Departamento de Matemáticas, Facultad de Ciencias
Universidad Autónoma de Madrid

Radiómica: las imágenes médicas como datos

Trabajo de fin de grado

Grado en Matemáticas

Autor: Carlos Arias Alcaide

Tutor: Amparo Baíllo Moreno
Davide Barbieri

Curso 2017-2018

Resumen

La radiómica, una nueva rama de investigación en la oncología, consiste en convertir imágenes médicas en una gran cantidad de valores numéricos y extraer información de ellos mediante técnicas estadísticas. Aunque el proceso radiómico es sencillo, existen múltiples factores más o menos controlables que hacen que una pequeña variación en alguno de los pasos, cambie los resultados del estudio radicalmente. Este trabajo se centra en analizar y reproducir, en un conjunto de 86 escáneres de tumores primarios de pulmón, los dos últimos pasos del proceso: extracción de características radiómicas (las variables estadísticas de interés en este campo) y análisis estadístico de las mismas. En la extracción de características se analiza el concepto que modelizan y el algoritmo de cada una de las 42 características obtenidas mediante IBEX, un programa informático específico para radiómica. En el análisis estadístico, se usa como técnica estadística fundamental el análisis de conglomerados tanto para la reducción de la dimensión de los datos como para la agrupación de los casos de pacientes. Se comprueba finalmente que, pese a realizarse el proceso radiómico en un conjunto muy reducido de datos, las relaciones entre los diferentes casos aparecen por sí solas, creándose grupos fácilmente justificables.

Abstract

Radiomics, a new brand of research in oncology, involves converting medical images into a large number of numeric values and extracting information from them using statistical techniques. Although the radiomic process is simple, there are considerably controllable factors that lead to huge variations in the results after just a small change in one of the steps. This work is focused on analysing and reproducing, in a set of 86 scanners of primary lung tumours, the last two steps of the radiomic process: the extraction of the radiomic features (the statistical variables of interest in this field) and the statistical analysis of them. In the extraction of the features, the concept to be modelled and the algorithm of each of the 42 features obtained through IBEX, a specific computer program for radiomics, are analysed. In the statistical analysis, the conglomerate analysis is used as a fundamental statistical technique both for the data dimensionality reduction and for the grouping of the patients cases. Finally, it is proved that, despite the radiomic process being performed on a very small set of data, the relationships among the different cases appear on their own, creating groups easily justifiable.

Índice general

CAPÍTULO 1
Introducción

La radiómica es una nueva rama de la oncología que surge de la unión de dos campos antes separados: las técnicas médicas de extracción de imágenes, usadas tradicionalmente en la detección y diagnóstico, y el análisis estadístico. Esta nueva rama consiste en la conversión de imágenes médicas en una gran cantidad de características numéricas y la posterior extracción de información de ellas mediante diversos procedimientos estadísticos. El objetivo final de la radiómica es el desarrollo de herramientas con mayor precisión en el diagnóstico y tratamiento.

El proceso seguido en radiómica consta de una serie de pasos:

1. **Adquisición de imágenes médicas.** Se somete al paciente a un conjunto de técnicas, como son la resonancia magnética o el escáner, para obtener las imágenes (secciones 2D).

2. **Delimitación de las regiones de interés.** En cada una de las imágenes adquiridas, el especialista radiólogo marca la zona concreta a estudiar, delimitando así cada una de las regiones de interés (ROI). En el caso de la oncología, las ROI son los tumores y, en algunas ocasiones, también se añade su alrededor.

3. **Reconstrucción del volumen de interés.** Empleando métodos matemáticos de interpolación, se crea una reconstrucción 3D a partir del conjunto de todas las ROI. Esta reconstrucción se denomina volumen de interés (VOI).

4. **Extracción de características y almacenamiento.** Usándose programas informáticos exclusivos para radiómica, se extraen las características radiómicas a partir de las ROI y el VOI obtenidos en cada paciente. Otra información como el historial clínico también puede incorporarse. Todo ello se almacena de forma ordenada en una base de datos junto con lo obtenido de otros pacientes.

5. **Análisis estadístico de las características.** La información de la base de datos se analiza mediante técnicas estadísticas multivariantes como clasificación no supervisada, clasificación supervisada, regresión multiple, etc. con el objetivo de encontrar relaciones entre los distintos pacientes.

