Wisconsin Prognostic Breast Cancer
Agrupamiento basado en modelos

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 cancer de mama, enfocándose en la identificación de la recurrencia de este cancer en específico. En esta oportunidad se hace uso del método de agrupamiento basada en modelos, buscando obtener conocimiento acerca de los datos en estudio mediante dicho método.

Keywords: Recurrencia, agrupamiento.

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.

Los objetivos de este estudio corresponden a:

- Analizar los datos correspondientes a la base de datos Wisconsin prognostic breast cancer
- Obtener conocimiento mediante el uso del método de agrupamiento basado en modelos.

2 Métodos y datos

1.1 Métodos utilizados

El método principal a utilizar es el método de *agrupamiento* basado en modelos, el cual busca maximizar el ajuste de los datos a las agrupaciones en modelos estadísticos multivariados. Cabe destacar que este método requiere de distintos parámetros, entre los cuales se encuentra la configuración de los modelos y el número de grupos a considerar, por lo que se usa el criterio BIC, junto al criterio ILC para determinar el número de grupos óptimo y la configuración recomendada.
Antes de su aplicación, se hace necesario en primer lugar identificar los atributos que aportan mayor información y que por lo tanto, son de mayor relevancia para el problema en cuestión. Para ello se realiza un análisis estadístico y un preprocesamiento de los distintos datos.

1.2 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 194 instancias 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	Tiempo de recurrencia si Resultado = R y Tiempo libre de enfermedad en caso contrario.
Radio *	Distancias desde el centro hasta los puntos del perímetro
Perímetro *	Longitud que envuelve la región celular muestreada mediante la imagen
Área *	Medida cuadrática de superficie de la región celular estudiada
Textura *	Desviación estándar de los valores de la escala de grises
Suavidad *	Variación local en longitudes de radio
Compacidad *	Perímetro al cuadrado dividido en el valor de área menos uno
Concavidad *	Gravedad de las partes cóncavas del contorno
Puntos Cóncavos *	Número de partes cóncavas del contorno
Simetría *	Medida de semejanza de los hemisferios de la célula
Dimensión Fractal *	Aproximación de la costa fractal menos uno
Tamaño del tumor	Tamaño del tumor extraído al momento de la cirugía
Ganglios Linfáticos	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 así 30 atributos

3 Resultados

El estudio para el agrupamiento basado en modelos es realizado a partir del criterio BIC y corroborado con el criterio ICL, a manera de poder determinar el número de grupos adecuado y la complejidad para los modelos.

Dado el análisis estadístico realizado, ver Anexo, se procede a explorar para cada una de las disposiciones de las variables los resultados del criterio del BIC. En este caso, el mejor valor del BIC se obtiene con las variables obtenidas a partir de las componentes principales, en combinación con la correlación de Pearson. Se concluye entonces que el mejor agrupamiento es a través de 2 grupos y con una configuración VEV.

Tabla 2. Mejores BIC obtenidos

	VEV,2	**EEV,2**	**EEE,1**
BIC	15165,1	14991,4	14971
BIC diff	0	-173,6	-194

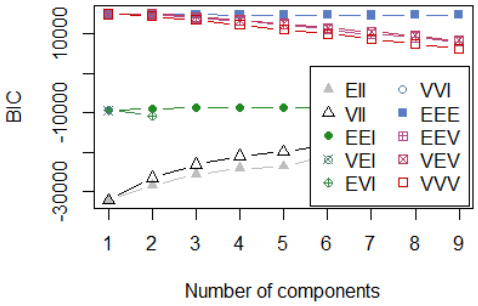

Fig. 1. Gráfico del BIC

Los resultados de las características del agrupamiento son corroboradas con el criterio del ICL, obteniéndose los mismos resultados.
A partir de los resultados anteriores, el agrupamiento realizado es representado en la Figura 3.

