Introducción al Reconocimiento de Patrones
Proyecto final

Marcelo Fiori

1 de marzo de 2010

Resumen

En este trabajo se ataca el problema de la detección de pólipos en la superficie del colon, a partir de imágenes de tomografía computada (técnica denominada *Virtual Colonoscopy*).

El enfoque principal se encuentra en la inclusión de nuevas características y en la evaluación de la influencia de las mismas para mejorar la clasificación.

El resto del documento está organizado como sigue: en la sección 2 se describen todas las características utilizadas, en la sección 3 se discute la disparidad entre clases y se evalúan algunas técnicas. Finalmente en las secciones 4 y 5 se presentan respectivamente los métodos de clasificación con sus resultados y las conclusiones.

1. Introducción

El cáncer colorectal es una de las mayores causas de muerte por cáncer en el mundo. La detección temprana de pólipos es fundamental para su tratamiento, permitiendo alcanzar tasas del 90 % de curabilidad. La técnica habitual para la detección de pólipos, debido a su elevada performance, es la colonoscopía óptica (técnica invasiva y extremadamente cara).

A mediados de los '90 surge la técnica denominada colonoscopía virtual. Esta técnica consiste en la reconstrucción 3D del colon a partir de cortes de tomografía computada (CT). Es por ende una técnica no invasiva, y relativamente barata. Existe investigación a nivel universitario relativa a la detección de pólipos, pero la cantidad de falsos positivos y falsos negativos producida por éstos métodos está muy por encima de los máximos aceptados en la práctica médica, especialmente en pólipos inferiores a $1cm$ de diámetro. Este trabajo se inscribe en esta temática, la detección automática de pólipos.

Se cuenta con 10 estudios de pacientes, para cada uno se tiene la segmentación de la superficie del colon (en formato *.vti*), la ubicación de los pólipos presentes, y si fueron encontrados mediante un estudio propio de Colonoscopía Virtual. En total, se cuenta con 17 pólipos, entre los cuales hay un *flat polyp*, y los tamaños de los pólipos varían entre $4mm$ y $13mm$.

Para realizar los cálculos de las características se utilizó el lenguaje *C++* con algunas herramientas del entorno *VTK*, y las tareas clasificación se realizaron en *Weka*.

1.1. Segmentación y reconstrucción 3D

Cada estudio consta de unas 400 imágenes CT de 512×512, con una resolución de $0{,}625mm$ por píxel en x e y, y una distancia de $1cm$ entre los cortes. A partir de estas imágenes se segmenta el colon, y luego se reconstruye la superficie 3D. Este proceso es llevado a cabo por la empresa *Echopixel*.

En la figura 1 se muestra la reconstrucción de un colon, y en la figura 2 se pueden ver dos pólipos sobre la superficie reconstruída.

