Wisconsin Prognostic Breast Cancer
Support Vector Machine

Dany Efrain Rubiano Jiménez
Universidad de Santiago de Chile

Abstract. La base de datos Wisconsin prognostic breast cancer alberga información acerca del seguimiento de distintos casos de cáncer de mama, enfocándose en la identificación de la recurrencia de este cáncer en específico. En esta oportunidad se hace uso del método SVM para la clasificación, buscando obtener conocimiento acerca de la recurrencia y no recurrencia del cáncer.

Keywords: Recurrencia, SVM.

1 Introducción

Una recurrencia o cáncer de mama recurrente es un cáncer de mama que vuelve a aparecer después de un determinado período en el que ya no fue detectado. Es un cáncer de mama que se ha propagado a otra parte del cuerpo, las células cancerosas pueden desprenderse del tumor original de la mama y alojarse en otras partes del cuerpo usando el torrente sanguíneo o el sistema linfático, una gran red de ganglios y vasos que eliminan bacterias, virus y desechos celulares.
Es por ello que el Dr Wolberg de la Universidad de Wisconsin, se ha dedicado desde 1984 a albergar registros del seguimiento de pacientes con cáncer de mama, con el fin de investigar qué características de la morfología, textura y dimensión celular, además de otros antecedentes reunidos en el proceso de cirugía de extracción de metástasis, inciden en la recurrencia
El problema radica en cómo clasificar un cáncer de mama en base a las características registradas, por lo que en el presente estudio se realiza la aplicación del método Support Vector Machine para la clasificación.

Los objetivos de este estudio corresponden a:

- Realizar un proceso de selección de las características más importantes.
- En base al ranking de características obtener una clasificación con SVM variando los kernels y parámetros disponibles.
- Comparar los resultados con métodos de clasificación anteriores.

La hipótesis que se aborda en el presente trabajo guarda relación con que mediante SVM, se obtiene una mejor clasificación para el conjunto de datos Wisconsin prognostic breast cáncer, que con Random Forest.

2 Métodos y datos

2.1 Métodos utilizados

El método principal a utilizar corresponde al de Support Vector Machine (SVM). Para la clasificación una SVM realiza la separación de las clases a través al aumento de dimensionalidad del problema basándose en un teorema de existencia. Las SVM utilizan una función kernel que define la distancia entre los datos nuevos y los vectores de soporte.
Para el presente estudio, se utiliza un kernel lineal, cuya medida de distancia utilizada es el producto punto. Además, se utiliza un kernel radial, el cual permite transformar el espacio de entrada en dimensiones superiores.

En lo que se refiere a la selección de características, se utiliza las funciones disponibles por el paquete de RWeka, además de la importancia de las variables a través de la técnica de Mean Decrease Gini resultante en la aplicación de la clasificación a través de Random Forest.
A las selecciones de los datos, se les aplica SVM radial y lineal por lo menos 3 veces usando validación cruzada de 10 folds, con Gamma en el rango de [2^(-15),2^3] y costo en el rango [2^(-5),2^15], según corresponda, tal como es recomendado en la literatura [4].

Por otro lado, en vista de que los datos están desbalanceados, ya que la clase no recurrente presenta una gran mayoría de instancias en comparación con la clase recurrente, se utiliza la técnica de balanceo de datos smote. Dicha técnica realiza una combinación entre un sobremuestreo de la clase minoritaria y un submuestreo de la clase mayoritaria.

2.2 Datos utilizados

El conjunto de datos reúne diferentes registros del seguimiento de un caso de cáncer de mama. Estos pacientes fueron vistos consecutivamente por el Dr. Wolberg desde 1984, e incluyen sólo aquellos casos que presentan cáncer de mama y ninguna evidencia de metástasis a distancia al momento del diagnóstico.
El conjunto de datos cuenta con 198 registros con 34 atributos y la clase. Esta última corresponde al resultado, recurrencia o no recurrencia del cáncer. La Tabla 1 muestra la descripción de los distintos atributos del conjunto de datos.

Tabla 1: Definición del conjunto de datos

Variable	**Descripción**
Identificador	Número único de identificación del paciente
Resultado	Etiqueta médica (R = recurrencia de cáncer / N = No recurrencia de cáncer)
Tiempo (t)	Tiempo de recurrencia si Resultado = R y Tiempo libre de enfermedad en caso contrario.
Radio (r) *	Distancias desde el centro hasta los puntos del perímetro
Perímetro (p) *	Longitud que envuelve la región celular muestreada mediante la imagen

Área (a) *	Medida cuadrática de superficie de la región celular estudiada
Textura (t) *	Desviación estándar de los valores de la escala de grises
Suavidad (s) *	Variación local en longitudes de radio
Compacidad (com) *	Perímetro al cuadrado dividido en el valor de área menos uno
Concavidad (con) *	Gravedad de las partes cóncavas del contorno
Puntos Cóncavos (pc) *	Número de partes cóncavas del contorno
Simetría (s) *	Medida de semejanza de los hemisferios de la célula
Dimensión Fractal (df) *	Aproximación de la costa fractal menos uno
Tamaño del tumor (ts)	Tamaño del tumor extraído al momento de la cirugía
Ganglios Linfáticos (ly)	Ganglios inflamados al momento de la cirugía

(*): Estos corresponden a atributos que son captados a partir de una imagen digitalizada de una aspiración con aguja fina (FNA) de una masa mamaria. En el conjunto de datos se presentan la media, desviación estándar y el promedio de los tres casos más altos (peor caso), obteniéndose de esta manera 30 atributos.