Aunque conceptualmente el proceso es sencillo, existe una gran cantidad de factores más o menos controlables a tener en cuenta, ya que pueden afectar a los re-

sultados finales del análisis. Por ejemplo, la extracción de imágenes utilizando distintos aparatos y ajustes, la delimitación de las ROI por distintos especialistas, etc. En [Aerts *et al.*, 2014] y [Gillies *et al.*, 2016] se pueden obtener más detalles sobre la radiómica, su proceso y los inconvenientes con que se encuentra.

Este Trabajo de Fin de Grado (TFG) consiste en una reproducción a pequeña escala de parte del proceso radiómico (extracción de características y análisis estadístico de las mismas) usando un conjunto de casos reales de tumor. La memoria se estructura en cuatro capítulos. En el primer capítulo se describe qué es la radiómica y el proceso seguido, además de explicar las herramientas empleadas para desarrollar este proyecto. En el segundo, se analizan las características radiómicas, es decir, las variables estadísticas de interés, y sus algoritmos de cálculo. El tercer capítulo expone todo el análisis estadístico de las características obtenidas en nuestro conjunto concreto de casos reales. En el cuarto, se comentan las conclusiones a las que se ha llegado.

1.1. Herramientas empleadas en nuestro estudio

En este proyecto empleamos casos de tumores primarios de pulmón (el tumor se originó en el pulmón y no es consecuencia de una metástasis) cuyas imágenes médicas fueron obtenidas con escáner. El escáner es una técnica no invasiva de extracción de imágenes que usa rayos X y en la que las imágenes se obtienen como un conjunto de secciones tomográficas. Estas secciones tomográficas son imágenes en 2D del interior del cuerpo que se presentan en distintos tonos de gris (los tonos más claros se corresponden con partes del cuerpo con mayor densidad, por ejemplo, huesos; y los tonos más oscuros, con partes menos densas, como los pulmones). Es en estas imágenes 2D donde el experto radiólogo delimita las ROI ayudándose de programas especializados. Además, los equipos modernos son capaces de crear una reconstrucción 3D (el VOI) a partir de las zonas señaladas (las ROI). Matemáticamente, tanto las secciones 2D como las reconstrucciones 3D en tonos de gris se tratan como mallas cuadriculadas donde cada uno de esos cuadrados o cubos (véase [Lorensen y Cline, 1987]) tiene un valor numérico (el tono de gris asociado) y representa una unidad mínima de imagen (pixel o voxel, respectivamente). Con estas mallas en 2D o 3D se calculan las características que describiremos en el siguiente capítulo.

En la extracción de las características, utilizamos el programa radiómico IBEX. Éste es un software libre y abierto creado con el objetivo de facilitar el trabajo colaborativo en la radiómica. Para nuestro estudio, nos hemos ayudado de su manual [Zhang *et al.*, 2015] y sus códigos programados en Matlab.

Finalmente, para el análisis hemos implementado código de programación en el lenguaje R, junto con Excel para la limpieza de la información recibida. Cabe destacar que todos nuestros códigos implementados en R se adjuntan como anexo al final del documento.

CAPÍTULO 2
Análisis de características radiómicas

En este capítulo se analizan las características radiómicas, es decir, las variables estadísticas de interés, que aparecen en nuestra reproducción del proceso radiómico y se explica la algoritmia realizada por el IBEX para su obtención. En [Zwanenburg *et al.*, 2018] se puede encontrar una breve descripción general de ellas junto con otras muchas más que no se emplean en este estudio.

En esta introducción del capítulo 2, hemos extraído las notaciones comunes usadas en las secciones 2.2, 2.3 y 2.4. Como hemos dicho en la sección 1.1, las imagénes en tonos de gris son tratadas como mallas cuadriculadas numéricas, así que, las notaciones extraídas son: N_v para la cantidad de vóxeles en el VOI, $N_p(ROI)$ para la cantidad de píxeles en la ROI de una sección tomográfica concreta y N_g la cantidad de posibles tonos de gris tanto para 3D como para 2D, ya que la cantidad de tonos de gris se fija antes de empezar a trabajar con el VOI o con las ROI. Éstos han sido los únicos valores cuya notación se ha podido unificar. No ha sido posible crear una notación común para indicar el tono de gris de cada píxel, debido a la disparidad de intereses en los cálculos de las características de los distintos grupos. En este caso, con el objetivo de simplificar las explicaciones, se ha decidido usar notaciones similares pero acordes a las necesidades de cada sección. También cabe aclarar que durante todo el texto se usan las expresiones "tono de gris" y "nivel de gris" indistintamente.

