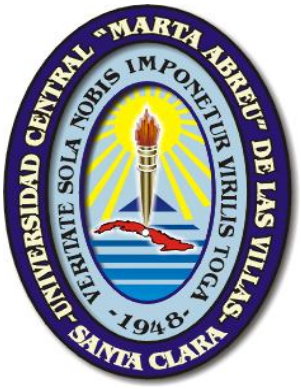
Universidad Central "Marta Abreu" de las Villas

Facultad Matemática Física y Computación

Licenciatura en Matemática



Trabajo de Diploma

"Métodos de extracción de características en datos de microarreglos de ADN para enfermedades oncológicas"

Autor: Miguel Alejandro Gutiérrez Arce.

Tutor: Msc. Yunier E. Tejeda Rodríguez,

Dr. Carlos Rodríguez Fadragas.

"Santa Clara, 2017"

"Año 59 de la Revolución"

A mis abuelos Nidia, Nelson, Esmildo, Consuelo y Ángel, en especial a este último por su preferencia hacia las Matemáticas.

Agradecimientos

Por mucho que me cueste redactar, no podría obviar esta página, pues sería un crimen olvidar aquellos que de alguna forma hicieron realidad este sueño. Por ello deseo agradecer ese apoyo desinteresado que me acompañó durante el extenso trayecto hacia mi formación:

"A mi madre, por esa constancia inagotable que permitió adentrarme en el mundo del estudio"

"A mi padre, quien depositó en mí el gen del análisis, siendo este imprescindible en el mundo de las Matemáticas"

"A mi abuela Consuelo y mis primas por brindarme todo su cariño"

"A mi tía Isabel, por su ayuda incondicional en labores de impresión, facilitando en muchas ocasiones, mi vida de trabajador y estudiante"

"A mis abuelitos Nidia y Nelson, por ser capaces de ofrecer lo que no tienen con tal de hacerme feliz"

"A mi hermosa Camila, por su cariño, comprensión y sobre todo paciencia ante el estrés que merita este proceso"

"A la Magariños, por colaborar en la revisión de esta tesis"

"A mis segundos padres Denis y Pablo, por su apoyo absoluto"

"A Inés, por desarrollar mis conocimientos pedagógicos en el Lázaro Cárdenas"

"A mi tutor Yunier, por adentrarme en el mundo de la Ciencia"

"A Jorge Luis Morales, por su aporte en el trabajo y sus recomendaciones para futuras investigaciones"

"A todos los profesores y compañeros que contribuyeron a mi formación como profesional"

Resumen

En esta investigación se presenta el diseño metodológico de la investigación. Se realiza una caracterización de tres métodos de extracción de características lineales y se comentan los principales problemas que presentan los datos de microarrays de ADN. Se realiza una caracterización de los algoritmos aleatorios, en particular la descomposición matricial CUR. Se propone tres metodologías para obtener un modelo de clasificación que pronostique diferentes enfermedades oncológicas. Por último, se hace una discusión de los resultados por dichas metodologías.

Abstract

This investigation presents the methodological design of the research. A characterization of three methods of extraction of linear characteristics is carried out and the main problems presented by DNA microarray data are commented. A characterization of the random algorithms, in particular the matrix decomposition CUR is performed. We propose three methodologies to obtain a classification model that predicts different oncological diseases. Finally, a discussion of the results is made by said methodologies.

Índice General

Dedicatoria ..... *I*

Agradecimiento ..... *II*

Resumen ..... *III*

Abstract ..... *IV*

Introducción *1*

1. Conjuntos de datos de microarrays de ADN para el cáncer *3*

1.1. Características de los datos microarray 4

1.2. Conclusiones del Capítulo 4

2. Métodos de extracción de características lineales en datos de microarrays de ADN *5*

2.1. Análisis de Componentes Principales 5

2.2. Análisis de Componentes Principales Supervisados 6

2.3. Mínimos Cuadrados Parciales. 8

2.4. Metodología propuesta 10

2.5. Resultados. 11

2.6. Discusión. 13

2.7. Conclusiones del Capítulo. 14

3. Algoritmos aleatorios en datos de microarrays de ADN. *15*

3.1. Descomposición matricial CUR 15

3.2. Metodología propuesta 16

3.2.1. Doble reducción de la dimensión 16

3.2.2. Doble reducción de la dimensión en forma paralela 18

3.3. Resultados. 19

3.4. Discusión 22

3.5. Conclusiones del Capítulo 22

Conclusiones Generales *24*

Recomendaciones *25*

Referencias *26*

Participación en eventos *30*

Anexos 31

Anexo 1: *Pseudocódigo para el cálculo de un modelo de clasificación por PCA, SPCA y PLS.* *31*

Anexo 2: *Gráficas del estadístico de prueba de la razón de verosimilitud para estimar el parámetro* θ. 32

Anexo 3: *Pseudocódigo para la propuesta* ***Doble reducción de la dimensión.*** 36

Índice General

Anexo 4: *Pseudocódigo para la propuesta* ***Doble reducción de la dimensión en forma paralela****.* 37
Anexo 5: *Matrices de confusión para el modelo PCA.* 39
Anexo 6: *Matrices de confusión para el modelo SPCA.* 40
Anexo 7: *Matrices de confusión para el modelo PLS.* 41
Anexo 8: *Matrices de confusión para el modelo CUR-PLS.* 42
Anexo 9: *Matrices de confusión para el modelo CUR-PLS-Par.* 43

Introducción

Durante las pasadas dos décadas, el advenimiento de conjuntos de datos de microarrays de ADN [1] han estimulado una nueva línea de investigación tanto en bioinformática como en aprendizaje de máquinas. Estos tipos de datos son utilizados para coleccionar información sobre tejidos y muestras de células respecto a diferentes expresiones de genes que puede ser útil para diagnosticar enfermedades o para distinguir tipos específicos de tumor. Aunque son empleadas muestras muy pequeñas (a menudo menor que 100 pacientes) para entrenamientos y pruebas, el número de características crece exponencialmente alcanzado de 6000 a 60000, cifras que crean una alta probabilidad de encontrar "negativos falsos" y "positivos falsos", representando un reto para los métodos estadísticos tradicionales [2].

Para resolver estos problemas de la alta dimensión se destacan los métodos de selección y extracción de características, siendo este último de interés en esta investigación. Los métodos de selección de características trabajan eliminando las características que son irrelevantes y redundantes, encontrándose usualmente tres variantes: métodos de filtros, envolventes y embebidos. En cambio, los métodos de extracción de características pueden ser tanto lineales como no lineales, centrándose en la construcción de nuevas variables que contengan la mayor información posible de las variables originales y sean a su vez, mucho más pequeñas.

Ambos métodos se aplican a un sólo conjunto de datos de ahí que se diga que estos trabajen de forma centralizada. Sin embargo, si se pudiera distribuir el conjunto de datos en diferentes subconjuntos y luego aplicar un método a cada uno de ellos combinando sus resultados se obtendría una reducción considerable en el tiempo de ejecución, una mejor interpretación en los datos y la precisión de clasificación no se vería afectada en exceso. La utilización de algoritmos aleatorios permitiría realizar tal distribución en forma paralela.

Por tal razón, se propone como problema de investigación:

¿Cómo reducir la dimensión en datos de microarrays de ADN para el cáncer que permita obtener un procedimiento para diagnosticar esta enfermedad?

Para resolver el problema de investigación, se plantea el siguiente objetivo general:

Reducir la dimensión en datos de microarrays de ADN para el cáncer mediante la distribución del conjunto de datos en diferentes subconjuntos a través de algoritmos aleatorios para obtener un procedimiento que diagnostique esta enfermedad.

Para dar cumplimiento al objetivo general, se trazan los siguientes objetivos específicos:

- Caracterizar los métodos de extracción de características lineales en datos de microarrays de ADN para el cáncer que traten el estudio de distinguir entre muestras cancerígenas y no cancerígenas.
- Caracterizar los algoritmos aleatorios en la aplicación de datos de microarrays de ADN para enfermedades oncológicas.
- Proponer una metodología que permita distribuir estos conjuntos de datos en diferentes subconjuntos y luego aplicar un método de extracción de característica lineal a cada uno de ellos combinando sus resultados para obtener un modelo de clasificación que pronostique esta enfermedad.
- Implementar la metodología propuesta a través de una función en el entorno de desarrollo integrado RStudio para que pueda ser empleada en el software R.

El resto del trabajo de diploma está estructurado como sigue. En el Capítulo 1 se describen los conjuntos de microarrays que se investigan. En el Capítulo 2 se caracterizan tres métodos de extracción de características lineales a partir de su fundamentación teórica, selección del número de componentes latentes y ordenamiento de las variables con respecto a la componente latente. Además, se muestra la metodología a proponer, resultados y discusión. Por último, se dan las conclusiones de este capítulo.

En el Capítulo 3 se caracterizan los algoritmos aleatorios en datos de microarrays de ADN, en particular la descomposición matricial CUR. Se proponen dos metodologías que utilizan la descomposición matricial anterior para la obtención de un modelo de clasificación. Presenta los resultados y la discusión de los mismos. Por último, se dan las conclusiones del capítulo.

1. Conjuntos de datos de microarrays de ADN para el cáncer

En este trabajo se estudian 6 de los 9 conjuntos de datos microarrays binarios estudiados por [3]. En la **Tabla 1** se da una descripción de estos conjuntos de datos.

Los conjuntos de datos Colon, DLBCL y Ovarian fueron descargados del repositorio *Kent Ridge Bio-Medical Repository*, from the Agency for Sciency , Technology and Research [4], el conjunto de datos CNS/Embrional-T fue descargado del repositorio *Dataset Repository*, from the Bioinformatics Research Group of Universidad Pablo de Olavide [5] mientras que los conjuntos de datos GLI-85 y SMK-CAN-187 fueron descargados del repositorio *Feature Selection Dataset*, from Arizona State University [6].

