REPORTE DE INVESTIGACION

del Instituto de Cibernética, Matemática y Física

Métodos de extracción de características en datos de microarrays de ADN para enfermedades oncológicas (Parte I)

Miguel A. Gutiérrez Arce,
Yunier E. Tejeda Rodríguez,
Carlos Rodríguez Fadragas

Octubre 2017 ICIMAF 2017-838

ICIMAF

Calle 15 No 551 e/ C y D, Vedado, La Habana 4 C.P. 10400, Cuba. Telf.: (537) 32 7764, 32 2688.
Fax: (537) 33 3373 Télex: 512230 ICIMAF CU.

ISSN 0138-8916

Reporte de Investigación del ICIMAF No. 838 (2017)
ISSN 0138-8916 Original paper

Reporte de Investigación del ICIMAF
Calle 15 No. 551 e/ C y D, Vedado,
La Habana, Cuba, CP 10400,
http://www.icimaf.cu

Métodos de extracción de características en datos de microarrays de ADN para enfermedades oncológicas (Parte I).

Miguel A. Gutiérrez Arce[1], Yunier E. Tejeda Rodríguez[2], Carlos Rodríguez Fadragas[3]

[2]**Dpto Matemática UCLV**, [3]**Dpto Física UCLV**

[1]mgarce@uclv.cu, [2]yunier@uclv.cu, [3]fadragas@uclv.edu.cu

Resumen: En este reporte 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 propone una metodología que consiste en tres etapas para obtener un modelo de clasificación que pronostique diferentes enfermedades oncológicas. Por último, se hace una discusión de los resultados por dicha metodología.

Palabras claves: Reducción de la dimensión**,** Análisis de componentes principales, Descomposición del valor singular, Análisis de componentes principales supervisado, Mínimos cuadrados parciales.

Abstract: This report 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. We propose a methodology that consists of three stages to obtain a classification model that predicts different oncological diseases. Finally, a discussion of the results is made by said methodology.

Keywords: Dimensional reduction, Principal component analysis, Singular value decomposition, Supervised principal components analysis, Partial least squares.

1. 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 tumores. Aunque son empleadas muestras muy pequeñas (por lo general menos de100 pacientes) para entrenamientos y pruebas, el número de características crece exponencialmente y van desde las 6000 hasta las 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 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 permitan distinguir entre muestras cancerígenas y no cancerígenas.
- Caracterizar los algoritmos aleatorios para datos de microarrays de ADN en enfermedades oncológicas.
- Proponer una metodología que permita distribuir los conjuntos de datos de microarrays en diferentes subconjuntos y luego aplicar un método de extracción de características lineales 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.

A continuación, se presenta como está estructurado este reporte. En la sección 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 dichas componentes. En la siguiente sección se describen los conjuntos de microarrays a investigar y seguido se muestra la metodología a proponer. En la sección 5 se presentan los resultados y en la sexta se realiza la discusión de los mismos. Por último, se dan las conclusiones.

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

En esta sección 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 ésta 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 [3, 4].

Según [5], PCA se obtiene por la descomposición propia de la matriz de covarianza o correlación. Esta descomposición se emplea 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, éstos 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,

$$X = U\Sigma V^t \tag{1}$$

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

$$U = XV\Sigma^{-1} = XW \tag{2}$$

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

$$\lambda_k > \frac{1}{r}\sum_{k}^{r}\lambda_k \tag{3}$$

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 [5]. Además, existe la posibilidad de elegir las componentes cuya varianza acumulativa explique un $Q100\%$ que garantice un alto comportamiento de la variabilidad total.

$$\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{j=1}^{p}\lambda_j} \leq Q \tag{4}$$