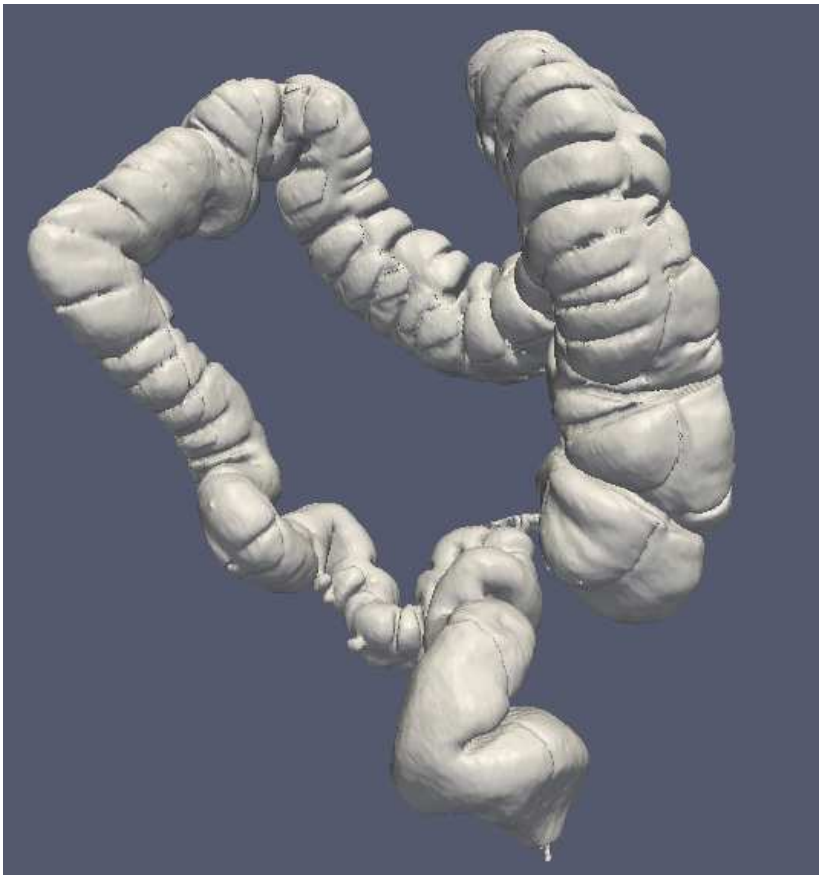

Figura 1: Reconstrucción de la superficie del colon.

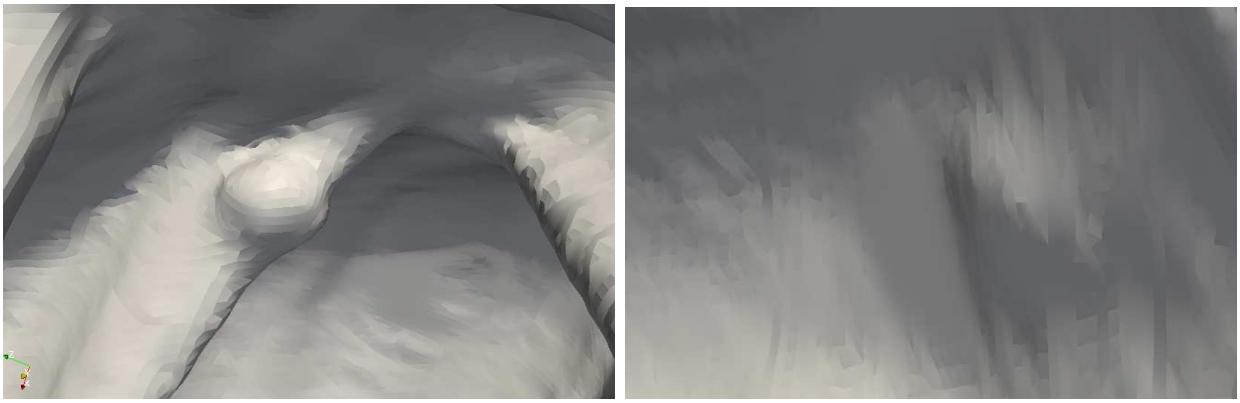

Figura 2: Pólipos sobre la reconstrucción de la superficie.

2. Características seleccionadas

Los pólipos presentan una forma geométrica particular, caracterizable, por ejemplo, por las curvaturas principales de la superficie. Dada la gran cantidad de puntos de la superficie reconstruída, resulta altamente costoso clasificar cada punto como pólipo o no pólipo, por lo tanto se realiza una primer selección de zonas, que por su geometría podrían ser pólipos. Las zonas candidatas a ser evaluadas y clasificadas son definidas a partir de una medida de curvatura llamada *Shape Index*, detallada en 2.1

2.1. Caractrísticas geométricas

Una buena medida de la forma local de una superficie es la denominada *Shape Index*, introducida por Koenderink en [4], y definida como sigue:

$$S = -\frac{2}{\pi} \arctan \frac{\kappa_{max} + \kappa_{min}}{\kappa_{max} - \kappa_{min}}$$

donde κ_{max} y κ_{min} son las curvaturas principales, máxima y mínima respectivamente. También se define el *curvedness* C, de la siguiente forma:

$$R = \sqrt{\frac{\kappa_{max}^2 + \kappa_{min}^2}{2}} \qquad C = \frac{2}{\pi} \ln R$$

Este cambio de coordenadas transforma el plano $(\kappa_{max}, \kappa_{min})$ en el plano (S, C). Mientras el valor de S es invariante en la escala y mide la forma local, el valor de C mide cuán pronunciada es la forma.

Todas las formas, salvo el plano (que tiene ambas curvaturas nulas), están determinadas por su valor de S. Por ejemplo un valor de $S = 0$ corresponde a un punto silla (curvaturas de igual módulo y opuesto signo), y un valor de $S = 1$ corresponde a las dos curvaturas iguales y positivas (como una esfera). La figuras 3 y 4 ilustran esta descripción y la completan.