Tabla 1: Descripción de conjuntos de datos binarios

Conjunto de datos	n	p	IR	F1	Referencia Original
CNS/Embrional-T	60	7129	1,86	0,45	[7]
Colon	62	2000	1,82	1,08	[8]
DLBCL	47	4026	1,04	2,91	[9]
GLI-85	85	22,283	2,27	2,35	[10]
Ovarian	253	15,154	1,78	6,94	[11]
SMK-CAN-187	187	19,993	1,08	0,41	[12]

Los parámetros n y p corresponden al número de muestras y genes respectivamente, en tanto, IR representa la tasa de desbalance definida como la cantidad de muestras de clases negativas dividido por la cantidad de muestras de clases positivas. Consecuentemente, F1 simboliza la máxima de las tasas discriminantes de Fisher [13] que puede ser calculada a través de la siguiente relación:

$$F1 = \max_{i,j=1,\dots n} \left\{ \frac{(\mu_i - \mu_j)^2}{\sigma_i^2 + \sigma_j^2} \right\} \; donde \; n = p \qquad (1)$$

Capítulo 1

1.1. Características de los datos microarray

Una de las características más comunes que presentan los datos microarray es la alta dimensión de sus datos, comúnmente conocido como problema "large p small n" lo que representa un reto para los métodos estadísticos tradicionales resultando difícil o imposible su aplicación. Además, existen otras características que hacen que la clasificación de datos microarray sea un desafío aún mayor para las técnicas computacionales tales como el desbalance de las clases, la complejidad de los datos o solapamiento de las clases, la presencia de datos "shift" y "outlier" así como datos faltantes [13].

1.2. Conclusiones del Capítulo

- Se muestran los conjuntos de datos de microarrays de ADN que serán de estudio en esta investigación.
- Los parámetros p, IR y F1 evidencian los problemas de alta dimensión, desbalance y solapamiento de las clases, respectivamente, en estos conjuntos de datos.
- La presencia de datos "shift", "outlier" y faltantes hacen que la clasificación de datos microarray sea un desafío aún mayor para las técnicas computacionales.

2. Métodos de extracción de características lineales en datos de microarrays de ADN

En este capítulo se presentan tres métodos de extracción de características lineales. El primero de ellos es el Análisis de Componentes Principales (PCA, por sus siglas en inglés), que es un método no supervisado, mientras que el Análisis de Componentes Principales Supervisado (SPCA, por sus siglas en inglés) y los Mínimos Cuadrados Parciales (PLS, por sus siglas en inglés) son supervisados.

2.1. Análisis de Componentes Principales

La idea central del PCA es reducir la dimensión de un conjunto de datos de variables interrelacionadas, mientras se conserve la mayoría de la información del presente conjunto. Esto se puede lograr construyendo nuevas variables que sean incorrelacionadas y ordenadas de modo tal, que las primeras absorban gran parte de la variabilidad total, siendo esta definida por la varianza de las variables creadas. Este método de extracción de características se describe en la totalidad de los libros de textos de análisis multivariante[14].

Según [15], PCA es obtenido de la descomposición propia de la matriz de covarianza o correlación. Esta descomposición es empleada solamente en algunas matrices cuadradas, fundamentalmente en matrices semi-definidas positivas. Una descomposición similar se aplica a toda matriz rectangular de valores reales: la descomposición del valor singular (SVD, por sus siglas en inglés). Este procedimiento se muestra a continuación: Sea X una matriz $n \times p$, y sea r el rango de X, luego se puede encontrar matrices U, Σ y V con las siguientes propiedades:

- $U_{n\times r}$ es una matriz cuyas columnas son los vectores propios (normalizados) de la matriz XX^t. Estos se catalogan como los vectores singulares izquierdos de X.
- $V_{p\times r}$ es una matriz cuyas columnas son los vectores propios (normalizados) de la matriz X^tX. Estos se identifican como los vectores singulares derechos de X.

- $\Sigma_{r\times r}$ es una matriz diagonal cuyos elementos de la diagonal principal tienen la forma $\lambda^{1/2}$, siendo λ valor propio de la matriz X^tX. Los elementos de Σ son llamados los valores singulares de X.

De manera que,

$$X = U\Sigma V^t \tag{2}$$

De la ecuación (1), se obtienen las componentes principales,

$$U = XV\Sigma^{-1} = XW \tag{3}$$

Dentro de los procedimientos para escoger cuántas componentes serán consideradas se encuentra la de seleccionar aquellas cuyo valor propio excede al promedio, es decir,

$$\lambda_k > \frac{1}{r}\sum_{k}^{r}\lambda_k \tag{4}$$

donde λ_k es el $k-ésimo$ valor propio correspondiente a la $k-ésima$ componente principal y r representa el rango de la matriz de datos en cuestión. En el caso de PCA usando la matriz de correlación se optan por las que tienen sus valores propios mayores que 1, sin embargo, esta técnica puede conducir a ignorar información importante [15]. Además, existe la posibilidad de elegir las componentes cuya varianza acumulativa explique un $Q100\%$ que garantice un alto comportamiento de la variabilidad total.

$$\frac{\sum_{i=1}^{k}\lambda_{\mathrm{i}}}{\sum_{j=1}^{p}\lambda_{\mathrm{j}}} \leq Q \tag{5}$$

En PCA, la correlación entre una variable y una componente se denomina "carga". Debido a que la suma de los cuadrados de los coeficientes de correlación entre una variable y todas las componentes es igual a 1, ocurre que las cargas al cuadrado otorgan la proporción de varianza de las variables explicadas por las componentes. Esta se utiliza para establecer un orden entre las variables según su trascendencia en un modelo de clasificación por PCA, a la cual denominamos como factor de importancia:

$$imp_j = \sum_{i=1}^{k} cor^2(x_j,u_i), \forall j = 1,\dots,p \tag{6}$$

2.2. Análisis de Componentes Principales Supervisados

El SPCA fue propuesto por [16] para problemas de regresión donde el número de variables es mucho mayor que el número de muestras. En lo adelante, se asumen a X e y

son matrices de orden $n \times p$ y $n \times 1$, respectivamente, siendo n las muestras u observaciones y p las variables o características. A continuación, se presenta el algoritmo de componentes principales supervisadas:

1. Calcular los coeficientes de regresión estandarizado univariante para cada característica.
2. Formar una matriz de datos reducidos consistiendo de solamente de aquellas características cuyo coeficiente univariante excede a un umbral θ en valor absoluto (θ es estimado por validación cruzada).
3. Calcular la primera (o primeras pocas) componentes principales de la matriz de datos reducidos.
4. Utilizar las componentes principales en un modelo de regresión para predecir el resultado.

Para desarrollar detalladamente este método se asume que las columnas de X (variables) estén centradas. Seguidamente se considera la descomposición del valor singular de X como se define en la sección 2.1.

Sea s el vector de p componentes que representa los coeficientes de regresión estandarizados para medir el efecto univariante de cada característica separadamente en Y:

$$s_j = \frac{x_j^t Y}{\|x_j\|} \tag{7}$$

con

$$\|x_j\| = \sqrt{x_j^t x_j} \tag{8}$$

Luego de haber calculado el vector s, se forma la matriz de datos reducidos X_θ, por aquellas variables de X que cumplan la condición $|s_j| > \theta$. A partir de esta matriz, se calculan las componentes principales supervisadas por medio de la descomposición del valor singular:

$$X_\theta = U_\theta \Sigma_\theta V_\theta^t \tag{9}$$

$$U_\theta = X_\theta \mathrm{V}_\theta \Sigma_\theta^{-1} = X_\theta \mathrm{W}_\theta \tag{10}$$

cuyas componentes se encuentran en la matriz de los vectores singulares izquierdos $U_\theta = (u_{\theta,1}, u_{\theta,2}, \ldots, u_{\theta,r})$, siendo $u_{\theta,1}$ la primera componente, $u_{\theta,2}$ la segunda componente así sucesivamente.

La primera componente principal supervisada $u_{\theta,1}$ se utiliza para ajustar un modelo de regresión lineal simple con la variable respuesta Y,

$$\hat{Y}^{spc,\theta} = \bar{Y} + \hat{\gamma} u_{\theta,1} \tag{11}$$

donde $\hat{\gamma} = u_{\theta,1}^t Y$ debido a que $u_{\theta,1}$ es un vector singular izquierdo de X_θ. Aunque se utiliza por lo general $u_{\theta,1}$ para ajustar este modelo, existe la posibilidad de utilizar más de una componente principal supervisada.

Teniendo en cuenta que las variables que pertenecen a la matriz de datos reducidos no son necesariamente importantes, se utiliza un factor de importancia basado en la correlación entre cada característica y $u_{\theta,1}$:

$$imp_j = cor(x_j, u_{\theta,1}), \forall j = 1, \ldots, p \tag{12}$$

Mientras mayor sea el valor absoluto de imp_j, mayor será la contribución de la variable x_j en la predicción de Y [17].

2.3. Mínimos Cuadrados Parciales.

En 1996 Wold propone el método PLS [18] y en su forma original se asociaban a los sistemas de ecuaciones estructurales (SEM, por sus siglas en inglés) [19]. La idea que perseguía Wold era dotar a la práctica estadística de una alternativa analítica para aquellas situaciones en que no se tenían las hipótesis básicas de la modelación estadística.

Años más tarde, se da una nueva formulación a PLS, llamada, regresión PLS [20-22] cuyas ideas se presentan a continuación.

Se supone que existen q variables $Y_1, \ldots, Y_q$ dependientes de p variables independientes $X_1, \ldots, X_p$. Se dispone de n observaciones y se desea ajustar un modelo de regresión. Los datos se resumen en forma matricial: $Y_{n\times q}$ y $X_{n\times p}$, respectivamente.

Una característica que tiene la regresión PLS es que se puede aplicar en situaciones donde el número de individuos, n sea menor que el número de variables, p. Esto no pasa

con las técnicas de regresión usual tales como la regresión clásica ya que la matriz de covarianza $\mathrm{X}^t X$ es singular.