Al final del capítulo se anexa la tabla 2.1 con la relación de características empleadas en nuestra reproducción.

2.1. Características morfológicas

Las características morfológicas son aquellas que describen aspectos geométricos del VOI, de ahí que, para su cálculo, no aparezcan en ningún caso los tonos de gris.

Estas características son calculadas directamente en la reconstrucción 3D del VOI y se corresponde con el bloque F1-Shape de nuestro análisis estadístico.

Número de vóxeles

Tomando como referencia la malla cuadriculada 3D, mencionada en 1.1, la característica número de vóxeles se corresponde con la cantidad de cubos que conforman el VOI en la malla.

La definición implementada en IBEX de esta caraterística no es tan simple y directa. IBEX tiene un parámetro llamado EdgeVoxelFraction, que se puede ajustar manualmente, y consiste en un peso aplicado a los vóxeles de los bordes del VOI, de tal manera que no cuenta lo mismo un voxel del borde que uno del interior. Por esta razón, IBEX calcula la característica número de vóxeles como:

$$TotalVoxel = TotalVoxel_{Interior} + EdgeVoxelFraction \cdot TotalVoxel_{Borde}$$

donde, en nuestro caso, EdgeVoxelFraction está fijado como 0,5.

En nuestro análisis estadístico, esta característica recibe la etiqueta F1.1.

Área de la superficie

El área de la superficie expresa la medida de extensión que ocupa la frontera del VOI en el plano 2D.

Nuestro programa radiómico IBEX calcula una estimación del área de la superficie usando las medidas de Minkowski para imágenes binarias en 3D (para saber más sobre estas medidas, véase [Legland *et al.*, 2007]).

El área de la superficie se corresponde con la característica de etiqueta F1.2.

Volumen

La característica volumen se trata de la medida del espacio contenido en el interior de la frontera del VOI.

IBEX calcula el volumen a partir de la aproximación siguiente:

$$Volumen = dim_X(Voxel) \cdot dim_Y(Voxel) \cdot dim_Z(Voxel) \cdot TotalVoxel.$$

En la fórmula anterior se emplean los términos $dim_X(Voxel)$, $dim_Y(Voxel)$ y $dim_Z(Voxel)$, ya que, dos escáneres distintos, no tienen por qué coincidir en el tamaño que sus vóxeles tengan. Esto se debe al ajuste que se hace del aparato en cuanto a cercanía con respecto al paciente o distancia entre sección y sección en el escáner realizado.

La característica F1.3 de nuestro análisis estadístico es el volumen del VOI.

SE HAN OMITIDO LAS PÁGINAS 5-28 CORRESPONDIENTES A LOS **ANÁLISIS DE CARACTERÍSTICAS Y ESTADÍSTICO**

CONTACTAR CON EL AUTOR DEL PROYECTO PARA MÁS INFORMACIÓN

CAPÍTULO 4
Conclusiones

La radiómica es un novedoso campo de investigación en la oncología donde la estadística juega un papel fundamental. Ésta pretende incorporar en el diagnóstico (y, consecuentemente, en el tratamiento) aspectos que en el pasado no eran tan relevantes como el tamaño, la forma, la textura, la rugosidad y otros aspectos no tan directos de los tumores. Con la reproducción del proceso radiómico, hecha a pequeña escala en este estudio, se puede observar que la radiómica es efectiva ya que consigue agrupar de una forma clasificatoria los distintos casos de tumores primarios de pulmón que hemos empleado para dicha reproducción.

En este estudio se ha realizado un análisis profundo del conjunto de características radiómicas establecidas por nuestro conjunto de datos reales. Este análisis aporta intuición acerca de la gran variedad de características que se pueden incorporar en los estudios radiómicos, el origen de éstas, su objetivo a cuantificar y sus algoritmos de cálculo. Se observa cómo realmente cada caso de tumor puede ser caracterizado a partir de dichos valores, cuyos algoritmos son ajustados eficazmente e implementados en programas propios de la radiómica como el IBEX.