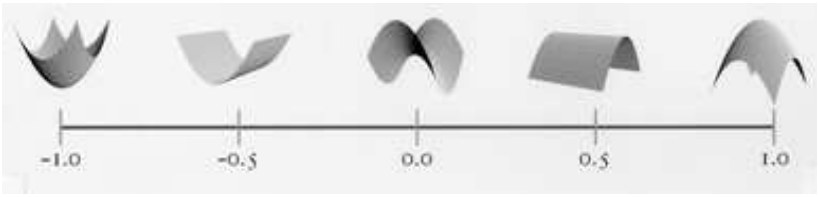

Figura 3: Algunas formas y su *Shape Index*

Los valores de interés para detectar pólipos son los de *Shape Index* cercanos a -1 en nuestro caso[1].

Dada la zona candidata Z_1, se considera además un anillo Z_2 que resulta de dilatar Z_1, para poder medir características de la zona relativas a su entorno (se buscan zonas que difieran considerablemente de su alrededor). Se calculan los histogramas de *Shape Index* sobre Z_1 y sobre Z_2, y se miden dos distancias entre estos histogramas: la distancia L_1 y la distancia de Kullback-Liebler (simetrizada). Además de estas dos características, se considera el área de Z_1 y su factor de forma, definido como la relación entre al área y el cuadrado del perímetro (normalizado por 4π). Esto totaliza cuatro características geométricas, no independientes entre sí.

[1] Esto es por la orientación de la normal a la superficie, con la orientación contraria, los valores de interés serían los cercanos a 1.

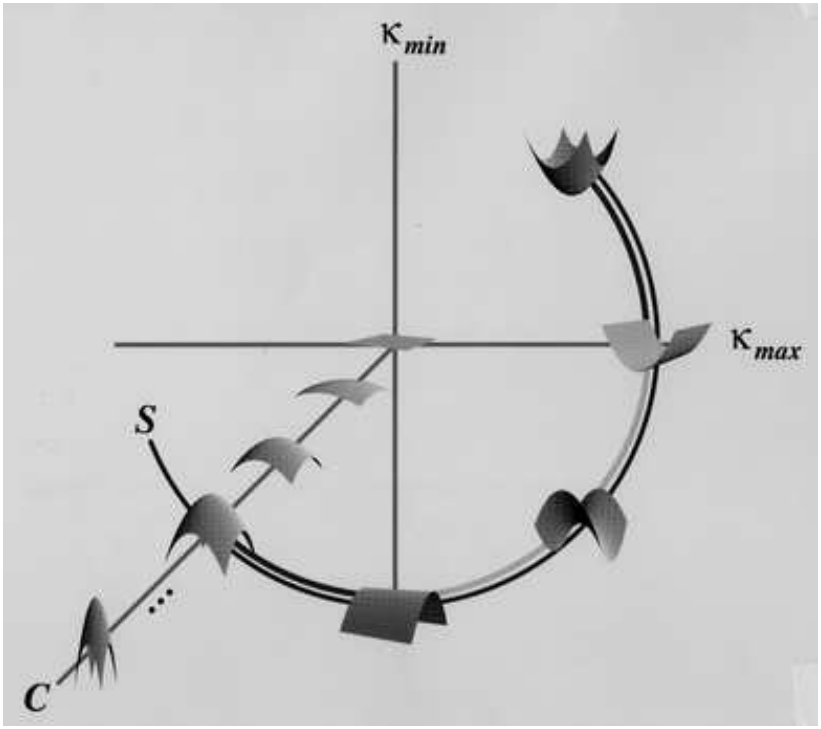

Figura 4: *Shape Index* y *Curvedness* en el plano $(\kappa_{max}, \kappa_{min})$

2.2. Características en nivel de gris

Es factible que los tejidos en los pólipos presenten distinta actividad celular que en el resto de las paredes del colon. En las imágenes de CT, esto se refleja en el nivel de gris (valores de absorbancia), por lo que es razonable introducir características relacionadas con estos valores.

Los valores de gris se encuentran en las imágenes de los cortes CT, y no en la superficie reconstruída con la que se estuvo trabajando hasta este punto. Para poder calcular características, primero se determina un volumen V_1 que incluye la zona candidata Z_1, y además contiene vóxels interiores al tejido (no en la superficie). Además, se determina un volumen V_2 alrededor de V_1, para poder comparar características de la zona candidata (V_1) con su entorno. La figura 5 muestra un ejemplo de estos volúmenes, en este caso, para un pólipo.

Una característica calculada es el promedio de los valores de gris en V_1, y también la diferencia entre este valor y el promedio calculado sobre V_2.

