Asociación Argentina
de Mecánica Computacional

Mecánica Computacional Vol XXVI, pp.2145-2151
Sergio A. Elaskar, Elvio A. Pilotta, Germán A. Torres (Eds.)
Córdoba, Argentina, Octubre 2007

ALGORITMO GOLOSO DE SELECCIÓN DE CARACTERÍSTICAS APLICADO A DATOS DE MICROARREGLO DE ADN

Agustina Bouchet, Mariela Azul Gonzalez, Juan Ignacio Pastore, Virginia Ballarin y Marcel Brun

Grupo de Procesamiento de Imágenes, Laboratorio de Procesos y Medicion de Señales, Facultad de Ingeniería, Universidad Nacional de Mar del Plata, Mar del Plata, Argentina

Palabras clave: Microarreglos de ADN, Selección de características

Resumen. El objetivo del análisis de la expresión genética es generar un algoritmo que permita asignar a cada patrón de expresión un fenotipo y que éste se corresponda con el fenotipo verdadero del paciente bajo analisis. Este análisis también puede ser útil para determinar los genes cuya expresión caracteriza los fenotipos patológicos. Para ello los especialistas deben construir arreglos experimentales que permitan confirmar los datos obtenidos a partir de los algoritmos. Uno de los mayores problemas que se presenta en el análisis estadístico de microarreglos es el de la gran dimensionalidad de los datos con respecto a la cantidad disponible de muestras. El error verdadero de un clasificador, diseñado a partir de los datos, disminuye cuando aumenta la cantidad de características, hasta cierto número óptimo a partir del cuál el error se incrementa nuevamente. Para solucionar este problema es necesario aplicar algoritmos de selección y extracción de características. En este trabajo presentamos un algoritmo goloso que puede seleccionar un conjunto de características, con baja correlación entre si, que permite predecir la variable dependiente. Este algoritmo mantiene el objetivo original de seleccionar genes altamente correlacionados a la variable dependiente, pero con un factor de peso dado por su máxima correlación a algún elemento del conjunto seleccionado en la iteración anterior. El funcionamiento del algoritmo fue probado mediante datos simulados. En todos los casos fue posible seleccionar las mejores características a partir de muchos candidatos en tiempo razonable, obteniendo resultados con alto nivel de predicción y menor tamaño.

1. INTRODUCCIÓN

La información acerca de la secuencia de aminoácidos que posee cada proteína se encuentra almacenada en el ácido dexoxiribonucleico (ADN) ubicado dentro del núcleo celular. Esta información compone el genotipo de las células. De todo el genotipo presente sólo se utiliza un pequeño porcentaje para sintetizar proteínas y esta expresión genética limitada permite que las células se diferencien entre si y respondan de manera diferencial a las diversas señales biológicas. La información que se expresa se denomina fenotipo y en algunas personas puede contener errores (mutaciones) que afectan a su salud.

El objetivo del análisis de la expresión genética es generar un algoritmo que permita asignar a cada patrón de expresión un fenotipo característico (sano, enfermo, enfermedad tipo 1, tipo 2, etc.) y que este fenotipo se corresponda con el fenotipo verdadero del paciente analizado Dougherty et al. (2004); Braga-Neto and Dougherty (2005). El estudio de los fenotipos de los pacientes se realiza mediante microarreglos genéticos. Estos microarreglos poseen sensores que detectan si un gen se esta expresando y además pueden estudiar los miles de genes que caracterizan el fenotipo de cada uno de los cientos de pacientes simultáneamente. Este análisis también puede ser útil para determinar los genes cuya expresión caracteriza los fenotipos patológicos. Esta información puede ayudar a determinar el tipo de enfermedad o la posible causa de la misma. Para ello los especialistas deben construir arreglos experimentales que permitan confirmar los datos que resultan de los algoritmos Quackenbush (2002); Dudoit et al. (2003); Hedenfalk et al. (2001).

Uno de los mayores problemas que se presenta en el análisis estadístico de microarreglos es la gran dimensionalidad de los datos respecto a la cantidad disponible de muestras. Este problema es inherentemente estadístico. Típicamente, el error verdadero de un clasificador, diseñado a partir de los datos, disminuye cuando aumenta la cantidad de características, hasta cierto número. A partir de esa cantidad óptima de características el error se incrementa nuevamente. Este fenómeno es a veces llamado maldición de la dimensionalidad.