Una vez cuantificados con las características los distintos casos reales, se procede a llevar a cabo el análisis estadístico cuyo objetivo es contemplar aspectos que a priori no son observables, como la comparación entre los datos para buscar propiedades en común. El concepto de medida de similaridad (y disimilaridad) y la técnica de análisis de conglomerados han sido cruciales en todo el proceso, apareciendo tanto en la elección de características transformadas relevantes durante el proceso de reducción de la dimensión como en la propia agrupación final de los datos. En las elecciones tomadas durante el estudio estadístico hemos recurrido al sentido común y la capacidad de interpretar los resultados; sin embargo, la supervisión de estas elecciones por parte de un experto oncólogo, especialmente en el análisis de conglomerados final donde se agrupan los pacientes, harían que el estudio se enriqueciera mucho más.

A lo largo de todo este estudio y en relación con la radiómica, hemos comprobado que surgen otras muchas ramas paralelas donde la matemática también tiene importancia como las técnicas de extracción de imágenes médicas, el procesamiento de dichas imágenes, la reconstrucción en 3D de los VOI a partir de las ROI en las secciones 2D, técnicas estadísticas de clasificación supervisada, etc. La aplicación de distintos algoritmos o herramientas de trabajo hace que surjan diferencias que pueden

ser bastante significativas. Como consecuencia, cualquier mejora en alguna de esas ramas involucradas en el proceso, hará que los resultados de los análisis radiómicos sean más fiables.

Anexo I

A continuación se adjuntan los códigos de programación usados. Debido a la extensión de línea que ocupan, se ha tenido que poner en horizontal las páginas, además de escribir en líneas distintas comandos que deberían ir escritos en una única línea.

Para poder usar estos códigos de programación, hay que escribir en la variable directorio, situada en cada uno de los encabezados, la ruta donde hemos situado la carpeta TFG completa (respetando la estructura de carpetas internas), en formato cadena "C:\\Users\\Carlos\\Desktop\\TFG".

La carpeta TFG contiene lo siguiente:

1. La carpeta 'R scripts' donde están todos los scripts que producirán el análisis.
2. La carpeta 'Datos' donde se encuentran los datos y su limpieza.
3. La carpeta 'Carga y transformacion' que es donde se cargarán los archivos de 'Script1.R'.
4. La carpeta 'Agrupaciones de caracteristicas transformadas' donde estarán los archivos que se generan tras cargar el código 'Script2.R'.
5. La carpeta 'Reduccion de dimension' en la que se generarán los archivos de 'Script3.R'.
6. La carpeta 'Agrupaciones de datos' donde se generarán los archivos de 'Script4A.R', 'Script4B.R' y 'Script4C.R'.

```
##############################################################################################
##                                                                                          ##
##                                FUNCIONES Y DATOS AUXILIARES                              ##
##                                                                                          ##
##############################################################################################

# Carga de librerias necesarias
library(e1071)

########------------------------------------------------------------------------########
########                      Variables generales necesarias                      ########
########------------------------------------------------------------------------########

# Lista con transformaciones
transformaciones <- list(Nombre = c('2', 'Originales', '1/2', 'Logaritmo',
                                    '-1/2', '-1', '-2'),
                         Valores = c(2, 1, 1/2, 0, -1/2, -1, -2),
                         Problema.negativos = c('FALSE', 'FALSE', 'TRUE', 'TRUE',
                                                'TRUE', 'FALSE', 'FALSE'),
                         Representacion = c('^2', '␣', '^(1/2)', 'log(', '^(-1/2)',
                                            '^(-1)', '^(-2)'),
                         Color = c('blue', 'green', 'yellow', 'yellow', 'yellow',
                                   'yellow', 'yellow'))

cantidad.transformaciones <- length(transformaciones$Nombre)

# Colores para nubes de puntos
colores.fondo <- rep(c('bisque','darkseagreen2'), 21)
colores.puntos <- rep(c('red','black'), 22)


########------------------------------------------------------------------------########
########                        Funciones personalizadas                          ########
########------------------------------------------------------------------------########

# Devuelve una string con coeficiente de asimetria de Pearson
```