La idea básica es hallar una descomposición en factores latentes T tal que:

$$\begin{aligned} \mathrm{Y} &= \mathrm{T}Q^t + F \\ \mathrm{X} &= \mathrm{T}P^t + E \end{aligned} \tag{13}$$

Donde T es una matriz de $n \times c$, que contiene las componentes latentes de las n observaciones. Por su parte, P, de $p \times c$, y Q, de $q \times c$, son matrices de coeficientes. E y F, son matrices de errores aleatorios de dimensiones $n \times p$ y $n \times q$, respectivamente.

PLS es un método que construye la matriz T para obtener una transformación lineal de X

$$T = XW \tag{14}$$

siendo W una matriz de ponderación orden $p \times c$. Esta matriz se obtiene a partir de la maximización del cuadrado de la covarianza entre la componente latente y la variable dependiente, sujeto a la restricción $w^t w = 1$ [21].Para esto, se utilizan varios algoritmos, entre ellos están NIPALS [23], KERNEL-PLS [24, 25], y SIMPLS [26]. En los paquetes pls [27], nipals [28] y plsgenomics [29] están implementados estos algoritmos.

Una vez obtenida la matriz T, se utiliza en la regresión en lugar de la matriz original. Finalmente, el modelo se expresa en las variables originales, haciendo la transformación "inversa". Esto es:

$$Q^t = (T^t T)^{-1} T^t Y \tag{15}$$

que no es más que la matriz de coeficientes para el modelo transformado. Al multiplicar Q^t por T, se obtiene la matriz de los coeficientes asociados a las variables originales:

$$B = WQ^t \tag{16}$$

El criterio para la selección del número de componentes es la minimización de la suma de cuadrados de los residuos. Los criterios más empleados son:

- Estimación de la suma de cuadrados de los residuos mediante validación cruzada
- Estimación de la suma de cuadrados de predicción PRESS (por sus siglas en inglés: **Pre**diction **S**um of **S**quares)

Al trabajar con el algoritmo SIMPLS en un problema de clasificación que estudia distinguir muestras cancerígenas y no cancerígenas, el vector de pesos $w_1 = (w_{11}, \ldots, w_{p1})^t$, que define la primera componente latente, puede ser empleado en el

ordenamiento de las variables de acuerdo a su relevancia en el modelo de clasificación. Este ordenamiento está propuesto en [30], donde demuestra que el estadístico-$\mathrm{BSS}_j/\mathrm{WSS}_j$ [31] es una función de ${w_1}^2$, la cual es estrictamente monótona. En este reporte se trabaja con la función que se encuentra en el paquete plsgenomics dada por:

$$f({w_1}^2) = -\sqrt{{w_1}^2} = -|w_1| \tag{17}$$

2.4. Metodología propuesta

La metodología que se propone a continuación consiste de tres etapas:

Etapa I: Reducción de la dimensión.

Etapa II: Construcción del modelo de clasificación.

Etapa III: Validación del modelo de clasificación.

En la primera etapa se reduce la dimensión de los datos empleando los métodos de extracción de características lineales siguientes:

- Análisis de Componentes Principales.
- Análisis de Componentes Principales Supervisados.
- Mínimos Cuadrados Parciales.

Luego de haber obtenido las k-ésimas componentes mediante los métodos anteriores, se construye un modelo de clasificación por Análisis Discriminante Lineal (LDA, por sus siglas en inglés) utilizándose, en dependencia del conjunto de datos, los siguientes criterios:

- Para los conjuntos de microarrays que no presentan conjunto de prueba se utiliza validación cruzada dejando uno fuera.
- Para los conjuntos de microarrays que tienen conjunto de prueba se construye el modelo de clasificación con el conjunto de entrenamiento y se emplea validación "holdout" [13] para el conjunto de prueba.

Por último, se obtiene la matriz de confusión para cada método de extracción y se emplean las medidas de sensibilidad (Se), especificidad (Es) y exactitud (Ex) para determinar cuan bueno es el modelo de clasificación.

Estas, se describen en términos de positivos verdaderos (PV), negativos verdaderos (NV), negativos falsos (NF) y positivos falsos (PF):

$$Se = \frac{PV}{PV + NF}; 0 \leqslant Se \leqslant 1 \quad (18)$$

$$Es = \frac{NV}{NV + PF};\ 0 \leqslant Es \leqslant 1 \quad (19)$$

$$Ex = \frac{PV + NV}{PV + NV + NF + PF};\ 0 \leqslant Ex \leqslant 1 \quad (20)$$

La sensibilidad y especificidad son medidas que permiten indicar que el modelo de clasificación soluciona el problema del desbalance de las clases, mientras que la exactitud muestra que dicho modelo enmienda el problema de la complejidad de los datos.

2.5. Resultados.

La metodología que se propone está implementada en el software R [32], la cual contempla las etapas siguientes: (i) Reducción de la dimensión, (ii) Construcción del modelo de clasificación y (iii) Validación del modelo de clasificación. En el **Anexo 1** se muestra el pseudocódigo para dicha implementación.

En la primera etapa se calculan k componentes para reducir la dimensión de los datos, siendo estas, combinaciones lineales de las variables originales. Para seleccionar dichas componentes, PCA utiliza el criterio dado por la ecuación (5) tomando como valor $Q = 0.75$, por su parte, PLS lo hace mediante una propuesta por validación cruzada de [30, 31] implementada en el paquete plsgenomics. En el caso de SPCA, es necesario estimar el parámetro θ por validación cruzada K-campos para determinar las componentes. Para los conjuntos Colon, DLBCL y SMK se estima θ por validación cruzada 5 campos mientras que para los conjuntos Ovarian y GLI-85 se usa validación cruzada 10 campos. En el paquete superpc [33] se utiliza el estadístico de prueba de la razón de verosimilitud para estimar el parámetro θ. Se puede apreciar en el **Anexo 2** como este estadístico de prueba para el parámetro θ estimado es significativo en los conjuntos Colon, DLBCL, Ovarian, GLI-85 y SMK a diferencia de lo que sucede con el conjunto CNS. En la **Tabla 2** se muestra el número de componentes latentes para los métodos PCA, SPCA y PLS, respectivamente, resolviendo el problema de la reducción de la dimensión.

Tabla 2: Número de componentes latentes para PCA, SPCA y PLS

Métodos	Colon	DLCBL	CNS	Ovarian	GLI85	SMK
PCA	6	18	19	3	35	23
SPCA	2	1	-	1	1	1
PLS	5	1	1	5	4	2

En la segunda etapa se calcula un modelo de clasificación por LDA tomando como variables predictoras las k componentes obtenidas por los métodos PCA, SPCA y PLS. Para esto se utiliza el paquete MASS [34].

En la tercera etapa se obtiene la matriz de confusión para cada método de extracción y se emplean las medidas de sensibilidad (Se), especificidad (Es) y exactitud (Ex) para determinar cuan bueno es el modelo de clasificación. En la **Tabla 3** y la **Figura 1** se muestran dichas medidas

Tabla 3: Resultados para el LDA en los conjuntos de datos.

Métodos	Medidas	Colon	DLCBL	CNS	Ovarian	GLI85	SMK
PCA	Ex	0.82	0.87	0.52	0.77	0.87	0.72
	Se	0.85	0.79	0.74	0.89	0.92	0.78
	Es	0.77	0.96	0.1	0.57	0.77	0.64
SPCA	Ex	0.87	0.98	-	0.97	0.96	**0.73**
	Se	0.93	0.96	-	**1**	0.98	**0.80**
	Es	0.77	1	-	0.91	**0.92**	**0.66**
PLS	Ex	**0.92**	**1**	**0.72**	**0.99**	**0.98**	**0.73**
	Se	**0.95**	**1**	**0.97**	0.99	**1**	**0.80**
	Es	**0.86**	**1**	**0.24**	**0.99**	**0.92**	0.64

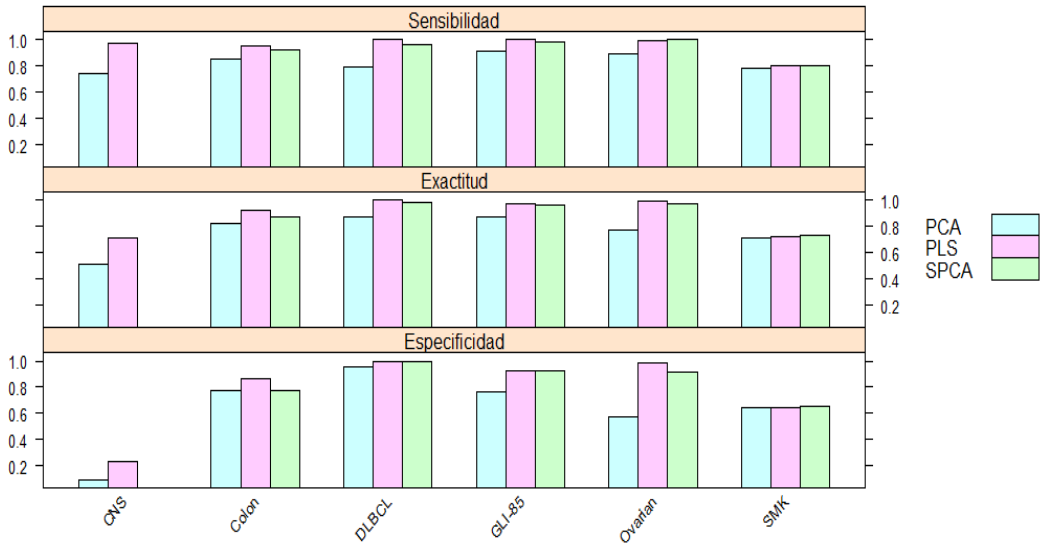


Figura 1: Resultados para el LDA en los conjuntos de datos.

2.6. Discusión.

En la sección anterior se implementan tres modelos de clasificación para solucionar los problemas de la alta dimensión, desbalance de las clases y el solapamiento de los datos. Estos modelos son aplicados a los conjuntos de datos de microarray presentados en el epígrafe 3.