Algunas soluciones al problema de dimensionalidad se pueden obtener aplicando algoritmos de selección y extracción de características Tao et al. (2004); Choudhary et al. (2006); Hashimoto et al. (2003). Usualmente, una lista inicial de unos pocos cientos o miles de genes se selecciona antes de diseñar clasificadores. Esta selección se realiza utilizando criterios simples basados en correlación o test de hipótesis Dougherty and Brun (2006). Uno de los problemas que se presenta al realizar esta selección inicial es que las características con mayor correlación con la variable dependiente (o clases) están usualmente correlacionadas entre ellas.

En este trabajo presentamos un algoritmo goloso que puede seleccionar un conjunto de características, con baja correlación entre si, que permite predecir la variable dependiente. Este algoritmo mantiene el objetivo original de seleccionar genes altamente correlacionados a la variable dependiente, pero con un factor de peso dado por su máxima correlación a algún elemento del conjunto seleccionado en la iteración anterior.

El funcionamiento del algoritmo fue probado mediante datos simulados con modelos utilizados previamente en otros trabajos de este tipo Dougherty and Brun (2006). En todos los casos fue posible seleccionar las mejores características a partir de miles de candidatos en tiempo razonable, obteniendo resultados con alto nivel de predicción y menor tamaño.

2. SELECCIÓN DE CARACTERÍSTICAS

Los métodos de selección de características se dividen en dos grandes grupos: métodos de envoltura (wrapper) y métodos de filtro (filter) Isabelle and Andre (2003); Tabus and Astola

(2005).

Los métodos de envoltura utilizan la precisión predictiva de un algoritmo de clasificación específico para determinar el poder discriminativo de un subconjunto de características. Este modelo es computacionalmente muy costoso para un conjunto de características muy grande y sólo se puede asegurar su utilidad para el clasificador utilizado.

Los métodos de filtro separan la selección de características de la clasificación y se basan en características generales del conjunto de datos de entrenamiento para seleccionar subconjuntos de características independientes de los algoritmos utilizados para clasificar. Los genes se ordenan de acuerdo a su relevancia individual o el poder discriminativo de las clases a las que pertenecen cada muestra. Luego se seleccionan los mejores genes de acuerdo al índice utilizado para evaluar las características previamente mencionadas.

Debido a la gran dimensionalidad del conjunto de datos los algoritmos de búsqueda deben evitar la búsqueda exhaustiva, por más que el conjunto encontrado sea subóptimo. Otro de los problemas que se presenta al realizar esta selección inicial es que las características con mayor correlación con la variable dependiente (o clases) están usualmente correlacionadas entre ellas. Esta correlación tiene dos consecuencias indeseables: a) la cantidad total de características necesaria para predecir la variable dependiente puede ser mucho mayor que el límite práctico de variables que pueden procesar los algoritmos; b) el uso de dos o más variables correlacionadas puede introducir confusión e impedir la obtención de resultados satisfactorios Dougherty and Brun (2006).

Una solución al problema es utilizar algoritmos llamados Análisis de Características Principales Ira et al. (2002) (Principal Feature Analysis PFA) que, a diferencia del Análisis de Componentes Principales (Principal Component Analysis PCA), no crea nuevas características como combinación lineal de las originales, sino que selecciona el mejor subconjunto de características ortogonales entre si. Una de las limitaciones de estos algoritmos es su gran costo computacional, lo cual los hace inapropiados para situaciones donde existen mas de 50000 variables (genes) en un mismo microarreglo. Otras técnicas sufren de las mismas limitaciones Yu and Liu (2004).

3. ALGORITMO

El algoritmo goloso presentado selecciona un conjunto de características que mejor predice la variable dependiente, y al mismo tiempo contiene una baja correlación entre las características seleccionadas. Este algoritmo mantiene el objetivo original de seleccionar genes altamente correlacionados a la variable dependiente, pero con un factor de peso dado por su máxima correlación a algún elemento del conjunto seleccionado en la iteración anterior.

En cada iteración k del algoritmo, una nueva característica f es agregada al conjunto anterior F_{k-1}. La selección de la característica a agregar se hace basada en su correlación a la variable dependiente $\mathbf{y}$ su máxima correlación a algún elemento de F_{k-1}. El rango de cada candidato f consiste en la correlación entre f y la variable dependiente, dividido por un término proporcional a la máxima correlación entre f y los miembros de F_{k-1}. El elemento con mayor rango es agregado al conjunto $F_k = F_{k-1} \cup \{f\}$, y un nuevo rango es computado para los restantes elementos, hasta incluir K elementos en F_K.