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:

$$imp_j = \sum_{i=1}^{k} cor^2(x_j, u_i), \forall j = 1, \dots, p \tag{5}$$

2.2. Análisis de Componentes Principales Supervisados.

El SPCA fue propuesto por [6] para problemas de regresión donde el número de variables es mucho mayor que el número de muestras. En lo adelante, se asume que 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 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) están 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:

$$s_j = \frac{x_j^t Y}{\|x_j\|} \tag{6}$$

con

$$\|x_j\| = \sqrt{x_j^t x_j} \tag{7}$$

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:

$$X_\theta = U_\theta \Sigma_\theta V_\theta^t \tag{8}$$

$$U_\theta = X_\theta \mathrm{V}_\theta \Sigma_\theta^{-1} = X_\theta \mathrm{W}_\theta \tag{9}$$

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,

$$\hat{Y}^{spc,\theta} = \bar{Y} + \hat{Y} u_{\theta,1} \tag{10}$$

donde $\hat{Y} = u_{\theta,1}^t Y$ 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 $u_{\theta,1}$:

$$imp_j = cor(x_j, u_{\theta,1}), \forall j = 1, \ldots, p \tag{11}$$

Mientras mayor sea el valor absoluto de imp_j, mayor será la contribución de la variable x_j en la predicción de Y [7].

2.3. Mínimos Cuadrados Parciales.

En 1966 Wold propone el método PLS [8] y en su forma original se asociaban a los sistemas de ecuaciones estructurales (SEM, por sus siglas en inglés) [9]. 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 [10-12] cuyas ideas se presentan a continuación.

Se supone que existen q variables $Y_1, \dots, Y_q$ dependientes de p variables independientes $X_1, \dots, X_p$. 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:

$$\begin{aligned} Y &= TQ^t + F \\ X &= TP^t + E \end{aligned} \tag{12}$$

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

$$T = XW \tag{13}$$

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 $w^t w = 1$ [11]. Para esto, se utilizan varios algoritmos, entre ellos están NIPALS [13], KERNEL-PLS [14, 15], y SIMPLS [16]. En los paquetes pls [17], nipals [18] y plsgenomics [19] 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", es decir,

$$Q^t = (T^t T)^{-1} T^t Y \tag{14}$$

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:

$$B = WQ^t \tag{15}$$

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 [13], donde demuestra que el estadístico- $\mathrm{BSS}_j/\mathrm{WSS}_j$ [14] 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:

$$f({w_1}^2) = -\sqrt{{w_1}^2} = -|w_1| \tag{16}$$

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

En este reporte de investigación se estudian 6 de los 9 conjuntos de datos microarrays binarios estudiados por [15]. En la **Tabla 1** se muestra una breve 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 [16], el conjunto de datos CNS/Embrional-T fue descargado del repositorio *Dataset Repository*, from the Bioinformatics Research Group of Universidad Pablo de Olavide [17] mientras que los conjuntos de datos GLI-85 y SMK-CAN-187 fueron descargados del repositorio *Feature Selection Dataset*, from Arizona State University [18].

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	[19]
Colon	62	2000	1,82	1,08	[20]
DLBCL	47	4026	1,04	2,91	[21]
GLI-85	85	22,283	2,27	2,35	[22]
Ovarian	253	15,154	1,78	6,94	[23]
SMK-CAN-187	187	19,993	1,08	0,41	[24]

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 [25] que puede ser calculada a través de la siguiente relación:

$$F1 = \max_{i,j=1,\dots n} \left\{ \frac{\left(\mu_i - \mu_j\right)^2}{\sigma_i^2 + \sigma_j^2} \right\} \; donde \; n = p \tag{17}$$

3.1. 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 [25].

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" [25] 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):

$$Se = \frac{PV}{PV + NF}; 0 \leqslant Se \leqslant 1 \tag{18}$$

$$Es = \frac{NV}{NV + PF};\ 0 \leqslant Es \leqslant 1 \tag{19}$$

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

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.

5. Resultados.

La metodología que se propone está implementada en el software R [26], 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 de 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 (4) tomando como valor $Q = 0.75$, por su parte, PLS lo hace mediante una propuesta por validación cruzada de [13, 14] implementada en el paquete plsgenomics. En el caso de SPCA, es necesario estimar el parámetro θ por validación cruzada K-campos con el objetivo de 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 [27] 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 [28].