Además, se incluyen medidas de textura. Primero se calcula, para cada volumen, las matrices de co-ocurrencia según varias direcciones [2]. Para cada dirección, se calculan algunas de las características de Haralick ([3]): entropía, energía, contraste, máxima probabilidad y SumMean, y luego se promedian estas características en las 7 direcciones elegidas. Las mismas características se miden también de manera relativa, es decir, realizando los mismos cálculos en V_2 y almacenando las diferencias entre, por ejemplo, la entropía en V_1 y la entropía en V_2.

Esto totaliza 12 características de gris y textura (6 absolutas y 6 relativas al entorno), que tampoco son independientes entre sí.

[2] Específicamente 7: $(1,0,0), (0,1,0), (0,0,1), (1,0,1), (1,1,0), (0,1,1)$ y $(1,1,1)$.

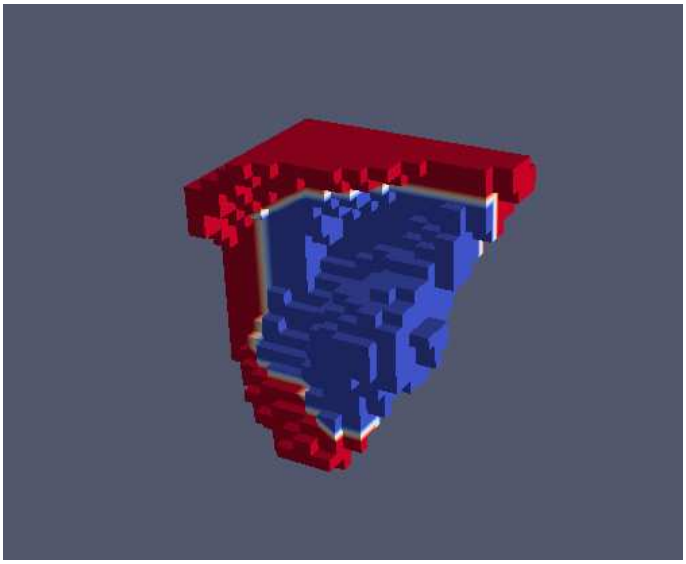

Figura 5: Volumen V_1 en azul y V_2 en rojo.

3. Disparidad entre clases

Luego de realizar la detección primaria que resulta en las zonas candidatas, se obtienen alrededor de 1300 zonas, entre las cuales se encuentran los 17 pólipos. Para contrarrestar la disparidad entre las clases, se realizaron pruebas de clasificación utilizando un muestreo de la clase predominante y generando nuevos patrones de la clase minoritaria, utilizando *SMOTE* ([2]). Algunos de los resultados, en particular clasificando con pocas características (y presumiblemente no las mejores), parecían excesivamente buenos. Sospechando que la causa podría ser la generación sintética de nuevos patrones, se procedió a realizar una evaluación del mismo, que se describe a continuación.

3.1. Evaluación de SMOTE

SMOTE (Synthetic Minority Over-sampling TEchnique) es un método para generar nuevos patrones de la clase minoritaria. Básicamente, las nuevas instancias se crean realizando combinaciones convexas entre los patrones originales.

Para evaluar la incidencia de la generación de nuevos patrones en la clasificación, se generaron conjuntos de datos de dos clases, con 5 características aleatorias, de distribución normal $\mathcal{N}(0, 1)$ para ambas clases. Teóricamente, cualquier clasificador debería rondar un 50 % de clasificación correcta.

Se crearon dos grupos de datos: uno con 100 instancias de la clase A y 2000 de la clase B, y otro grupo con 20 instancias de A y 2000 de B. Para cada grupo se realizó el siguiente experimento: se generaron nuevos patrones de la clase A utilizando SMOTE, y se muestrearon patrones de la clase B para equiparar la cantidad de patrones por clase. Luego se clasificaron los patrones utilizando el clasificador *NaiveBayes* de *Weka*, repitiendo 1000 veces los sorteos para distintas semillas.

Se repitió este proceso variando la proporción de nuevos patrones generados mediante SMOTE (desde 0 % a 2000 % en un grupo, y desde 0 % a 10000 % en el otro).

Los resultados se presentan en las figuras siguientes, donde se muestran el *True Positive Rate* y el área bajo la curva ROC, en función de la proporción de datos sintéticamente generados. Las figuras 6 y 7 corresponden a los resultados obtenidos a partir del primer grupo (donde la clase A contaba con 100 instancias originalmente), y las figuras 8 y 9 muestran los resultados para el segundo grupo (donde la clase A contaba con 20 patrones al inicio).