SE HAN OMITIDO LAS PÁGINAS 33-63 CORRESPONDIENTES AL **CÓDIGO PROGRAMADO EN R PARA EL ANÁLISIS ESTADÍSTICO**

CONTACTAR CON EL AUTOR DEL PROYECTO PARA MÁS INFORMACIÓN

```
# solo sale 1. En caso de que por ejemplo salieran 2 pregrupos, se procederia de la si-
# guiente forma.
autovectores <- JordanN$vectors[,1:2]

# 3.) Normalizacion de los 2 primeros autovectores con la norma del maximo
autovectores.normalizados <- matrix(0, numero.datos,2)
autovectores.normalizados[,1] <- autovectores[,1] * 1/norm(as.matrix(autovectores[,1]),
                                                           type = "M")
autovectores.normalizados[,2] <- autovectores[,2] * 1/norm(as.matrix(autovectores[,2]),
                                                           type = "M")

# 4.) Clasificamos los datos
preclusters.esp <- c(NULL)
for (i in 1:numero.datos){
  preclusters.esp <- c( preclusters.esp, which.max(autovectores.normalizados[i,]) )
}
rm(i) # Borrar variables auxiliares

# 5.) Agrupacion de los datos cuyos agrupamientos tienen menos de M datos
M <- 2
clusters.esp <- preclusters.esp
for (i in unique(clusters.esp)){
  if( length( clusters.esp[ clusters.esp == i] ) < M ){
    clusters.esp[ clusters.esp == i] <- unique(clusters.esp)[1]-1
  }
}

rm(i)

# Representacion de pairs con el cluster

pdf('Agrupaciones␣de␣datos/Espectral␣con␣disimilaridad␣gower.pdf')
pairs(datos.reducidos, col = clusters.esp, pch = 16)
dev.off()

print('Se␣realizo␣el␣analisis␣de␣conglomerados␣con␣algoritmo␣espectral')
```

Bibliografía

[Aerts *et al.*, 2014] Aerts, H. J., Velazquez, E. R., Leijenaar, R. T., Parmar, C., Grossmann, P., Carvalho, S., Bussink, J., Monshouwer, R., Haibe-Kains, B., Rietveld, D., Hoebers, F., Rietbergen, M. M., Leemans, C. R., Dekker, A., Quackenbush, J., Gillies, R. J., y Lambin, P. (2014). Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications*, 5:4006.

[Amadasun y King, 1989] Amadasun, M. y King, R. (1989). Textural features corresponding to textural properties. *IEEE Transactions on Systems, Man and Cybernetics*, 19:1264–1274.

[Cocci *et al.*, 2015] Cocci, G., Barbieri, D., Citti, G., y Sarti, A. (2015). Cortical spatio-temporal dimensionality reduction for visual grouping. *Neural Computation*, 27:1252–1293.

[Gillies *et al.*, 2016] Gillies, R. J., Kinahan, P. E., y Hricak, H. (2016). Radiomics: Images are more than pictures, they are data. *Radiology*, 278:563–577.

[Gonzalez y Woods, 2017] Gonzalez, R. C. y Woods, R. E. (2017). *Digital Image Processing*. Pearson.

[Hastie *et al.*, 2009] Hastie, T., Tibshirani, R., y Friedman, J. (2009). *The Elements of Statistical Learning. Data Mining, Inference and Prediction*. Springer.

[Legland *et al.*, 2007] Legland, D., Kiêu, K., y Devaux, M.-F. (2007). Computation of Minkowski measures on 2D and 3D binary images. *Image Analysis and Stereology*, 26:83–92.

[Lorensen y Cline, 1987] Lorensen, W. E. y Cline, H. E. (1987). Marching cubes: A high resolution 3D surface construction algorithm. *Computer Graphics*, 21:163–169.

[Mardia *et al.*, 1989] Mardia, K. V., Kent, J. T., y Bibby, J. M. (1989). *Multivariate Analysis*. Academic Press.

[Peña, 2002] Peña, D. (2002). *Análisis de Datos Multivariantes*. S.A. McGraw-Hill/Interamericana de España.

[Rice, 2007] Rice, J. A. (2007). *Mathematical Statistics and Data Analysis*. Thomson Brooks/Cole.

[Ross, 2007] Ross, S. M. (2007). *Introducción a la Estadística*. Editorial Reverté.

[Zhang *et al.*, 2015] Zhang, L., Fried, D. V., Fave, X. J., Hunter, L. A., Yang, J., y Court, L. E. (2015). ibex: An open infrastructure software platform to facilitate collaborative work in radiomics. *Medical Physics*, 42:1340–1353.

[Zwanenburg *et al.*, 2018] Zwanenburg, A., Leger, S., Vallières, M., y Löck, S. (2018). Image biomarker standarisation initiative. https://arxiv.org/abs/1612.07003.