La **Tabla 3** muestra los valores de las medidas de sensibilidad, especificidad y exactitud en PCA, SPCA y PLS para cada conjunto. En la misma se puede observar que en PCA se aprecian conjuntos como el CNS, que presenta un 10% en especificidad, indicando que este modelo ostenta un gran problema a la hora de predecir los "negativos verdaderos" (pacientes sanos que son correctamente identificados), y los "positivos verdaderos" (pacientes enfermos que son correctamente identificados), pues este conjunto de datos presenta un 74% en sensibilidad. La exactitud del PCA fue de 52%, siendo este muy bajo, por lo tanto, se pude decir que el PCA es ineficiente para darle solución al problema de solapamiento.

Por su parte en el método SPCA se obtienen valores de 100% en especificidad y sensibilidad. Un ejemplo de máximo por ciento en especificidad se refleja en el conjunto DLCBL, demostrando con ello una predicción exacta en cuanto a pacientes que no padecen la enfermedad. En tanto, el conjunto Ovarian identifica correctamente a los individuos que se encuentran en el grupo "enfermos". Sin embargo, el conjunto

SMK consta de resultados que se hallan distantes del por ciento ideal, afectando en gran medida la predicción de este modelo. Algo muy similar sucede para el mismo conjunto en PLS.

Por tanto, los datos cuyo valor de F1 sea bajo, es decir, que presentan un alto nivel de solapamiento, influirían de forma negativa en los modelos predictores.

A modo de comparación, los resultados obtenidos por PCA son mucho más discretos que los de SPCA y PLS, mientras que este último excede en exactitud a los anteriores presentando en determinados conjuntos valores por encima del 95% e incluso del 100% en el caso de DLCBL. Esto puede ser observado claramente en la **Figura 1**.

2.7. Conclusiones del Capítulo.

- En este trabajo de diploma se presentó una caracterización de varios métodos de extracción de características lineales tales como PCA, SPCA y PLS respectivamente. Estos se centran en la construcción de nuevas variables de forma tal, que un número reducido de ellas resuman los datos originales tanto como sea posible.
- La metodología propuesta evidencia mediante tres etapas el procedimiento a seguir en esta investigación.
- El resultado obtenido ilustra la validación de tres modelos de acuerdo a las medidas de sensibilidad, especificidad y exactitud. Además, a partir de estas métricas se considera el modelo más indicado en la predicción de muestras cancerígenas y no cancerígenas.

3. Algoritmos aleatorios en datos de microarrays de ADN.

Los algoritmos aleatorios para problemas matriciales muy grandes han recibido gran atención en los años recientes, principalmente en el análisis de microarrays de ADN. Estos algoritmos se refieren a una clase de algoritmos de proyección aleatoria y de muestreo aleatorio desarrollados recientemente [35]. Los algoritmos de proyección aleatoria resuelven el problema de aproximación de mínimos cuadrados mientras que los algoritmos de muestreo aleatorio resuelven el problema de aproximación de matrices de bajo rango [35]. Este último, es de nuestro interés en la investigación en el cual la descomposición matricial CUR [36] juega un papel muy importante.

3.1. Descomposición matricial CUR.

La descomposición matricial CUR se emplea para nombrar aquellas descomposiciones matriciales de bajo rango que son explícitamente expresadas en términos de un número pequeño de columnas y/o filas de una matriz de datos, A. Estas descomposiciones permiten aproximar la matriz A por medio del producto de tres matrices C, U y R donde C y R contienen algunas columnas y filas de A, respectivamente, mientras U es una matriz que se construye cuidadosamente de manera que garantice dicha aproximación.

Se conocen varias descomposiciones CUR que se diferencian en las cotas de error obtenidas y en el criterio para elegir las columnas y filas que forman las matrices C y R [37-41].

En [38] se propone la descomposición matricial CUR, la cual elige las columnas a incluir en C (similarmente en R) a partir de un factor de importancia para cada columna de la matriz A. Dicho factor se define a partir de la matriz A y un parámetro de entrada dado por el rango k, como se muestra a continuación:

$$\pi_j = \frac{1}{k}\sum_{p=1}^{k}(v_j^p)^2, \forall j = 1,\ldots,n \qquad (1)$$

donde v_j^p es la j-ésima componente del p-ésimo vector singular derecho de A.

A continuación, se muestrea aleatoriamente un pequeño número de columnas de A usando ese factor de importancia como una distribución de probabilidad.

El algoritmo básico para seleccionar las columnas de una matriz denominado ColumnSelect [38] toma como entrada cualquier matriz de orden $m \times n$, un parámetro de rango k y un parámetro de error ϵ.

El resultado teórico más importante que avala dicho algoritmo establece que con probabilidad al menos del 99%, esta elección de columnas satisface que

$$\|A - P_C A\|_F \leqslant \left(1 + \frac{\epsilon}{2}\right) \|A - A_k\|_F \qquad (2)$$

donde $P_C A$ denota la matriz de proyección sobre el espacio columna generado por C y A_k es la matriz de rango k más próxima a A en norma de Frobenius (ver en [42] la demostración).

De esta forma el resultado garantiza que si A es una matriz cercana a una matriz de rango k entonces, con alta probabilidad, el subespacio generado por las columnas de A está próximo al subespacio generado por las columnas de C. Esto justifica el hecho de poder utilizar un método en forma paralela distribuyendo la matriz de datos A en varias matrices C [43].

En [44] se mencionan los métodos experimentales "random", "exact.num.random", "top.scores", "ortho.top.scores" y "highest.ranks" los cuales se encuentran implementados en el paquete rCUR [45]. En dicho trabajo comentan que tales métodos proporcionan la misma precisión que el algoritmo ColumSelect.

3.2. Metodología propuesta

En esta sección se proponen dos metodologías que utilizan la descomposición matricial CUR para la obtención de un modelo de clasificación. La primera reduce la dimensión de los datos por CUR y luego emplea los métodos PCA, SPCA y PLS. Esta metodología es una continuación de la metodología presentada en el epígrafe 2.4. En lo adelante, cuando se refiera a esta metodología se hará por **Doble reducción de la dimensión**. La segunda es una realización en forma paralela de **Doble reducción de la dimensión**.

3.2.1. Doble reducción de la dimensión.

A continuación, se presenta la metodología que consiste de cinco etapas para la obtención un modelo de clasificación:

Etapa I: Reducción de la dimensión por CUR.

Etapa II: Reducción de la dimensión por métodos de extracción de características.

Etapa III: Construcción del modelo de clasificación.

Etapa IV: Validación del modelo de clasificación.

Etapa V: Ordenamiento de las variables respecto a la componente latente.

En la primera etapa se reduce la dimensión de los datos mediante la descomposición matricial CUR para obtener la matriz C.

A partir de la matriz C se reduce la dimensión empleando los métodos de extracción de características lineales siguientes:

- Análisis de Componentes Principales.
- Análisis de Componentes Principales Supervisados.
- Mínimos Cuadrados Parciales.

Luego de obtener las k-ésimas componentes mediante los métodos anteriores, se construye un modelo de clasificación por Análisis Discriminante Lineal (LDA, por sus siglas en inglés), utilizando en dependencia del conjunto de datos los siguientes criterios:

- Para los conjuntos de microarrays que no presentan conjunto de prueba se utiliza validación cruzada dejando uno fuera.
- Para los conjuntos de microarrays que tienen conjunto de prueba se construye el modelo de clasificación con el conjunto de entrenamiento y se emplea validación "holdout" [13] para el conjunto de prueba.

Posteriormente se obtiene la matriz de confusión para cada método de extracción y se emplean las medidas de sensibilidad (Se), especificidad (Es) y exactitud (Ex) para determinar cuan bueno es el modelo de clasificación.

Estas, se describen en términos de positivos verdaderos (PV), negativos verdaderos (NV), negativos falsos (NF) y positivos falsos (PF):

$$Se = \frac{PV}{PV + NF};\ 0 \leqslant Se \leqslant 1 \tag{3}$$

$$Es = \frac{NV}{NV + PF};\ 0 \leqslant Es \leqslant 1 \tag{4}$$

$$Ex = \frac{PV + NV}{PV + NV + NF + PF};\ 0 \leqslant Ex \leqslant 1 \tag{5}$$

La sensibilidad y especificidad son medidas que permiten indicar que el modelo de clasificación soluciona el problema del desbalance de las clases, mientras que la exactitud muestra que dicho modelo enmienda el problema de la complejidad de los datos.

En correspondencia al método de extracción de característica empleado, se ordenan las variables de acuerdo a su factor de importancia:

- PCA: $imp_j = \sum_{i=1}^{k} cor^2(x_j, u_i), \forall j = 1, \dots, p$
- SPCA: $imp_j = cor(x_j, u_{\theta,1}), \forall j = 1, \dots, p$
- PLS: $imp_j = -|w_{j1}|, \forall j = 1, \dots, p$

3.2.2. Doble reducción de la dimensión en forma paralela.

La metodología **Doble reducción de la dimensión** propuesta en la sección anterior considera un modelo que tal vez no sea el más idóneo en la predicción. Para prescindir de esta incertidumbre se presenta una metodología cuya idea central es transformar el problema de clasificación de gran dimensión en varios problemas de clasificación de menores dimensiones, distribuyendo los datos en forma paralela con el objetivo de encontrar una cantidad $k < p$ de variables que tengan una alta capacidad predictiva. Dicha metodología referida por **Doble reducción de la dimensión en forma paralela** cuenta con seis etapas, realizando las cuatro primeras en forma paralela. A continuación, se presenta dicha metodología:

Etapa I: Reducción de la dimensión por CUR.

Etapa II: Reducción de la dimensión por métodos de extracción de características.

Etapa III: Construcción de los m modelos de clasificación.

Etapa IV: Validación de los m modelos de clasificación.

Etapa V: Selección del mejor modelo.

Etapa VI: Ordenamiento de las variables respecto a la componente latente.

La primera etapa consiste en reducir la dimensión de los datos mediante la descomposición matricial CUR, obteniéndose m matrices C de modo que sus columnas satisfagan la desigualdad (2).