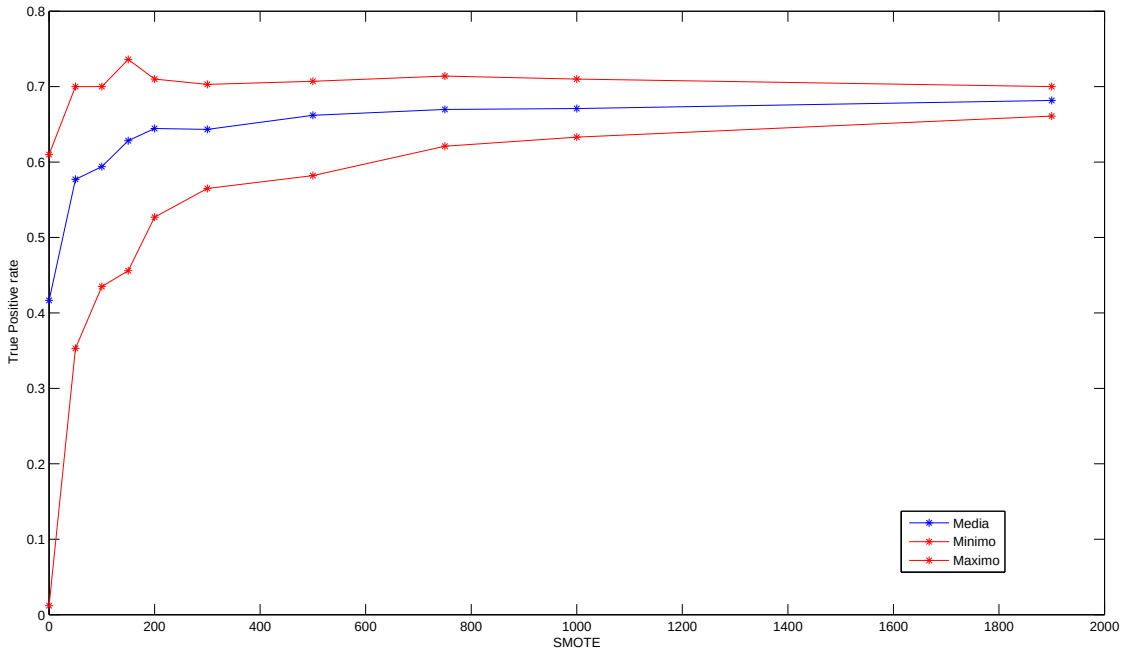

Figura 6: TPR vs Proporción SMOTE, para el grupo 1 (inicialmente 100 patrones en clase A y 2000 en B)

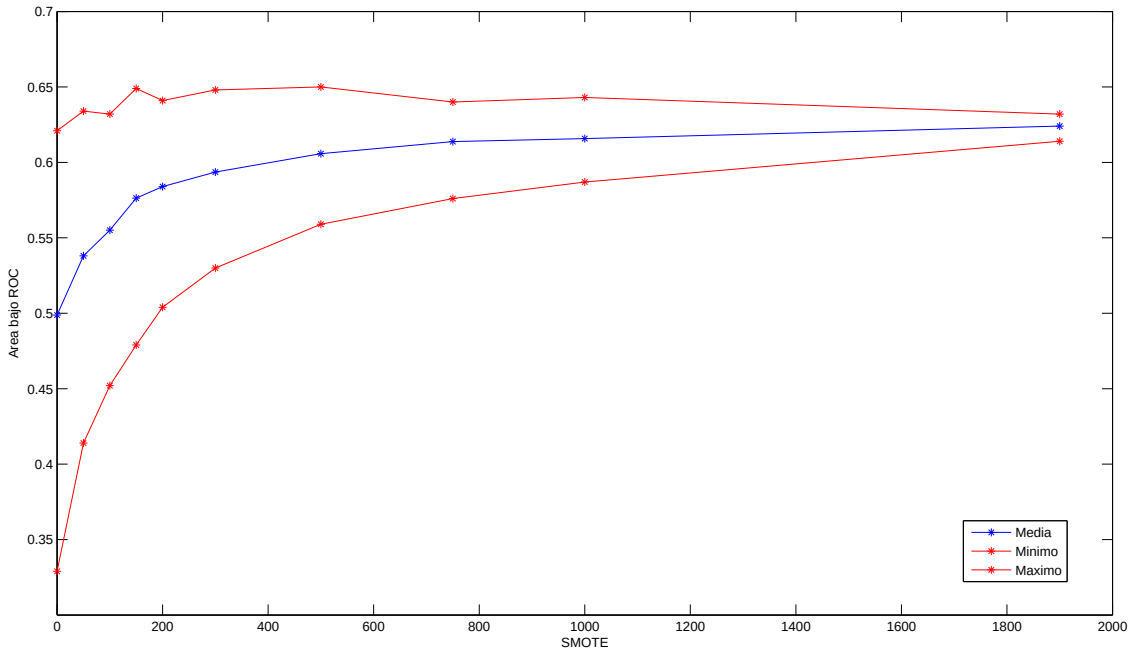

Figura 7: Area ROC vs Proporción SMOTE, para el grupo 1 (100 patrones en clase A y 2000 en B)