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. En los **Anexos 3**, **4** y **5** se encuentran las matrices de confusión para los modelos PCA, SPCA y PLS, respectivamente.

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

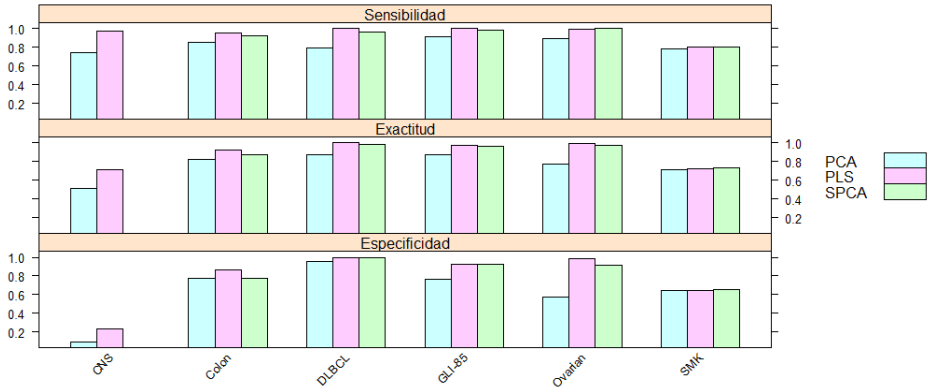

Figura 1: Resultados para el LDA en los conjuntos de datos.

6. Discusión.

En la sección anterior se implementaron 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 presenta 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 del 52%, siendo este muy bajo, por lo tanto, se puede 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**.

7. Conclusiones.

- En este reporte 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.
- Se muestran los conjuntos de datos de microarrays de ADN y se comentan sus principales problemas.
- 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.

RECONOCIMIENTOS

Los autores desean agradecer las contribuciones que ha hecho el MSc. Jorge Luis Morales Martínez y a todas aquellas personas que han contribuido de alguna manera para que sea posible la realización del presente trabajo. Para todos ellos, MUCHAS GRACIAS…"

Referencia

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. Jollife, I.T., *Principal Component Analysis*. 2 ed. Springer Series in Statistics, ed. Springer. 2002, New York: Springer-Verlag.
4. Mardia, K.V., Kent, J. T. and Bibby, J.M., *Multivariate Analysis*. 1979, Londres: Academic Press.
5. Williams, H.A.a.L.J., *Principal component analysis.* WIREs Computational Statistics, 2010. **2**: p. 433-460.
6. 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.
7. 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.
8. 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.
9. Joreskög, K.G., *A general method for analysis of covariance structures.* Biometrika, 1970. **57**: p. 239-251.
10. 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.
11. 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.
12. Wold S, S.M.a.E.L., *PLS-regression a basic tool of chemometrics* Chemometr. Intell. Lab., 2001. **58**: p. 109–130.
13. Boulesteix, A.-L., *PLS Dimension Reduction for Classification with Microarray Data.* Statistical Applications in Genetics and Molecular Biology, 2004. **3**(1): p. 1-32.
14. 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.

15. 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.
16. Dataset, K.R.B.-M. 2014.
17. Dataset Repository, B.R.G. 2014.
18. University, F.S.D.a.A.S. 2014.
19. 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.
20. 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.
21. 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.
22. 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.
23. 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.
24. 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.
25. Bolón-Canedo, V., *Novel feature selection methods for high dimensional data*, in *Department of Computer Science*. 2014, UNIVERSITY OF A CORUÑA.
26. Team, R.C., *R: A Language and Environment for Statistical Computing*, R.F.f.S. Computing, Editor. 2016: Vienna, Austria.
27. Tibshirani, E.B.a.R. *superpc: Supervised principal components*. 2012; Available from: https://CRAN.R-project.org/package=superpc.
28. Ripley, W.N.V.a.B.D., *Modern Applied Statistics with S*. Fourth ed. 2002, New York: Springer.

Anexos:

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.