Para un conjunto seleccionado de características, $F_{k-1} = f_{i_1}, ..., f_{i_{k-1}}$, cada características todavia no seleccionada, $f \in F - F_{k-1}$, recibe un puntaje asociado al cociente

$$R(f) = \frac{D(\mathbf{x}_f, \mathbf{y})}{0{,}01 + \max\limits_{g \in F_{k-1}} D(\mathbf{x}_f, \mathbf{x}_g)} \tag{1}$$

donde $\mathbf{x}_f$ y $\mathbf{x}_g$ representan el vector de expresión de las características f y g, respectivamente, y $D(\mathbf{x}, \mathbf{y})$ puede ser la correlacion $\rho(\mathbf{x}, \mathbf{y})$ u otra medida de similitud.

Esta ecuación mantiene el objetivo original de seleccionar genes altamente correlacionados al objetivo $\mathbf{y}$ (variable independiente), pero al mismo tiempo considerando la redundancia existente con las características previamente seleccionadas. Los pasos a seguir para la aplicación de este algoritmo son los siguientes (para simplificar la notación, cuando sea posible usaremos $\mathbf{x}_i$ para denotar $\mathbf{x}_{f_i}$):

1. Inicializa F como lista de todas las características, $F = \{f_1, \ldots, f_N\}$
2. Inicializa la lista de características F_0 vacia, $F_0 = \emptyset$
3. Calcula la correlación $D(\mathbf{x}_i, \mathbf{y})$ para todas las características $f_i \in F$
4. Selecciona la característica f_{i_0} mas correlacionada al objetivo: $D(i_o) = D(\mathbf{x}_{i_0}, \mathbf{y}) \leq D(\mathbf{x}_i, \mathbf{y})$ para todo $i = 1, \ldots, N$.
5. Agrega f_{i_0} a la lista $F_1 = \{f_{i_0}\}$
6. Para el paso $k \geq 2$, para cada característica f_i en $F - F_{k-1}$, calcula su correlación maxima $MD(i)$ con alguna característica en F_{k-1}, y su maxima correlacion $D(i)$ al objetivo:

$$MD(i) = \max_{g \in F_{k-1}} D(\mathbf{x}_i, \mathbf{x}_g)$$

$$D(i) = D(\mathbf{x}_{i_0}, \mathbf{y}) \leq D(\mathbf{x}_i, \mathbf{y}) \text{ para todo } i = 1, \ldots, N$$

7. Selecciona la característica i_0 con mayor cociente $D(i)/(0{,}01 + MD(i))$ entre las características a $f_i \in F - F_{k-1}$
8. Repite el paso 6 hasta seleccionar K características en F_K.

4. RESULTADOS EXPERIMENTALES

Para evaluar la habilidad del algoritmo en su tarea de seleccionar características utilizamos un modelo sintético de datos basado en dos clases modeladas con una distribución Gaussiana multivaluada, con matriz de covarianza en bloques, para simular la existencia de características altamente correlacionadas.

Como en otros estudios Dougherty and Brun (2006), una estructura en bloques de la matriz de covarianza permite que características en el mismo bloque estén altamente correlaciondas, y características en bloques diferentes tengan correlación cercana a cero. Todas las características tienen la misma varianza, de tal forma que los elementos de la diagonal, en la matriz de covarianza, tienen idéntico valor σ^2. Para definir los valores de correlación, las d características se dividen igualmente en K grupos, con cada grupo teniendo d/K características. La matriz de covarianza contiene valores iguales a cero fuera de los bloques, valores σ^2 en la diagonal, y valores $\rho \cdot \sigma^2$ en los bloques.

Los parámetros con los que fue utilizado el algoritmo son $\rho = 0{,}6$, $\sigma = 1{,}3$, $d = 256$ y $K = 16$, simulando la situación en que existe un gran número de características, con alto grado de correlación: cada uno de los 16 grupos contiene 16 características altamente correlacionadas.

Para cada experimento uno de estos valores es variado dentro de cierto rango para estudiar su efecto en los resultados.

Para cada valor de los parámetros, el algoritmo seleccionó una cantidad K de características, las cuales fueron usadas para evaluar el error de un clasificador lineal. El clasificador lineal se entrenó usando un discriminante lineal sobre las características seleccionadas. Este clasificador asume que las dos clases tienen distribución normal con matrices de covarianza similares, generando un hiperplano que separa las dos clases Braga-Neto and Dougherty (2005). El error se estimó aplicando el clasificador generado a partir de una gran cantidad de muestras, y contando la proporción de muestras mal clasificadas. El entrenamiento y la estimación de error fueron realizadas con una gran cantidad de datos para evitar error de estimación. Este proceso fue repetido 100 veces para cada valor de K, para promediar los resultados y generar los gráficos de error. De esta forma, los gráficos muestran la capacidad discriminatoria de las características seleccionadas, y no refleja posibles limitaciones de entrenamiento y estimación de error.