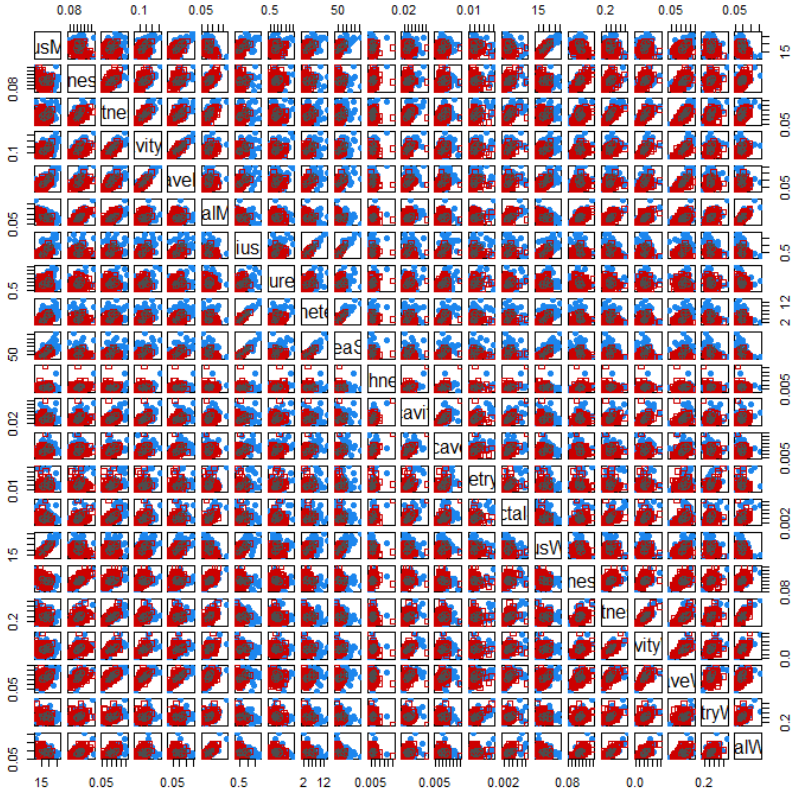

Fig. 2. Mejor agrupamiento con los atributos seleccionados, según las métricas BIC e ICL.

Dado que este es un método de agrupamiento, a modo de comparación se registran los resultados de la aplicación del agrupamiento por k-medias, con k=2, con la presentación de la tabla de contingencia que refleja la clasificación en grupos respectiva. Así mismo, se presenta la tabla de contingencia respectiva al agrupamiento basado en modelos

Tabla 3. Tabla de contingencia del agrupamiento basado en modelos.

	N	**R**
N	100	48
R	32	14

Tabla 4. Tabla de contingencia del agrupamiento de k-medias

	N	**R**
N	116	32
R	29	17

Luego, se presenta la comparación de las clasificaciones de los distintos agrupamiento con las clases originales de las instancias del conjunto de datos.

Tabla 5. Nivel de acierto entre grupo y clase en porcentaje.

	Modelo VEV, 2	**K-medias, k = 2**
N	66,22%	76,82%
K	29,78%	36,17%

4 Discusión

Dadas las distintas configuraciones de variables obtenidas a partir del análisis estadístico, se procedió a realizar una exploración de cada una con el criterio del BIC, agregándose otro procedimiento de selección de variables para el análisis de agrupamiento. En esta ocasión, tal como se mencionó en el apartado anterior, el mayor valor del BIC se obtuvo a partir del análisis de las componentes principales, en conjunto con el análisis de la correlación de Pearson.

El criterio del BIC en esta ocasión resulta en la agrupación en 2 clusters, lo cual se equipara con el conocimiento a priori del conjunto de datos, el cual cataloga las instancias en las clases recurrentes y no recurrentes.

Buscando cumplir con el principio de parsimonia, se podría optar por tomar el modelo con una configuración EEV-2, en vez de la actual, que es la que indica un mayor valor de BIC, bajando la complejidad del agrupamiento.