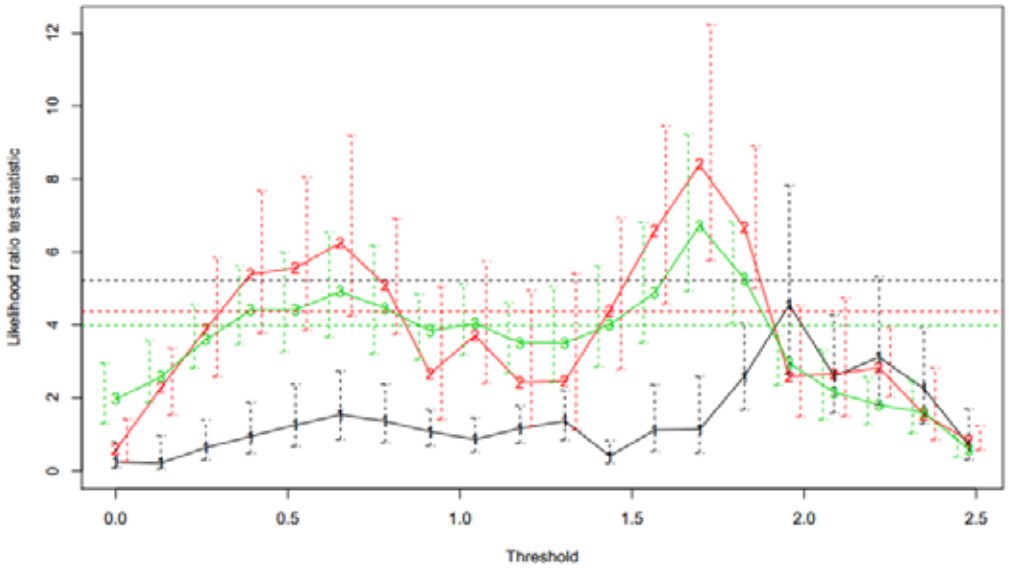

Figura 2: Validación cruzada 5 campos para el conjunto de Colon.

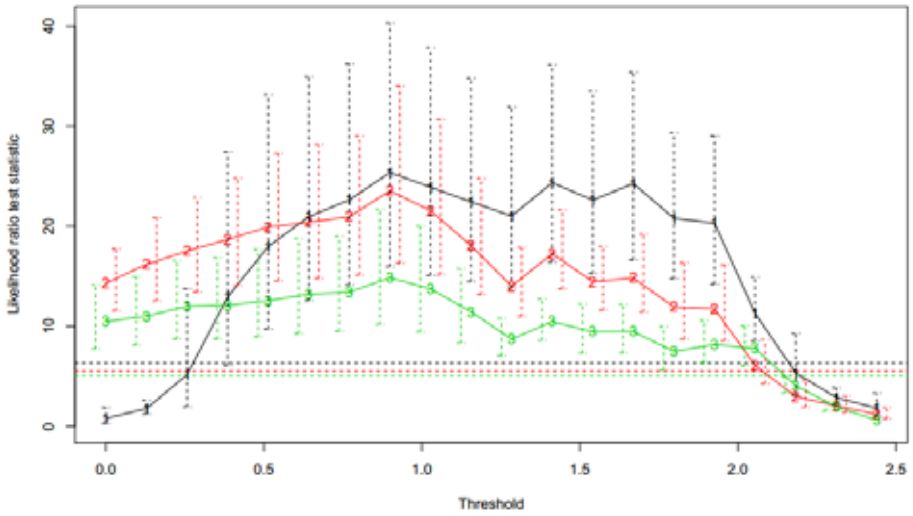

Figura 3: Validación cruzada 5 campos para el conjunto DLBCL.