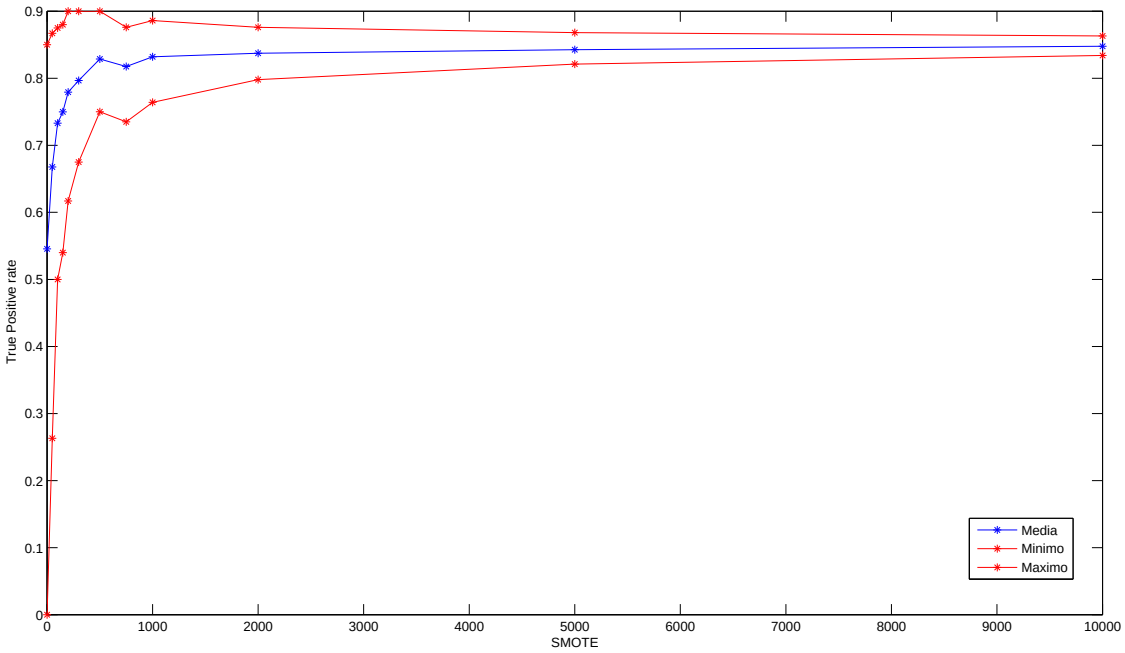

Figura 8: TPR vs Proporción SMOTE, para el grupo 2 (inicialmente 20 patrones en clase A y 2000 en B)