De la tabla de contingencia, se aprecia un alto número de falsos positivos y negativos, de tal manera que esta agrupación con respecto al conocimiento previo de la clasificación del conjunto de datos, otorga un acierto del 66,22% para los casos no recurrentes, y un 36,17% para los recurrentes, resultados bajos que se pueden deber al análisis de las variables a considerar.

Como ser aprecia en la tabla 4 los niveles de acierto en cuanto al grupo y la clase proporcionada por el conjunto de datos muestran la paridad que existen en los resultados del agrupamiento basados en modelos y el agrupamiento de k-medias. Esto puede ser debido a la configuración utilizada para los modelos, en donde se mantuvo el volumen variable, la forma igual y la orientación variable.

5 Conclusiones

Durante el desarrollo de la experiencia el análisis estadísitco en conjunto con la exploración del criterio BIC, permitió la reducción de las dimensiones involucradas en el problema, bajando desde 34 dimensiones a 22, lo que posibilitó obtener los resultados presentados.

El modelo obtenido considera que la cantidad de grupos que mejor separa a los datos coincide con la cantidad de clases que se tienen como conocimiento a priori, por lo cual se puede inferir que dentro de cada clase no se presentan otros tipos de “subclases”, que podrían derivar en el caso de las recurrencias, a tener dentro de estas a ciertos tipos de recurrencias.

En lo que respecta a la extracción de conocimiento, este no varía en gran medida entre el método de k-medias y el agrupamiento basado en modelos, por lo que el método a elegir, puede ser según el principio de parsimonia, buscando disminuir la complejidad.
Según los resultados, es importante destacar que los métodos comprendidos actuarán de una forma deseada siempre y cuando el conjunto de datos esté considerando las cualidades o atributos que permitan diferenciar correctamente los grupos buscados. Por lo tanto, es de suma importancia realizar un correcto análisis estadístico previo para discriminar las variables que otorgan mayor información. En este caso, dicho proceso presentó una gran dificultad debido a las características propias del conjunto de datos y a la gran cantidad de atributos que presenta.

6 Referencias

1. Chacón, M. Taller de minería de datos avanzada, Capitulo 1: Agrupamiento basado en Modelos.
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.

7 Anexo

7.1 Correlación de Pearson

En primer lugar se aplica una correlación de Pearson, para analizar la relación lineal entre cada una de las variables. Debido a que se cuentan con 34 atributos en total, en esta ocasión no se presentan todos los resultados, destacando solamente aquellos que presentan un mayor valor.

Tabla 6. Correlación de Pearson.

Variable	Coef. de correlación	Variable
RadiusMean	0.99	PerimeterMean
	0,99	AreaMean
	0,91	PerimeterWorst
	0,89	AraeWorst

De la información anterior se destaca que debido a la alta correlación entre el RadioMean y las otras variables, se pueden eliminar estas últimas, conservando solo la variable RadioMean, ya qque están aportando casi la misma información.

7.2 Test de Normalidad

Se aplicó un test de normalidad de Shapiro-Wilk sobre cada una de las variables para estudiar su distribución. Los resultados indican que las únicas variables que se distribuyen normal dado que los resultados del test muestran que el p-valor es mayor a la significancia son TextureWorst y ConcaveWorst con p-valores de 1.020204e-01 y 6.582631e-01 respectivamente

7.3 Correlación de Spearman

Ahora se aplica una correlación de Spearman, teniendo en cuenta que la mayoría de las variables no se distribuyen como una normal, para estudiar la relación lineal entre la clase y cada una de las variables. En la Tabla 7 se presentan aquellas que presentan un coeficiente de correlación más alto, por lo cual podrían otorgar mayor información para la clasificación.

Tabla 7. Correlación de Spearman.