Una vez obtenidas las m matrices C se reduce la dimensión empleando los métodos de extracción de características lineales siguientes:

- Análisis de Componentes Principales.
- Análisis de Componentes Principales Supervisados.
- Mínimos Cuadrados Parciales.

Luego de obtener las k-ésimas componentes mediante los métodos anteriores, se construye un modelo de clasificación por LDA a cada matriz de datos reducidos, utilizando en dependencia del conjunto de datos los siguientes criterios:

- Para los conjuntos de microarrays que no presentan conjunto de prueba se utiliza validación cruzada dejando uno fuera.
- Para los conjuntos de microarrays que tienen conjunto de prueba se construye el modelo de clasificación con el conjunto de entrenamiento y se emplea validación "holdout" [13] para el conjunto de prueba.

Posteriormente se obtiene la matriz de confusión para el método de extracción empleado en cada uno de los m modelos, y se utilizan las medidas de sensibilidad (Se), especificidad (Es) y exactitud (Ex) para determinar cuan bueno es el modelo de clasificación.

A partir de la validación de cada modelo se selecciona el de menor error de mala clasificación, definido como:

$$E = \min\{E_i\}_{i=1}^{m} = \min\{1 - Ex_i\}_{i=1}^{m} \qquad (6)$$

Por tanto, este modelo presenta el mayor por ciento en exactitud.

En correspondencia al método de extracción de característica empleado, se ordenan las variables de acuerdo a su factor de importancia:

- PCA: $imp_j = \sum_{i=1}^{k} cor^2(x_j, u_i), \forall j = 1, \dots, p$
- SPCA: $imp_j = cor(x_j, u_{\theta,1}), \forall j = 1, \dots, p$
- PLS: $imp_j = -|\mathrm{w}_{j1}|, \forall j = 1, \dots, p$

3.3. Resultados.

En la sección 2.6 se realizó una comparación entre los resultados obtenidos por los modelos PCA, SPCA y PLS, siendo este último superior a los primeros en cuanto a los

valores de sensibilidad, especificidad y exactitud. Por esta razón, en esta sección se implementan en el software R [32] las metodologías **Doble reducción de la dimensión** y **Doble reducción de la dimensión en forma paralela** utilizando el método PLS, denotadas como CUR-PLS y CUR-PLS-Par, respectivamente. En los **Anexos 3** y **4** se muestran el pseudocódigo para ambas implementaciones.

En la primera etapa que contempla la reducción de la dimensión por CUR, se utiliza el 10% del total de variables en cada conjunto de datos para determinar la matriz C. Con este fin, CUR-PLS selecciona aquellas variables con factores de importancia mayores por medio del método "top.scores". Para la ejecución del CUR-PLS-Par se crea un clúster de 4 procesadores a través de los paquetes foreach [46] y doSNOW [47] calculando 100 matrices C en forma paralela mediante el método "random".

Una vez realizada la primera etapa, se procede a calcular k componentes latentes por PLS para reducir la dimensión. Para ello, se trabaja con una propuesta por validación cruzada de [30, 31] implementada en el paquete plsgenomics [29].

Luego de calcular las k componentes se pasa a la obtención un modelo de clasificación por LDA. Para lograr esto, CUR-PLS toma como variables predictoras estas k componentes obtenidas a partir de la matriz C. En cambio, CUR-PLS-Par lo hace de un modo diferente. A partir del clúster de 4 procesadores se calculan en forma paralela 100 modelos tomando como variables predictoras las k componentes obtenidas en las 100 matrices C. Seguido de esto, se realiza una validación cruzada deja-uno-fuera para cada modelo y se selecciona aquel que tenga el menor error de mala clasificación. Para la obtención de los modelos se utiliza el paquete MASS [34]. En la **Tabla 4** se muestra el número de componentes para los modelos PCA, SPCA, PLS, CUR-PLS y CUR-PLS-Par. Además, se evidencia cómo se resuelve el problema de la reducción de la dimensión.

Tabla 4: Número de componentes latentes para los modelos

Métodos	Colon	DLCBL	CNS	Ovarian	GLI85	SMK
PLS	5	1	1	5	4	2
CUR-PLS	1	1	1	7	2	1
CUR-PLS-Par	6	1	3	6	4	2

Por último, se valida el modelo de clasificación obtenido por CUR-PLS, así como CUR-PLS-Par empleando las medidas de sensibilidad (*Se*), especificidad (*Es*) y exactitud (*Ex*) para determinar cuan bueno es el modelo de clasificación. En la **Tabla 5** y la **Figura 2** se muestran dichas medidas

Tabla 5: Resultados para el LDA en los conjuntos de datos.

Métodos	Medidas	Colon	DLCBL	CNS	Ovarian	GLI85	SMK
PLS	Ex	**0.92**	**1**	0.72	0.99	0.98	0.73
	Se	**0.95**	**1**	0.97	0.99	**1**	0.80
	Es	**0.86**	**1**	0.24	0.99	0.92	0.64
CUR-PLS	Ex	0.82	0.91	0.77	0.98	0.89	0.74
	Se	0.88	0.92	0.87	0.99	0.98	0.80
	Es	0.73	0.91	0.57	0.97	0.69	0.67
CUR-PLS-Par	Ex	**0.92**	**1**	**0.93**	**1**	**0.99**	**0.75**
	Se	**0.95**	**1**	**0.92**	**1**	**1**	**0.82**
	Es	**0.86**	**1**	**0.95**	**1**	**0.96**	**0.68**

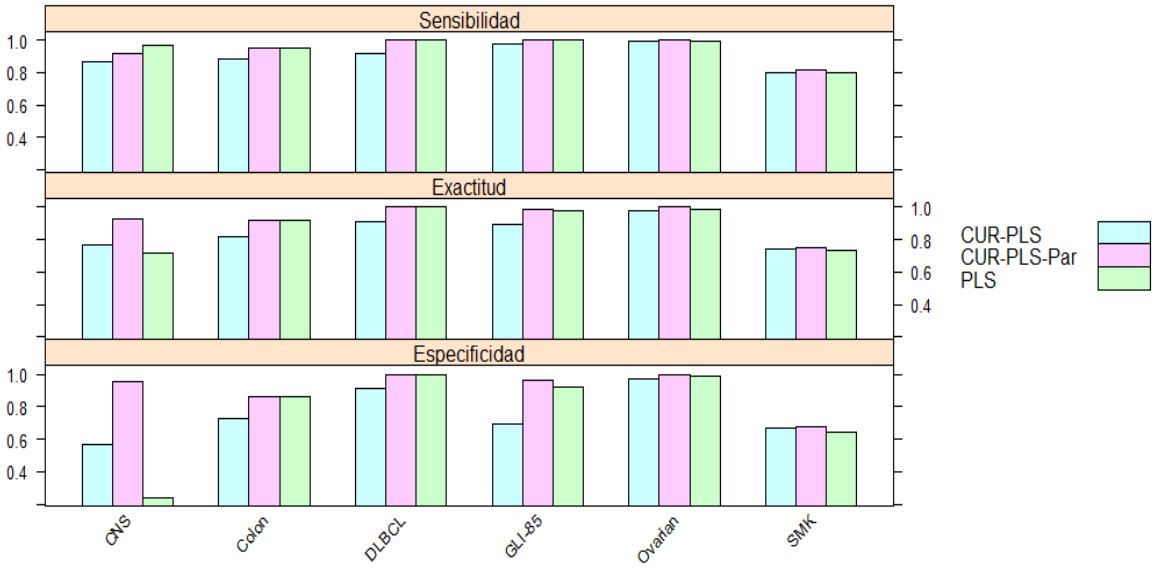


Figura 2: Resultados para el LDA en los conjuntos de datos.

3.4. Discusión.

En la **Tabla 5** se puede apreciar que el modelo CUR-PLS no presenta valores en sensibilidad, especificidad y exactitud por encima de los exhibidos por PLS, salvo en determinados conjuntos como son CNS y SMK. Esto demuestra que la metodología **Doble reducción de la dimensión** no es del todo eficiente en la predicción de enfermedades aun cuando el modelo es más simple debido a que el número de variables representa el 10% de las originales.

En el modelo CUR-PLS-Par se observan mejores resultados que el CUR-PLS, dado que en los conjuntos de datos que se estudian, las medidas empleadas para validar dicho modelo son superiores. Por su parte, el modelo PLS da resultados muy similares, incluso iguales en determinados conjuntos al modelo CUR-PLS-Par. La diferencia entre ambos modelos radica en el número de variables originales a reducir, pues PLS trabaja con todas las variables mientras que CUR-PLS-Par lo hace con el 10% de ellas.

A modo de conclusión se puede decir que el modelo CUR-PLS-Par es el mejor resolviendo los problemas de reducción de la dimensión, desbalance y solapamiento de las clases. Además, la programación paralela brinda un aporte sustancial en este resultado, pues se selecciona de 100 modelos el de mayor exactitud.

3.5. Conclusiones del Capítulo.

- Se caracterizan los algoritmos aleatorios en la aplicación de los datos de microarray de ADN en particular la descomposición CUR.
- En este reporte se proponen las metodologías **Doble reducción de la dimensión** y **Doble reducción de la dimensión en forma paralela** para la obtención de un modelo de clasificación.
- Las metodologías **Doble reducción de la dimensión** y **Doble reducción de la dimensión en forma paralela** se implementaron en el entorno de desarrollo integrado RStudio para que pueda ser empleada en el software R.
- Los resultados evidencian como el modelo CUR-PLS-Par resuelve los problemas de reducción de la dimensión, desbalance y solapamiento de las clases.

- La programación paralela brinda un aporte sustancial en la predicción de muestras cancerígenas y no cancerígenas.

- Se muestra los conjuntos de datos a emplear y sus características.
- Los métodos PCA, SPCA y PLS son caracterizados, así como la descomposición matricial CUR en la aplicación de los datos de microarray de ADN.
- Tres metodologías fueron implementadas en el entorno de desarrollo integrado RStudio para que pueda ser empleada en el software R.
- Los resultados definen cuál de los modelos presenta mayor exactitud en la predicción de enfermedades oncológicas.