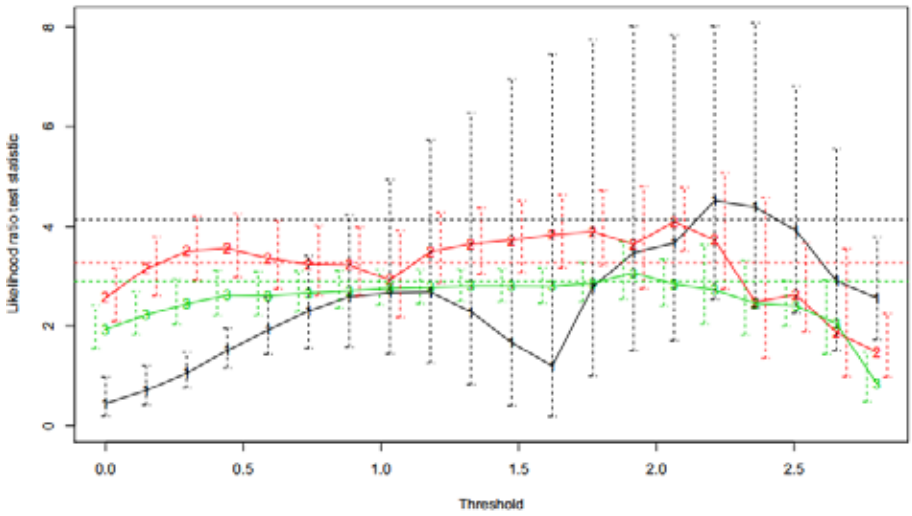

Figura 4: Validación cruzada 5 campos para el conjunto SMK.

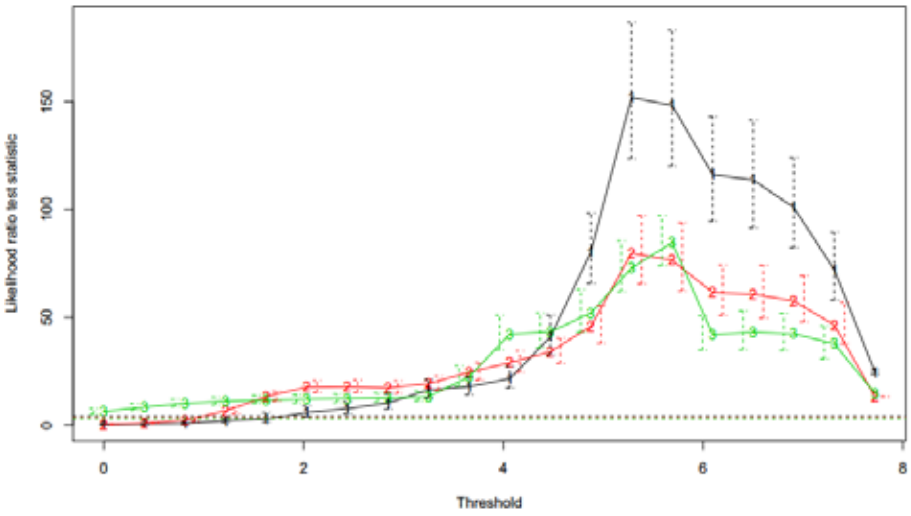

Figura 5: Validación cruzada 10 campos para el conjunto Ovarian.