Para el preprocesamiento de los datos, se elimina la variable identificador, debido a que no aporta ninguna información acerca del problema en cuestión. Además de esto, se eliminan 4 registros que presentan valores perdidos, quedando en total 194 registros disponibles para el estudio en cuestión.

Dado que el conjunto de datos se encuentra desbalanceado, se aplica smote, resultando 184 registros, 92 por cada clase.
Para la aplicación de SVM, el conjunto de datos se divide en una muestra del 70% para el entrenamiento y un 30% para las pruebas.

3 Resultados

En primer lugar, se aplica SVM sin ninguna selección de variables previa, a modo de comparar los resultados sobre el conjunto de datos sin balancear y una vez aplicado smote. Los resultados se presentan en la Tabla 3.1

Tabla 3.1: *SVM sobre el conjunto de datos desbalanceado y con smote.*

	Kernel Lineal			**Kernel Radial**			
	Costo	**Error**	**AUC**	**Costo**	**Gamma**	**Error**	**AUC**
Datos desbalanceados	2	0,220	0,585	8	0,003	0,225	0,488
Smote	16	0,276	0,814	2	0,125	0,207	0.9074

A partir de estos resultados, se opta por seguir utilizando smote para posteriores aplicaciones de SVM.
En lo que respecta a la selección de variables, se utiliza el criterio Maan Decrease Accuracy, procediente de la aplicación resultante de la clasificación mediante Random

Forest en estudios anteriores. El ranking de variables obtenido se puede observar en la Figura 3.1.

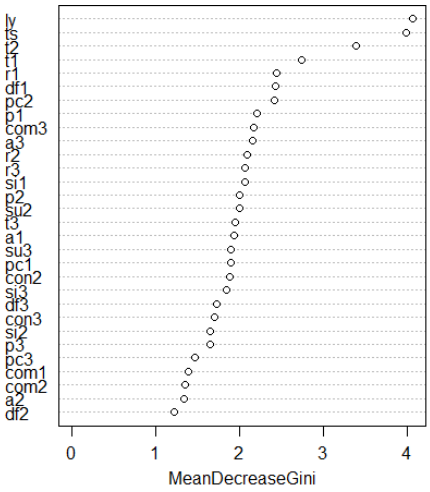

***Figura 3.3**: Importancia de variables*

Buscando cumplir el principio de parsimonia, se utilizan en primer lugar las 9 primeras variables con mayor importancia, disminuyendo la cantidad de variables consideradas una vez aplicada la SVM. Los resultados de este proceso se reflejan en la Tabla 3.2.

***Tabla 3.2:** SVM con selección de variables a través de Mean Decrease Gini.*

	Kernel Lineal			**Kernel Radial**			
	Costo	**Error**	**AUC**	**Costo**	**Gamma**	**Error**	**AUC**
ly,t2,ts,df1,a1,a3,p1,su3,si1	0,031	0,330	0,55	8	0,125	0,192	0,759
ly,t2,ts,df1,a1,a3,p1,su3	0,125	0,330	0,57	16	0,125	0,184	0,833
ly,t2,ts,df1,a1,a3,p1	0,031	0,338	0,61	32	0,5	0,253	0,796
ly,t2,ts,df1,a1,a3	0,031	0,307	0,61	1	1	0,246	0,777
ly,t2,ts,df1,a1	0,25	0,315	0,57	2	2	0,253	0,833
ly,t2,ts,df1	0,5	0,361	0,61	2	16	0,207	0,777
ly,t2,ts	0,25	0,369	0,61	0,5	16	0,223	0,777

Weka a través de la función *InfoGainAttributeEval*, otorga una evaluación de la ganancia de información de los atributos, a través de la cual se presenta la selección de características mostrada en la Tabla 3.3.

Tabla 3.3: *Selección de variables a través de las funciones de Weka*

Variables	Importancia
ly	0,1568
df1	0,1289
ts	0,1051
r2	0,0680
a3	0,0678
r3	0,0674
si1	0,0625

A partir de la selección de características mediante las herramientas que otorga Weka, buscando cumplir el principio de parsimonia, se utilizan en primer lugar las 7 primeras variables consideradas, disminuyendo la cantidad de variables una vez aplicada la SVM. Los resultados de este proceso se reflejan en la Tabla 3.4.