- Implementar las metodologías ***Doble reducción de la dimensión*** y ***Doble reducción de la dimensión en forma paralela*** para PCA y SPCA.
- Aplicar las metodologías antes mencionadas en la presencia de datos "outlier".
- Resolver los siguientes problemas: (i) clasificar los diferentes tipos de cáncer e (ii) identificar subtipos de cáncer que pueden progresar agresivamente.

1. Hira, Z.M., *Dimensionality Reduction Methods For Microarray Cancer Data Using Prior Knowledge*, in *Department of Computing*. 2016, Imperial College London. p. 1-219.
2. Jianqing Fan, F.H.a.H.L., *Challenges of Big Data Analysis*. 2013. p. 1-38.
3. V. Bolón-Canedo, N.S.-M., A. Alonso-Betanzos, J.M. Benítez and F. Herrera, *A review of microarray datasets and applied feature selection methods.* Information Sciences, 2014. **282**: p. 111-135.
4. Dataset, K.R.B.-M. 2014.
5. Dataset Repository, B.R.G. 2014.
6. University, F.S.D.a.A.S. 2014.
7. S. Pomeroy, P.T., M. Gaasenbeek, L. Sturla, M. Angelo, M. McLaughlin, J. Kim, L. Goumnerova, P. Black, C. Lau, et al *Prediction of central nervous system embryonal tumour outcome based on gene expression.* Nature, 2002. **415**: p. 436-442.
8. U. Alon, N.B., D. Notterman, K. Gish, S. Ybarra, D. Mack, A. Levine,, *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays.* Proc. Nat. Acad. Sci, 1999. **96** p. 6745–6750.
9. A. Alizadeh, M.E., R. Davis, C. Ma, I. Lossos, A. Rosenwald, J. Boldrick, H. Sabet, T. Tran, X. Yu, et al, *Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling.* Nature, 2000. **403**: p. 503–511.
10. W. Freije, F.C.-V., Z. Fang, S. Horvath, T. Cloughesy, L. Liau, P. Mischel, S. Nelson, *Gene expression profiling of gliomas strongly predicts survival.* Cancer Res. , 2004. **64**: p. 6503–6510.
11. E. Petricoin, A.A., B. Hitt, P. Levine, V. Fusaro, S. Steinberg, G. Mills, C. Simone, D. Fishman, E. Kohn, et al, *Use of proteomic patterns in serumto identify ovarian cancer.* Lancet, 2002. **359**: p. 572–577.
12. A. Spira, J.B., V. Shah, K. Steiling, G. Liu, F. Schembri, S. Gilman, Y. Dumas, P. Calner, P. Sebastiani, et al, *Airway epithelial gene expression in the*

diagnostic evaluation of smokers with suspect lung cancer. Nat. Med., 2007. **13**: p. 361–366.

13. Bolón-Canedo, V., *Novel feature selection methods for high dimensional data*, in *Department of Computer Science*. 2014, UNIVERSITY OF A CORUÑA.
14. Jollife, I.T., *Principal Component Analysis*. 2 ed. Springer Series in Statistics, ed. Springer. 2002, New York: Springer-Verlag.
15. Williams, H.A.a.L.J., *Principal component analysis.* WIREs Computational Statistics, 2010. **2**: p. 433-460.
16. Eric Bair, T.H., Debashis Paul, and Robert Tibshirani, *Prediction by Supervised Principal Components.* Journal of the American Statistical Association, 2006. **101**: p. 119-138.
17. Jun Bin, F.-F.A., Nian Liu, Zhi-Min Zhang, Yi-Zeng Liang, Ru-Xin Shu and Kai Yang, *Supervised principal components: a new method for multivariate spectral analysis.* Journal of Chemometrics, 2013. **27**: p. 457-467.
18. Wold, H., *Estimation of principal components and related models by iterative least squares*, in *Multivariate Analysis*, P.R. Krishnaiah, Editor. 1966, Academic Press: Nueva York.
19. Joreskög, K.G., *A general method for analysis of covariance structures.* Biometrika, 1970. **57**: p. 239-251.
20. M. Sjöström, S.W., W. Lindberg, J.-A. Persson and H. Martens, *A Multivariate Calibration Problem in Analytical Chemistry Solved by Partial Least-Squares Models in Latent Variables.* Analytica Chimica Acta, 1983. **150**: p. 61-70.
21. Strimmer, A.-L.B.a.K., *Partial Least Squares: A Versatile Tool for the Analysis of High-dimensional Genomic Data.* Bioinformatics, 2007. **8**: p. 32-44.
22. Wold S, S.M.a.E.L., *PLS-regression a basic tool of chemometrics* Chemometr. Intell. Lab., 2001. **58**: p. 109–130.
23. Wold, H., *Nonlinear Iterative Partial Least Squares (NIPALS) Modelling: Some Current Developments*, in *Multivariate Analysis*, P.R. Krishnaiah, Editor. 1973, Academic Press: New York.

24. Braak, S.d.J.a.C.J.F.T., *Comments on the PLS kernel algoritm*. Journal of Chemometrics, 1994. **8**: p. 169-174.
25. Macgregor, B.S.D.a.J.F., *Improved PLS Algorithms*. Journal of Chemometrics, 1997. **11**: p. 73-85.
26. Jong, S.d., *SIMPLS: An alternative approach to partial least squares regression*. Chemometrics and Intelligent Laboratory Systems, 1993. **18**: p. 251-263.
27. Liland, B.-H.M.a.R.W.a.K.H. *pls: Partial Least Squares and Principal Component regression*. 2011; Available from: https://CRAN.R-project.org/package=pls.
28. VARMUZA, K.a.F., P., *Introduction to multivariate statistical analysis in chemometrics*. 2008, Boca Raton: CRC Press.
29. Anne-Laure Boulesteix, G.D., Sophie Lambert-Lacroix, Julie Peyre and Korbinian Strimmer. *plsgenomics: PLS Analyses for Genomics*. 2015; Available from: https://CRAN.R-project.org/package=plsgenomics.
30. Boulesteix, A.-L., *PLS Dimension Reduction for Classification with Microarray Data*. Statistical Applications in Genetics and Molecular Biology, 2004. **3**(1): p. 1-32.
31. Boulesteix, A.-L.I., *Dimension Reduction and Classication with High-Dimensional Microarray Data*, in *F akultät für Mathematik, Informatik und Statistik*. 2004, Ludwig-Maximilian-Universitä at München. p. 1-116.
32. Team, R.C., *R: A Language and Environment for Statistical Computing*, R.F.f.S. Computing, Editor. 2016: Vienna, Austria.
33. Tibshirani, E.B.a.R. *superpc: Supervised principal components*. 2012; Available from: https://CRAN.R-project.org/package=superpc.
34. Ripley, W.N.V.a.B.D., *Modern Applied Statistics with S*. Fourth ed. 2002, New York: Springer.
35. Mahoney, M.W., *Randomized algorithms for matrices and data*. 2001, Stanford University. p. 1-54.
36. Martinsson, P.-G., *Randomized methods for matrix computations and analysis of high dimensional data*. 2016. p. 1-55.

37. A. Frieze, R.K.a.S.V., *Fast Monte-Carlo algorithms for finding low-rank approximations.* J. ACM,, 2004. **51**: p. 1025–1041.
38. Drineas, M.W.M.a.P., *CUR matrix decompositions for improved data analysis.* PNAS, 2009. **106**: p. 697-702.
39. P. Drineas, R.K.a.M.W.M., *Fast Monte-Carlo algorithms for matrices III: Computing a compressed approximate matrix decomposition.* SIAM J Comput, 2006. **36**: p. 184–206.
40. Stewart, G.W., *Four algorithms for the efficient computation of truncated QR approximations to a sparse matrix*. Numer. Math, 1999. **83**: p. 313-323.
41. Tyrtyshnikov, S.A.G.a.E.E., *The maximum-volume concept in approximation by low-rank matrices.* Contemporary Mathematics, 2001. **280**: p. 47–51.
42. P. Drineas, M.W.M.a.S.M., *Relative-error CUR matrix decompositions.* SIAM J Matrix Anal Appl, 2008. **30**: p. 844-881.
43. V. Bolón-Canedo, N.S.-M.a.A.A.-B., *Distributed feature selection: An application to microarray data classification.* Applied Soft Computing, 2015. **30**: p. 136-150.
44. A. Bodor, I.C., M. W. Mahoney and N. Solymosi *rCUR: an R package for CUR matrix decomposition.* BMC Bioinformatics, 2012. **13** p. 1-6.
45. Solymosi, A.B.a.N. *rCUR: CUR decomposition package*. 2012; Available from: http://CRAN.R-project.org/package=rCUR.
46. Weston, R.A.a.S. *foreach: Foreach looping construct for R.* 2014; Available from: https://CRAN.R-project.org/package=foreach.
47. Weston, R.A.a.S. *doSNOW: Foreach parallel adaptor for the snow package*. 2014; Available from: https://CRAN.R-project.org/package=doSNOW.

- Fórum Científico Estudiantil a nivel de Facultad, obteniendo la categoría de "Relevante", el 26 de abril de 2017.
- Fórum a nivel de Universidad.

Anexo 1: *Pseudocódigo para el cálculo de un modelo de clasificación por PCA, SPCA y PLS.*

Datos a entrar:

X: Datos de microarrays.

Y: Variable dependiente binaria. La clase "1" representa que el paciente no padece la enfermedad. La clase "2" representa que el paciente si padece la enfermedad.

type: Una variable que toma los valores "PCA" y "SPCA".

Hacer:

type = "PCA"

if (type = "PCA") then

  Y_K = PCA(X)

else if (type = "SPCA") then

  Y_K = SPCA(Y, X)

else

  Y_K = PLS(Y, X)

end if

$\hat{Y}$ = LDA (Y, Y_K)

mc = matriz.confusion (Y, $\hat{Y}$)

Ex = sum(diag(mc))/sum(mc)

Se = mc [2,2]/sum(mc [2,])

Es = mc [1,1]/sum(mc [1,])

Comentarios:

PCA: Análisis de Componentes Principales.

