarios. Para los cálculos se utilizaron dos computadoras portátiles: Dell Inspiron M531R-5535, con Windows 8.1 de 64-bit, procesador AMD A-10-5745M a 2.10 GHz y 8Gb de RAM y Dell Inspiron 3579, con Windows 10 de 64-bit, procesador Intel Core i7-8750H, 32Gb de RAM y placa de video dedicada Nvidia Geforce GTX 1050Ti. Como lenguaje base de programación se utilizó Python (3.6.5) desde la distribución de Anaconda 5.2.0 64-bit. Como visualizador de imágenes se utilizó Radiant DICOM Viewer 64-bit. El procesado digital realizado en cada mamografía cuenta con las etapas de segmentación, extracción de características y luego clasificación.

3.2. Segmentación

El objeto o región de interés (ROI) es la mama propiamente dicha. Por eso en esta etapa se eliminó, de cada mamografía, las regiones no relevantes: etiqueta, fondo, implante mamario, músculo pectoral y piel. Para eliminar la etiqueta superior que indica la proyección y lateralidad se utilizó: binarizado, etiquetado y multiplicación. Sólo para las imágenes con implante mamario se realizó la sustracción del mismo utilizando: crecimiento de regiones, imagen negativa, dilatación y multiplicación. Para las imágenes MLO se requiere la extracción del músculo pectoral. Ha resultado la extracción más compleja debido a que se parte de la hipótesis de que el músculo es la región de mayor intensidad de la imagen. Para esto se propone como solución la aplicación de técnicas de umbralizado y ajuste de curva. Pero existen excepciones en donde esta región es la menos intensa. A su vez en los casos positivos no siempre el umbralizado y ajuste devuelven un resultado óptimo. Por esto se utilizó un algoritmo automático adaptable: Operador gradiente para detectar bordes, Suma ponderada para aumentar el contraste, una modificación de la Umbralización de Ridler Calvard [14] para ubicar los puntos generadores en el crecimiento de regiones, ajuste de curva de segundo grado o recta según corresponda, contorno, multiplicación, dilatación y erosión. La modificación de la Umbralización de Ridler Calvard consiste en aplicar el algoritmo sólo a los pixeles mayores que el promedio.

3.3. Extracción de Características

Una vez definida la ROI, se extrae información relevante de cada mamografía, útil para la clasificación. Las características calculadas se agrupan según su tipo (ver detalles en [2]):

- *Cantidad de píxeles de la ROI.* Dado que en general las mamas de mayor tamaño presentan una proporción de tejido graso superior.
- *Análisis de histograma global.* Del histograma global (con niveles de gris de 0 a 65536) se consideró: media, mediana, desviación estándar, asimetría, curtosis y moda.
- *Análisis de histograma local.* Basado en [6] se consideraron la desviación estándar y la asimetría de los histogramas locales separados en cinco partes según la distancia a la línea de la piel.
- *Entropía de permutaciones.* Inspiradas en [1], a celdas de 100×100 píxeles, se calculó el promedio de las entropías de permutaciones de bloques 3×2 con retardo de solapado de 1.
- *Entropía de bloques del gradiente binarizado.* A partir del binarizado de la imagen resultado del filtro de gradiente se calcula la entropía del histograma de las 512 (2^9) configuraciones binarias 3×3.
- *Análisis fractal.* Con el algoritmo multinivel de Otsu, se calculan cuatro umbrales y con éstos ocho máscaras a las que se le calcula la DF. Considerando los valores de gris en las máscaras y en los bordes de éstas, se calcularon: promedio, desviación estándar, curtosis, asimetría y entropía, ver[4].
- *Descriptores clásicos de Haralick.* Se utilizó un comando de la librería mahotas de Python: *mahotas.features.haralick()*para calcular 13 características de Haralick [5] en cada dirección.
- *Características extraídas de información DICOM.* De la información contenida en los archivos DICOM se extrajeron las características[16]: edad, lateralidad, Kilo voltaje pico de salida, entre otras.

3.4. Clasificación

Considerando la tabla de datos de tamaño 6613×198, donde cada fila corresponde a una imagen y cada columna a una característica, se normalizaron las características numéricas y se codificaron las categóricas para poder ajustar los modelos. Se dividieron los datos de manera aleatoria en conjunto de entrenamiento (80 % de los ejemplos) y conjunto de evaluación (20 % restante) para calcular las métricas de rendimiento. Se evaluaron cuatro diferentes algoritmos inteligentes de clasificación: Casificador de Vectores de Soporte (SVC de la familia SVM [12]), Regresión Logística [13], Random Forest [15] y k-NN [8]. A cada uno de éstos se le realizó una optimización de los hiperparámetros mediante validación cruzada.