Tabla 3.4: *SVM con selección de variables a través de Mean Decrease Gini.*

	Kernel Lineal			**Kernel Radial**			
	Costo	**Error**	**AUC**	**Costo**	**Gamma**	**Error**	**AUC**
ly,df1,ts,r2,a3,r3,si1	0,031	0,276	0,611	32	0,25	0,176	0,888
ly,df1,ts,r2,a3,r3	0,25	0,315	0,574	32	0,25	0,2	0,796
ly,df1,ts,r2,a3	0,031	0,323	0,592	32	0,25	0,169	0,796
ly,df1,ts,r2	0,125	0,338	0,537	32	0,015	0,26	0,629
ly,df1,ts	0,031	0,369	0,592	1	8	0,253	0,759

Finalmente, el mejor clasificador teniendo en cuenta las características del problema y buscando el principio de parsimonia, es el obtenido a través de las herramientas que provee Weka considerando 7 variables. El detalle de la matriz de confusión de la clasificación se presenta en la Tabla 3.6, mientras que para la comparación de resultados, la Tabla 3.6 muestra la matriz de confusión generada de la clasificación mediante Random Forest basada en la selección de variables mediante el criterio Mean decrease Gini en conjunto con la aplicación de smote.

Tabla 3.5: *Matriz de confusión mejor clasificación.*

	N	**R**
N	24	3
R	3	24
Accuracy = 88,88%		

Tabla 3.6: *Matriz de confusión de la clasificación mediante Randon Forest.*

	N	**R**
N	58	8
R	8	58
Accuracy = 87,87%		

4 Discusión

En primer lugar, en la comparación de la clasificación obtenida con SVM para los datos originales desbalanceados con respecto los balanceados con smote, se observa como aumenta considerablemente el área bajo la curva ROC después de la aplicación de smote, mejorando así la clasificación.
Es importante destacar que los parámetros gamma y costo no deben ser simplemente elegidos al azar. El costo es el peso que se le da a cada observación a la hora de clasificar es decir, es la regularización del impacto que se le otorga a las variables de holgura del problema de optimización. Gamma por su parte, otorga una relación en función del kernel utilizado, permitiendo encontrar los subespacios que puedan diferenciar los puntos en el espacio, permitiendo además la adición de una mayor complejidad a la hora de separar observaciones. Los resultados muestran que el gamma se movía en un rango entre 0,015 hasta un máximo de 16, mientras que el costo en el caso del kernel radial se situaba en un rango de 0,5 a 32. Por otro parte, para el kernel lineal, el costo no sobrepasaba el 1.
En una comparación de resultados entre el uso de los tipos de kernel, aquel que entregó mejores resultados en términos del área bajo la curva es el kernel radial. Esto demuestra que se hace necesario aumentar la complejidad al momento de separar las observaciones.
Se destaca los buenos resultados alcanzados con los dos criterios de selección de características utilizados. Es importante señalar que la reducción de características podría ser mayor según los resultados obtenidos, ya que a pesar de que a medida que se reduce el número de variables consideradas, el área bajo la curva también lo hace, esta disminución no es de gran significancia, de modo que con 3 varianles el área bajo la curva ROC para ambos criterios de selección, se registra en torno a 0,75.
A partir de ello, se destaca que las variables que otorgan mayor información corresponden al estado de los ganglios linfáticos (ly) y el tamaño del tumor (ts), seguida de otras variables relacionadas con la textura y dimensión celular.
Cabe destacar que debido a la naturaleza del problema, en el que se busca conocimiento acerca de la recurrencia del cáncer de mama, resulta necesario incurrir en un modelo que tenga una mayor exactitud, es por ello que se opta por elegir la clasificación señalada en la Tabla 3.5, y no otro que sea menos complejo pero castigado en la reducción de exactitud.
Por último, los resultados de la clasificación mediante SVM, en comparación con lo obtenido a través de Random Forest no difieren significantemente entre sí, al tomar en cuenta la medida de exactitud. Sin embargo, se debe resaltar que las variables consideradas para ambos métodos son medianamente distintas, dados los criterios de selección considerados.

5 Conclusiones

En lo que refiere a la hipótesis planteada al comienzo del presente trabajo, los resultados no permiten su validación, o invalidación. Esto es porque los resultados obtenidos de clasificación con SVM y Random Forest son similares en términos de exactitud. Sin

embargo, cabe destacar que en la aplicación de ambos métodos sin la presencia de Smote, la clasificación a través de SVM logra resultados significativamente mejores a los obtenidos con Random Forest, en el que el error de clasificación rondaba el 90%. Esto puede deberse a que SVM busca asegurar el óptimo, convergiendo al mínimo global.

Se puede comprobar en base a los modelos generados que las variables que más inciden en la recurrencia corresponden al estado de los ganglios linfático, en conjunto con el tamaño del tumor. A su vez, la complejidad disminuye drásticamente al considerar sólo 7 variables de las 32 originales, obteniendo de todas maneras buenos resultados. Por supuesto este conocimiento debe ser validado por un experto.

6 Referencias

1. Chacón, M. Taller de minería de datos avanzada, Capitulo V: Support Vector Machime (SVM).
2. R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
3. William H. Wolberg. Computer-Derived Nuclear "Grade" and Breast Cancer Prognosis. Departments of Surgery, Human Oncology, and Computer Sciences University of Wisconsin, Madison.
4. Hsu, C. , Chang, C. , Lin, . C. A Practical Guide to Support Vector Classification.