SPCA: Análisis de Componentes Principales Supervisados.

PLS: Mínimos Cuadrados Parciales.

Y_K: K primeras componentes latentes.

LDA: Análisis Discriminante Lineal.

$\hat{Y}$: Estimación de la variable dependiente.

mc: Matriz de confusión de orden 2x2.

Ex: Exactitud.

Se: Sensibilidad.

Es: Especificidad.

Anexo 2: *Gráficas del estadístico de prueba de la razón de verosimilitud para estimar el parámetro θ.*

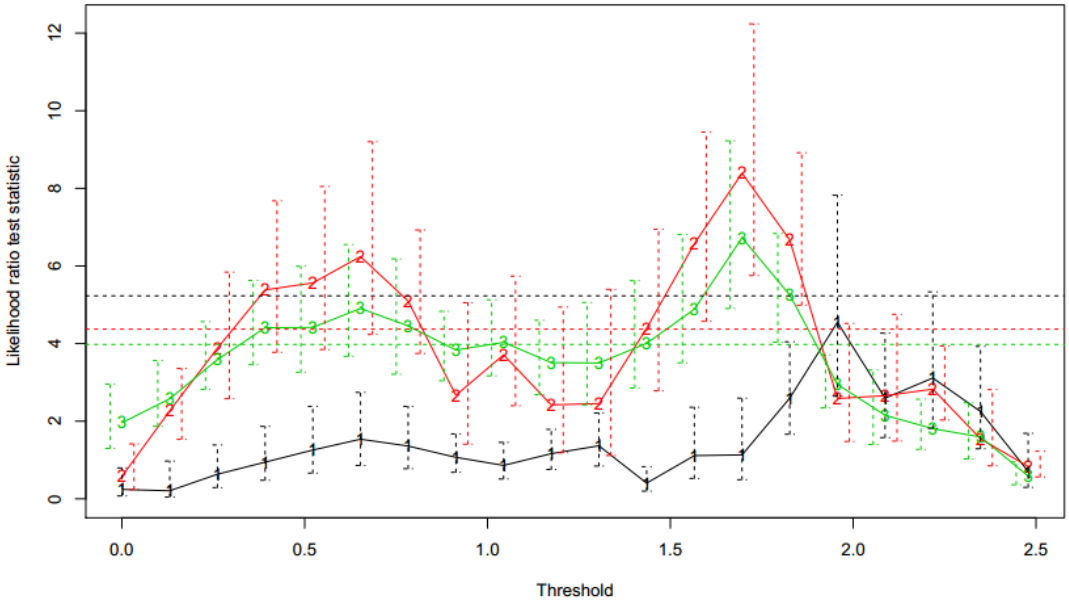


Figura 3: Validación cruzada 5 campos para el conjunto de Colon.

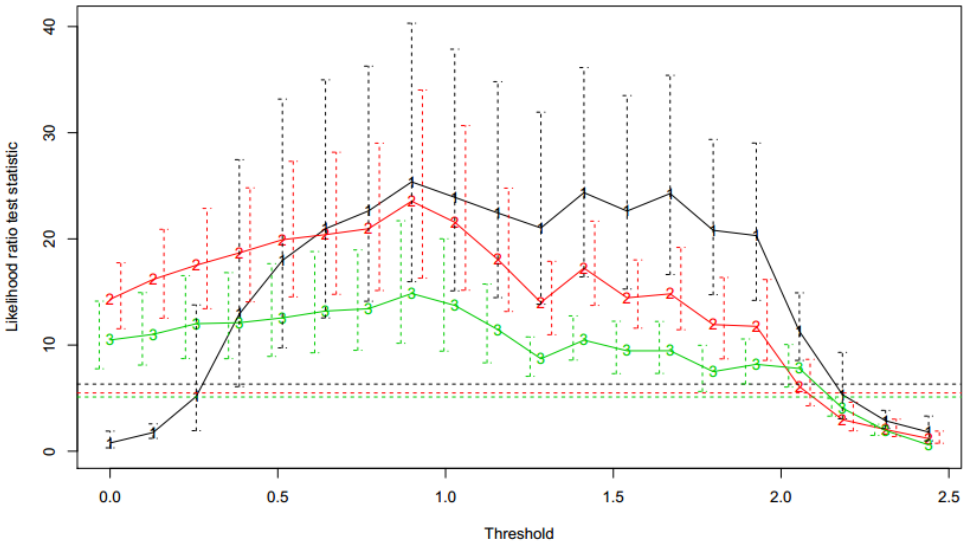


Figura 4: Validación cruzada 5 campos para el conjunto DLBCL.

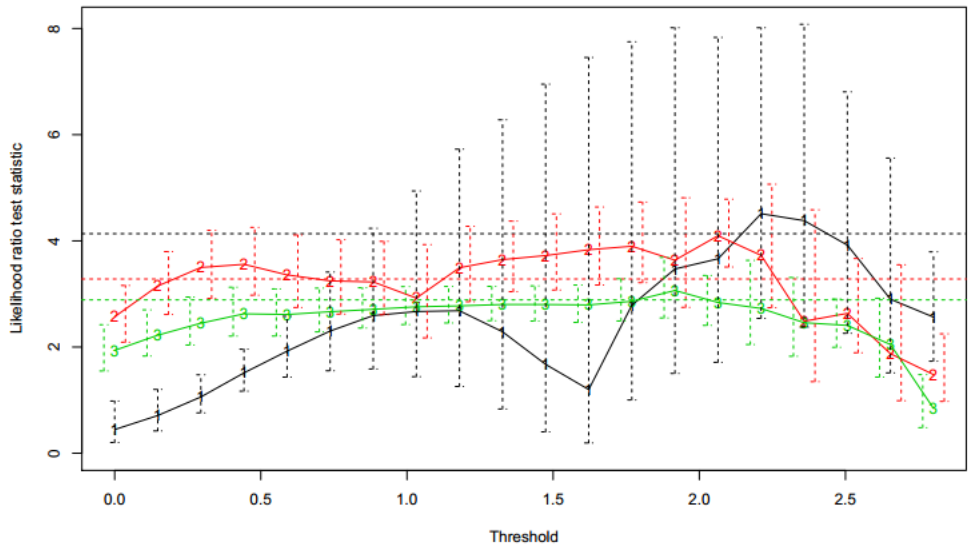


Figura 5: Validación cruzada 5 campos para el conjunto SMK.

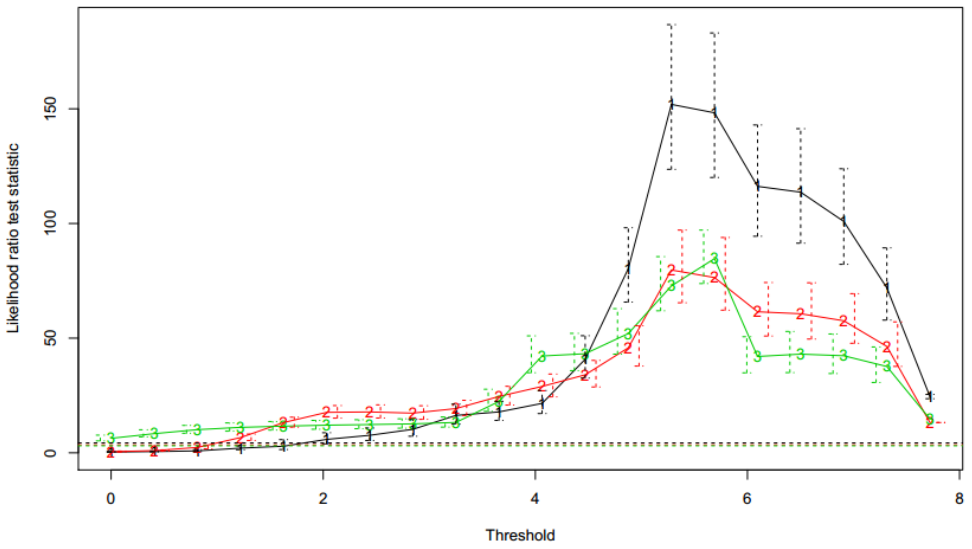


Figura 6: Validación cruzada 10 campos para el conjunto Ovarian.

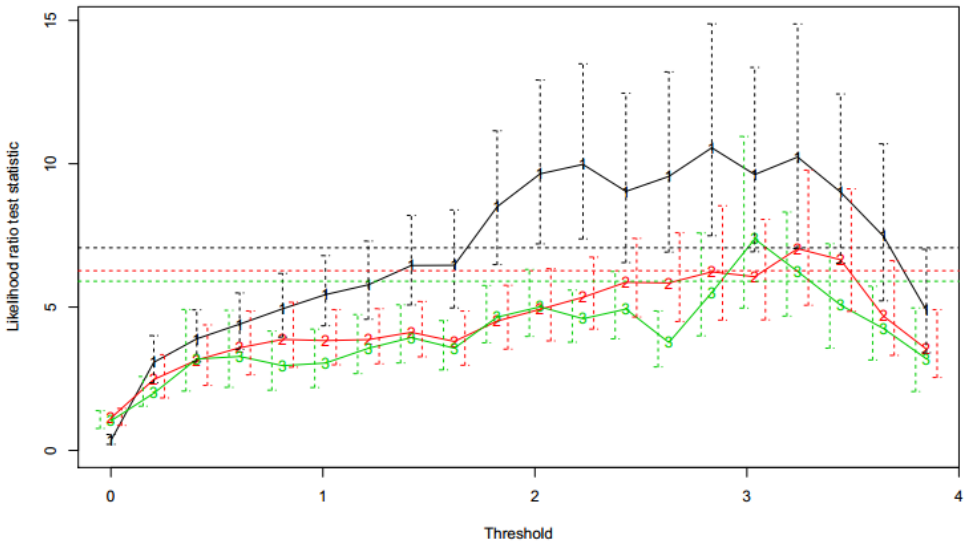


Figura 7: Validación cruzada 10 campos para el conjunto GLI-85.