4. Resultados

La Tabla 4.1 resume los resultados de cada clasificador utilizando la base completa. Ha de notarse que el mejor promedio de aciertos (accurancy) se obtuvo con SVC. Analizando la Tabla 4.2, que resume los resultados del SVC para cada clase por separado, se puede advertir que el valor de recall es mucho menor para las clases a y d, es decir que la proporción de clasificación incorrecta es mayor en esos casos. Esto puede deberse a la diferencia en la cantidad de muestras de las clases. Se probaron técnicas que abordan esa problemática pero no se obtuvieron mejores resultados, por lo que se propone a futuro agrandar la base y balancearla.

Tabla 4.1: Comparación de los resultados obtenidos con los diferentes algoritmos.

	Score	SVC	LR	RF	k-NN
0	Test Acurancy	0.764	0.753	0.746	0.720
1	Mean Cross Validation	0.754	0.747	0.722	0.717

Tabla 4.2: Comparación de los resultados obtenidos de SVC según la clase.

	precision	recall	F1-score	support
a	0.73	0.45	0.56	95
b	0.73	0.83	0.78	518
c	0.81	0.79	0.80	570
d	0.74	0.60	0.66	97

A modo exploratorio se clasificó con SVC y k-NN considerando conjuntos de características por vez. De los resultados obtenidos se puede destacar que tanto las características extraídas del histograma como las relacionadas con la dimensión fractal, son útiles para una buena clasificación de los datos, obteniendo un promedio de aciertos de 0.72 y 0.71 respectivamente. Si bien estos valores no varían significativamente respecto a la evaluación con todas las características, hay una disminución en la generalización de la clasificación pues el error aumenta para las clases a y d.

También se realizó una reducción de dimensionalidad del espacio de características utilizando el algoritmo de ACP (Análisis de Componentes Principales) con el cual se obtuvieron 30 componentes ortogonales que explicaron el 95 % de la variabilidad de los datos, reduciendo los tiempos de cómputo.

REFERENCIAS

[1] A. ANTONELLI, G. J. MESCHINO, AND V. L. BALLARIN, *Cuantificación de la densidad mamaria mediante Entropía de Permutación en mamografías*, SABI , 2017.

[2] M BERTERO AND G. GRIFFA, *Clasificación automática de la densidad mamaria*, Proyecto Integrador de Ingeniería Biomédica de la Universidad Nacional de Córdoba (Argentina) , 2018

[3] J. W. BYNG, N. F. BOYD, E. FISHELL, R. A. JONGK AND M. J. YAFFE , *Automated analysis of mammographic densities* , 2013.

[4] A. F. COSTA AND G. HUMPIRE-MAMANI AND A. J. M. TRAINA, *An efficient algorithm for fractal analysis of textures*, Graphics, Patterns and Images, 2012 25th SIBGRAPI Conference on, IEEE, 39–46, 2012.

[5] R. M. HARALICK AND K. SHANMUGAM AND I. DINSTEIN,*Textural features for image classification*, IEEE Transactions on systems, man, and cybernetics, 3, 6, 610–621, 1973.

[6] N. KARSSEMEIJER, *Automated classification of parenchymal patterns in mammograms*, University Hospital Nijmegen, Department of Radiology, PO Box 9101, 6500 HB Nijmegen,The Netherlands, 7, 14, 355–377, 1998.

[7] C. MATA, J. FREIXENET, X. LLADÓ AND A. OLIVER,*Texture Descriptors applied to Digital Mammography*,University of Western Australia, 2009.

[8] A. MOUJAHID, I. INZA Y P. LARRAÑAGA, *Clasificadores K-NN*, Departamento de Ciencias de la Computación e Inteligencia Artificial Universidad del País Vasco–Euskal Herriko Unibertsitatea, 2010.

[9] S. PETROUDI AND M. BRADY , *Breast Density Characterization using Texton Distributions* , 33rd Annual International Conference of the IEEE EMBS Boston, Massachusetts USA , 2011.

[10] E.A. SICKLES , C.J. D'ORSI AND L.W. BASSETT,*BI-RADS del ACR: Mamografía*, Reston, VA, Colegio Estadounidense de Radiología, 2013.

[11] J. A. SOLVES LLORENS, *Análisis de la densidad de mama asistido por ordenador*,Trabajo de investigación del Master de Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia, 2012.

[12] E. J. C. SUÁREZ, *Tutorial sobre máquinas de vectores soporte (SVM)*, 2014.

[13] *Introducción a la regresión logística* , BIOESTADISTICA (55 - 10536) , Departamento de Estadística Universidad Carlos III de Madrid , 2010.

[14] *Método Ridler Calvard y Otsu*, Última vez visitada Octubre de 2018. https://prezi.com/povqek3wduzr/metodo-ridler-calvar-y-metodo-otsu/

[15] *Random Forest Simple Explanation*, Última vez visitada Octubre de 2018. https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d .

[16] *Innolitics*, 2018. https://dicom.innolitics.com/ciods/digital-mammography-x-ray-image