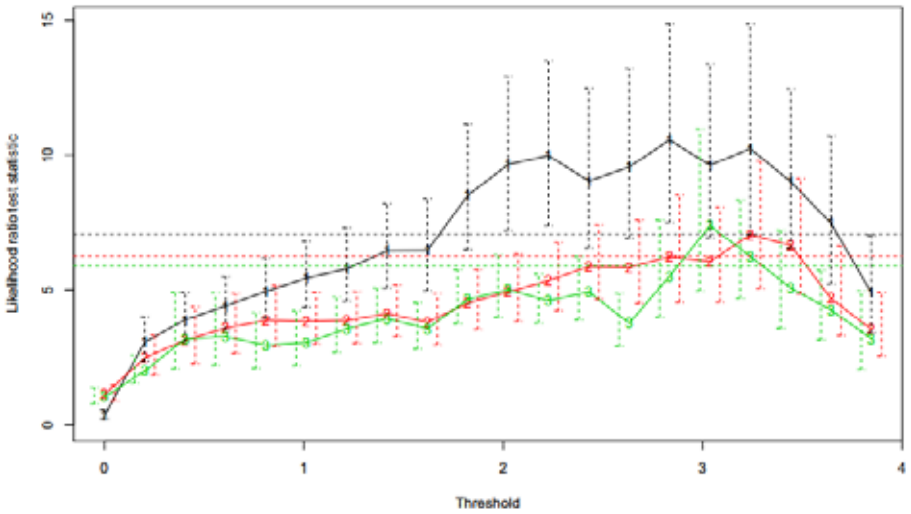

Figura 6: Validación cruzada 10 campos para el conjunto GLI-85.

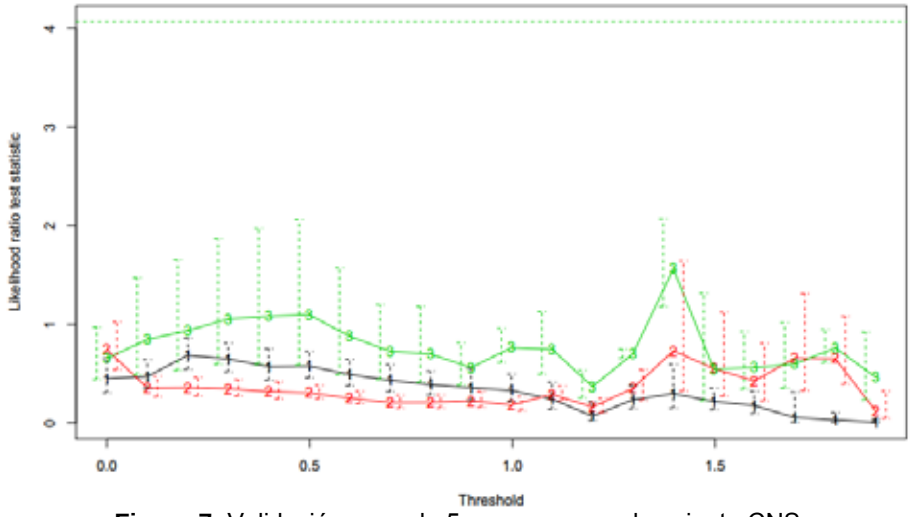

Figura 7: Validación cruzada 5 campos para el conjunto CNS.

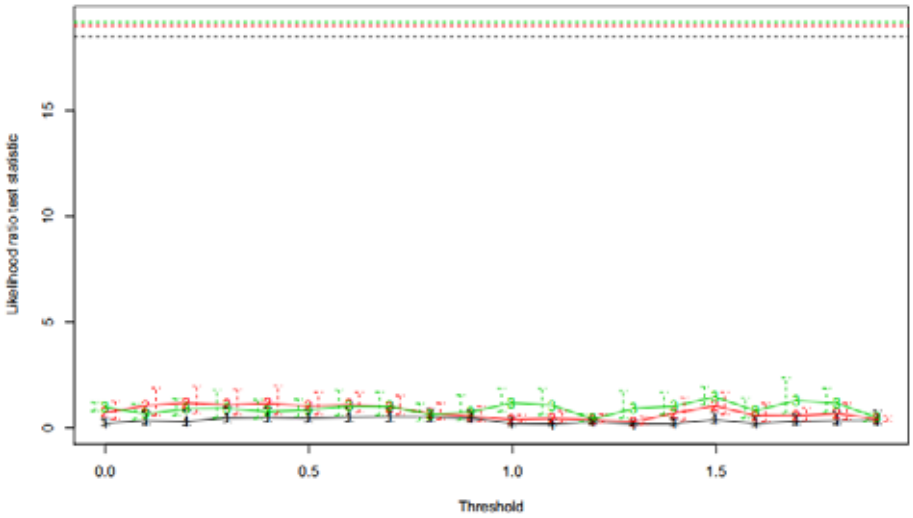

Figura 8: Validación cruzada 10 campos para el conjunto CNS.

Anexo 3: Matriz 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 4: Matriz 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 5: Matriz 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