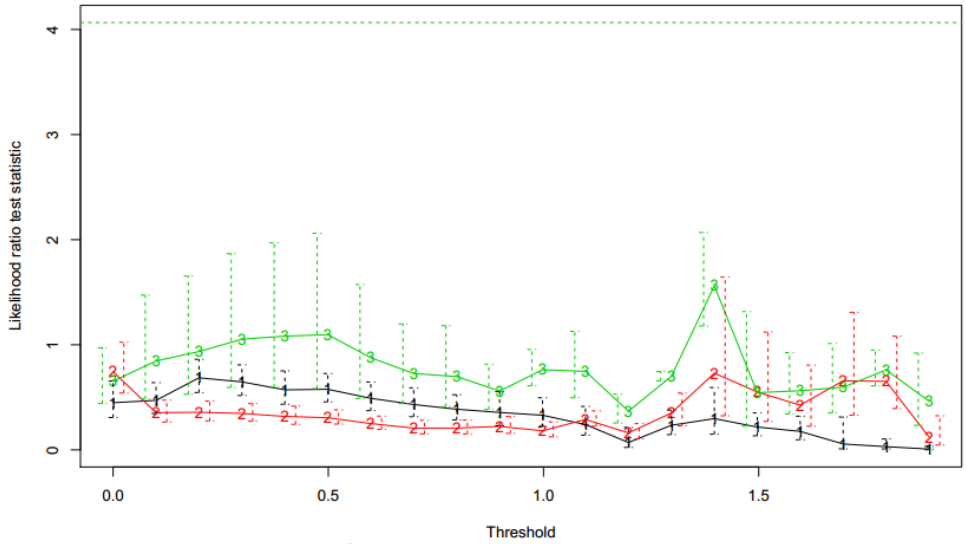


Figura 8: Validación cruzada 5 campos para el conjunto CNS.

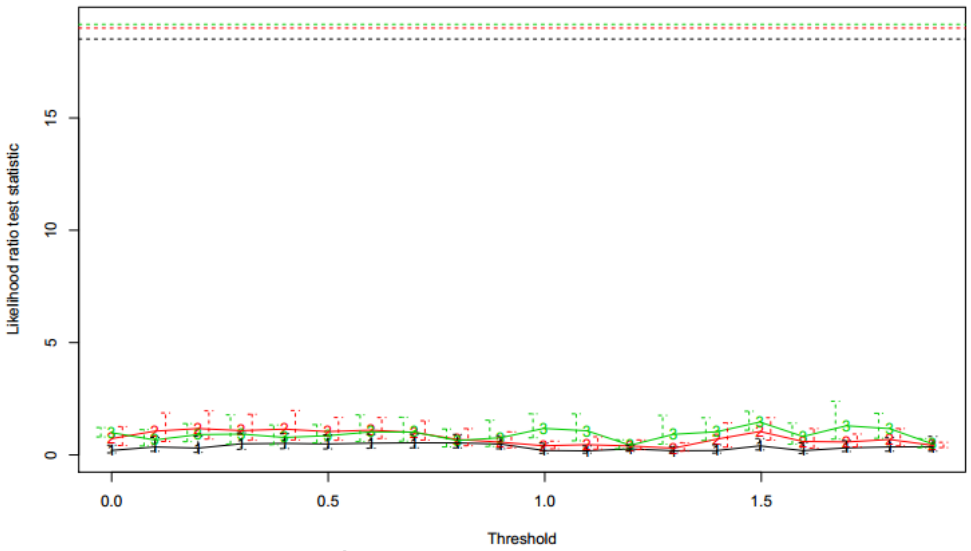


Figura 9: Validación cruzada 10 campos para el conjunto CNS.

Anexo 3: *Pseudocódigo para la propuesta* ***Doble reducción de la dimensión.***

Datos a entrar:

X: Datos de microarrays.

Y: Variable dependiente binaria. La clase "1" representa que el paciente no padece la enfermedad. La clase "2" representa que el paciente si padece la enfermedad.

type: Una variable que toma los valores "PCA" y "SPCA".

Hacer:

type = "PCA"

C = CUR(X)

if (type = "PCA") then

  Y_K = PCA(C)

else if (type = "SPCA") then

  Y_K = SPCA(Y, C)

else

  Y_K = PLS(Y, C)

end if

$\hat{Y}$ = LDA (Y, Y_K)

mc = matriz.confusion (Y, $\hat{Y}$)

Ex = sum(diag(mc))/sum(mc)

Se = mc [2,2]/sum(mc [2,])

Es = mc [1,1]/sum(mc [1,])

Comentarios:

C: Matriz obtenida por el algoritmo COLUMNSELECT aplicado a la matriz X.

PCA: Análisis de Componentes Principales.

SPCA: Análisis de Componentes Principales Supervisados.

PLS: Mínimos Cuadrados Parciales.

Y_K: K primeras componentes latentes.

LDA: Análisis Discriminante Lineal.

$\hat{Y}$: Estimación de la variable dependiente.

mc: Matriz de confusión de orden 2x2.

Ex: Exactitud.
Se: Sensibilidad.
Es: Especificidad.

Anexo 4: *Pseudocódigo para la propuesta* ***Doble reducción de la dimensión en forma paralela****.*

Datos a entrar:

X: Datos de microarrays.

Y: Variable dependiente binaria. La clase "1" representa que el paciente no padece la enfermedad. La clase "2" representa que el paciente si padece la enfermedad.

type: Una variable que toma los valores "PCA" y "SPCA".

Hacer:

type = "PCA"

for ($i = 1$ to m) do

  C_i= CUR(X)

  if (type = "PCA") then

    Y_{k_i} = PCA (C_i)

  else if (type = "SPCA") then

    Y_{k_i} = SPCA(Y, C_i)

  else

    Y_{k_i} = PLS(Y,C_i)

  end if

  $\hat{Y}_i$ = LDA (Y, Y_{k_i})

  mc_i = matriz.confusion (Y, $\hat{Y}_i$)

  Ex_i = sum(diag(mc_i))/sum(mc_i)

  Se_i = mc_i [2,2]/sum(mc_i [2,])

  Es_i = mc_i [1,1]/sum(mc_i [1,])

end for

best = min($1 - Ex_1$, $1 - Ex_2$, …, $1 - Ex_m$)

$\hat{Y}_{best}$

Comentarios:

C_i: Matriz i-ésima obtenida por el algoritmo COLUMNSELECT aplicado a la matriz X.

PCA: Análisis de Componentes Principales.

SPCA: Análisis de Componentes Principales Supervisados.

PLS: Mínimos Cuadrados Parciales.

Y_{k_i}: K primeras componentes latentes de la matriz C_i.

LDA: Análisis Discriminante Lineal.

$\hat{Y}_i$: Estimación i-ésima de la variable dependiente.

mc_i: Matriz i-ésima de confusión cuyo orden es 2x2.

Ex_i: Exactitud del i-ésimo modelo.

Se_i: Sensibilidad del i-ésimo modelo.

Es_i: Especificidad i-ésimo modelo.

best: Menor error de mala clasificación.

$\hat{Y}_{best}$: Mejor modelo de clasificación.

Anexo 5: *Matrices de confusión para el modelo PCA.*

Colon	FALSO	TRUE
FALSO	17	5
TRUE	6	35

DLBCL	FALSO	TRUE
FALSO	22	1
TRUE	5	19

CNS	FALSO	TRUE
FALSO	2	19
TRUE	10	29

Ovarian	FALSO	TRUE
FALSO	52	39
TRUE	18	144

GLI-85	FALSO	TRUE
FALSO	20	6
TRUE	5	54

SMK	FALSO	TRUE
FALSO	58	32
TRUE	21	76

Anexo 6: *Matrices de confusión para el modelo SPCA.*

Colon	FALSO	TRUE
FALSO	17	5
TRUE	3	37

DLBCL	FALSO	TRUE
FALSO	23	0
TRUE	1	23

Ovarian	FALSO	TRUE
FALSO	83	8
TRUE	0	162

SMK	FALSO	TRUE
FALSO	59	31
TRUE	19	78

GLI-85	FALSO	TRUE
FALSO	24	2
TRUE	1	58

Anexo 7: *Matrices de confusión para el modelo PLS.*

Colon	FALSO	TRUE
FALSO	19	3
TRUE	2	38

DLBCL	FALSO	TRUE
FALSO	23	0
TRUE	0	24

CNS	FALSO	TRUE
FALSO	5	16
TRUE	1	38

Ovarian	FALSO	TRUE
FALSO	90	1
TRUE	1	161

GLI-85	FALSO	TRUE
FALSO	24	2
TRUE	0	59

SMK	FALSO	TRUE
FALSO	58	32
TRUE	19	78

Anexo 8: *Matrices de confusión para el modelo CUR-PLS.*

Colon	FALSO	TRUE
FALSO	16	6
TRUE	5	35

DLBCL	FALSO	TRUE
FALSO	21	2
TRUE	2	22

CNS	FALSO	TRUE
FALSO	12	9
TRUE	5	34

Ovarian	FALSO	TRUE
FALSO	88	3
TRUE	2	160

GLI-85	FALSO	TRUE
FALSO	18	8
TRUE	1	58

SMK	FALSO	TRUE
FALSO	60	30
TRUE	19	78

Anexo 9: *Matrices de confusión para el modelo CUR-PLS-Par.*

Colon	FALSO	TRUE
FALSO	19	3
TRUE	2	38

DLBCL	FALSO	TRUE
FALSO	23	0
TRUE	0	24

CNS	FALSO	TRUE
FALSO	20	1
TRUE	3	36

Ovarian	FALSO	TRUE
FALSO	91	0
TRUE	0	162

GLI-85	FALSO	TRUE
FALSO	25	1
TRUE	0	59

SMK	FALSO	TRUE
FALSO	61	29
TRUE	17	80