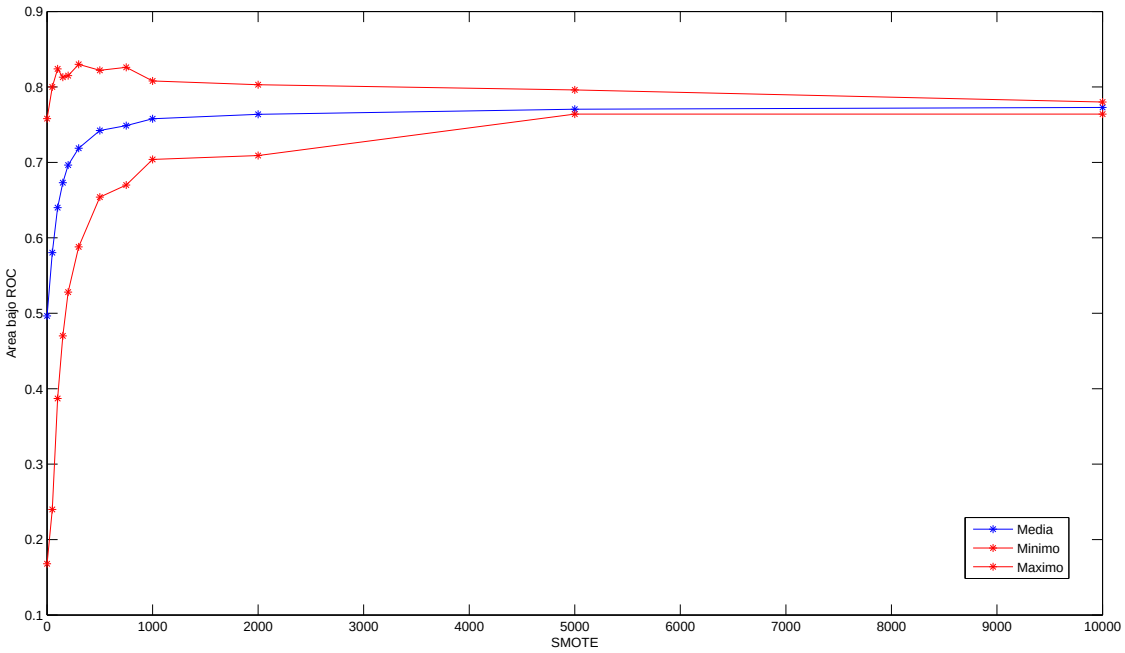

Figura 9: Area ROC vs Proporción SMOTE, para el grupo 2 (100 patrones en clase A y 2000 en B)

Dos observaciones importantes surgen de las figuras. La primera, que se ve en cada figura individualmente, es que a medida que se aumenta la proporción de patrones generados con SMOTE, la performance del clasificador aumenta, llegando a un área bajo la curva ROC de 0,77 en un caso (se recuerda que las características originales son valores aleatorios). La segunda, es que este efecto empeora a medida que la clase minoritaria tiene menos patrones inicialmente, y esto se observa comparando los resultados entre los grupos. Esto es previsible, ya que se cuenta con menos patrones para realizar las combinaciones, y por lo tanto los datos generados son mucho más dependientes entre sí.

Por estas razones, que en el caso de este trabajo llevarían a resultados de clasificación demasiado optimistas, se decidió no utilizar esta técnica de generación sintética de patrones.

3.2. Muestreo de la clase predominante

Dada la discusión de la parte anterior, para obtener resultados más justos y no tan favorables artificialmente hacia los objetivos planteados, se procedió simplemente muestreando patrones de la clase predominante, hasta equiparar los patrones por clase. La semilla de los sorteos se fue variando, repitiendo para 1000 valores distintos de semilla, y efectuando la clasificación cada vez.

4. Clasificación y Resultados

Se probaron tres clasificadores: *NaiveBayes*, *AdaBoost* con árboles *C4.5* y *SVM* de una clase.

Una de las preguntas iniciales a responder era si las características de gris y textura aportan positivamente para la clasificación[3]. Los resultados se presentan entonces de manera de contestar esta pregunta. A saber, se muestran parámetros de clasificación utilizando sólo características geométricas, sólo de gris y textura, y todas juntas. Además, se separan las características absolutas y relativas de gris y textura.

Para SVM de una clase se utilizó la librería *libsvm* ([1]) sobre *Weka*. Se entrenó con la mitad de los patrones de la clase No Pólipos, y se evaluó con la otra mitad por un lado, y con el total de los pólipos por otro. Utilizando un kernel polinómico, y ajustando los parámetros para clasificar razonablemente los pólipos (más de 15 de 17 clasificados como outliers), los resultados no fueron buenos al clasificar los no pólipos (alrededor de 300 de un total de 600 fueron clasificados como outliers).

Los resultados de clasificación utilizando *NaiveBayes* y *AdaBoost* con *C4.5* se presentan en las tablas 1 y 2 respectivamente. Se muestra el True Positive Rate promedio (sobre los 1000 sorteos con semillas distintas), el False Positive Rate promedio, y el área bajo la curva ROC promedio. Todos los valores fueron obtenidos mediante *10-cross validation*.

	Geom	Gris(abs)	Gris(rel)	Geom + Gris(abs)	Geom + Gris(rel)	Todas
TP	0.895	0.739	0.951	0.912	0.975	0.956
FP	0.136	0.531	0.203	0.168	0.118	0.139
Area ROC	0.941	0.613	0.917	0.932	0.964	0.955

Cuadro 1: Resultados de clasificación utilizando *NaiveBayes*

	Geom	Gris(abs)	Gris(rel)	Geom + Gris(abs)	Geom + Gris(rel)	Todas
TP	0.838	0.439	0.887	0.824	0.896	0.892
FP	0.145	0.324	0.166	0.148	0.128	0.129
Area ROC	0.887	0.529	0.878	0.880	0.908	0.906