Variable	rho
Time	-0,34
RadiusMean	0,15
PerimeterMean	0,15
AreaMean	0,16
ConcaveMean	0,10
FractalMean	-0,10
RadiusSE	0,14
PerimeterSE	0,14
AreaSE	0,15
RadiusWorst	0,19
PerimeterWorst	0,18
AreaWorst	0,19
TumorSize	0,23
Lymph	0,24

7.4 Análisis de componentes principales

Tomando en cuenta las tres primeras componentes, se registra el 65% de la información, la siguiente figura refleja la varianza que tiene cada una de las componentes.

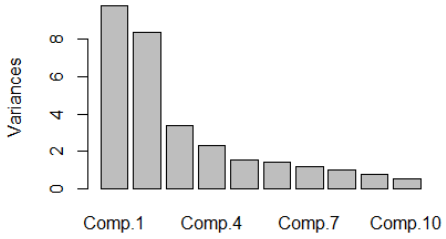

Fig. 3. Varianza de cada una de las 10 primeras componentes principales

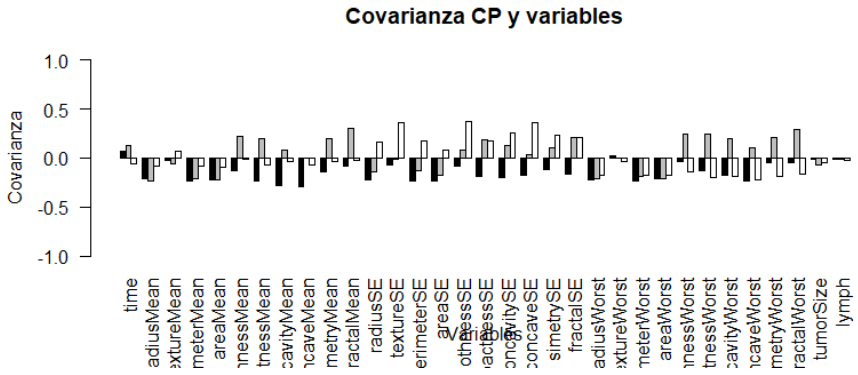

Fig. 4. Covarianza de las componentes principales y la varianza

De los datos representados por la Figura 4, se obtiene que las variables que tienen una mayor varianza según las componentes son: radiusMean, perimeterMean, areaMean, smoothnessMean, compactnessMean, concavityMean, concaveMean, simetryMean, fractalMean, radiusSE, textureSE, perimeterSE, areaSE, smoothnessSE, concavitySE ,concaveSE, simetrySE, fractalSE, radiusWorst, perimeterWorst, areaWorst, smoothnessWorst, compactnessWorst, concavityWorst, concaveWorst, simetryWorst, y fractalWorst.

7.5 Exploración del análisis estadístico en combinación con criterio BIC

A manera de resumen en la siguiente tabla se muestra la exploración del análisis estadístico comprendido anteriormente combinando las variables seleccionadas según la información obtenida para cada caso. Dicha exploración se evalúa con el criterio BIC, para tener aquel valor de BIC máximo que genere un mejor agrupamiento.

Tabla 8. Exploración del análisis estadísstico y criterio BIC.

Consideración	Valor BIC	Configuración
Variables considerando Spearman	-10455,39	VEV,2
Variables considerando Spearman con Pearson	-4864,66	VVV,2
Variables obtenidas de las componentes principales	9465,86	VEV,2

Variables combinando componentes principales con Pearson	15165,1	VEV,2
Variables considerando solo los atributos de las medias	-2055,707	EEE,5
Variables considerando solo los atributos de las medias combinado con Pearson	507,18	EEE,3
Variables combinando componentes principales con Pearson y Spearman	-866,35	VVV,2

Finalmente, se puede obtener que la mejor consideración de las variables es en la combinación de los resultados de las componentes principales y la correlación de Pearson entre las variables, obteniendo un mayor valor de BIC en comparación con las otras disposiciones.