Cuadro 2: Resultados de clasificación utilizando *AdaBoost*

[3]Al ver que, utilizando SMOTE, características aleatorias aportaban positivamente al resultado final, se decidió realizar análisis incluído en 3.1

Si bien es necesario realizar una selección de características más profunda (y ciertamente mejorar cada una de ellas, relizando mejores cálculos, mejores distancias, etc), entre las combinaciones presentadas se destaca la unión de características geométricas con características de gris y textura relativas. Este conjunto funciona, tanto con *NaiveBayes* como con *AdaBoost*, ligeramente mejor que todas las características juntas. Este hecho se condice con los resultados de clasificación utilizando solamente las características de gris y textura absolutas, que son definitivamente malos.

En la figura 10 se presenta el histograma del área bajo la curva ROC (AUC) sobre los 1000 sorteos, utilizando *NaiveBayes* con características geométricas y características de gris y textura relativas (la combinación que maximiza el AUC promedio en las tablas). Se puede ver que, si bien el promedio es de 0,964, la distribución está lejos de ser simétrica, y la mayoría del área está cerca $AUC = 1$. En particular, la mediana es 0,972 y la moda es 1.

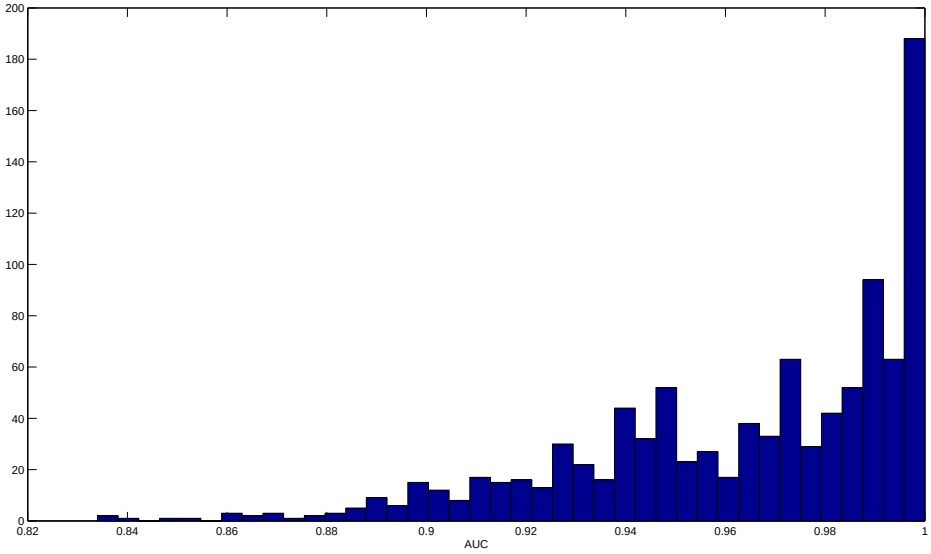

Figura 10: Histograma de AUC

5. Conclusiones

Los resultados de clasificación son prometedores, sobretodo teniendo en cuenta que se aún pueden refinar las características por un lado (tanto su cálculo como su selección, en particular, los resultados presentan una gran sensibilidad respecto a la forma de determinar los volúmenes V_1 y V_2 definidos en 2.2), y el método de clasificación por otro.

Los mayores aportes de este trabajo son principalmente dos. El primero, es la inclusión de características de Haralick de textura, y fundamentalmente la consideración relativa de las mismas (buscar diferencias de textura con un entorno); esto es nuevo en el tópico, y se comprobó en este trabajo que mejora la clasificación. El segundo es más bien un aviso, y es que hay que ser cuidadoso al generar datos sintéticamente con SMOTE y sacar conclusiones apresuradas sobre las características utilizadas. Esto no significa que no se pueda utilizar SMOTE, sino que las medidas de clasificación al utilizarlo, no son del todo justas. Quizás se pueda medir, al usar SMOTE, el nivel de discriminación de una característica comparándolo con una característica aleatoria utilizando la misma proporción de SMOTE, es decir, la ganancia de una característica respecto a una aleatoria, para las mismas condiciones de evaluación.

Referencias

[1] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[2] N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[3] Robert M. Haralick, K. Shanmugam, and Its'hak Dinstein. Textural features for image classification. *Systems, Man and Cybernetics, IEEE Transactions on*, 3(6):610–621, 1973.

[4] Jan J. Koenderink. *Solid shape*. MIT Press, Cambridge, MA, USA, 1990.