Exp	**d**	**K**	**MedidaA**	**Medida B**	σ	ρ
A	256	16	Corr	Corr	1 4 6 8	0.8
B	256	16	Corr	Corr	3	0.3 0.5 0.7 0.9

Tabla 1: Lista de experimentos y parámetros usados

Las figuras 1 a) y 1 b) muestran los resultados del error calculado, usando un tamaño razonablemente alto de 1000 muestras para entrenar el clasificador y 1000 muestras para calcular el error, para los experimentos listados en la tabla 4. Los dos gráficos muestran como el error se estabiliza después de 16 características. En todos los casos, las primeras 16 características seleccionadas por el algoritmo pertenecen a diferentes bloques. Para mas de 16, las características adicionales están altamente correlacionadas con las primeras 16, y no aportan poder de discriminación, por lo que el error se estabiliza.

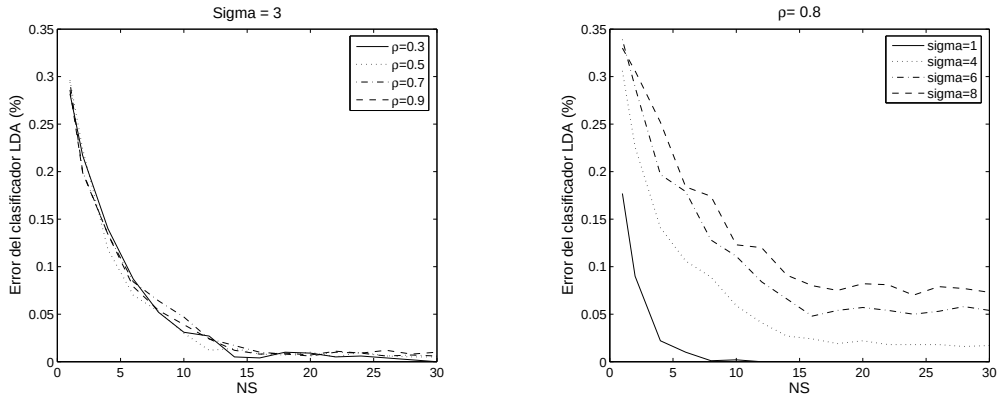
El algoritmo propuesto fue contrastado contra selección de características basada en la métrica BSS/WSS, computada como la proporción entre la distancia entre las clases y el tamaño de las clases Tabus and Astola (2005):

$$\frac{BSS(i)}{WSS(i)} = \frac{\sum_c \eta_c (\overline{x}_j^{(c)} - \overline{x}_j)}{\sum_c \eta_c \left(\sigma_j^{(c)}\right)^2}$$

donde c es la clase, η_c es la cantidad de elementos de la clase c, $\overline{X}_j$ es el valor medio de de la característica j, $\overline{X}_j^{(c)}$ es el valor medio de de la característica en la clase c, y $\sigma_j^{(c)}$ es la varianza de la característica j en la clase c.

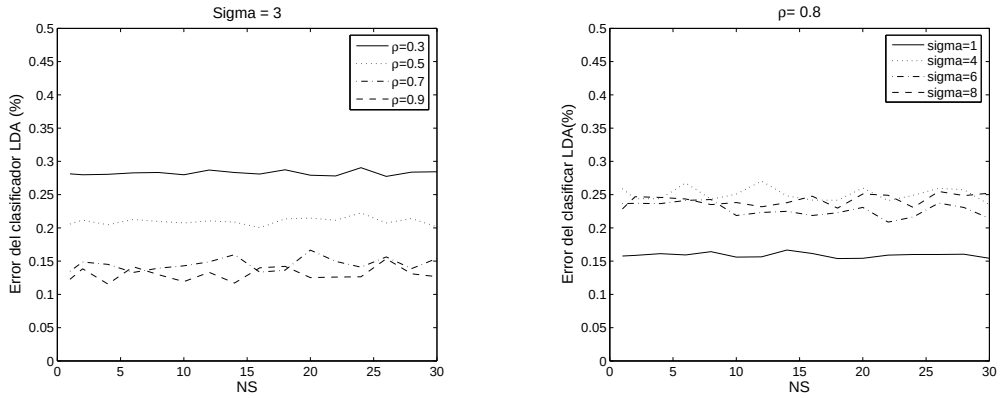
Las figuras 2 a) y 2 b) muestran los resultados obtenidos para los mismos experimentos de la tabla 4. En este caso vemos que para cualquier número de características el error es constante, usualmente más alto que los valores obtenidos con el algoritmo goloso, excepto para cantidades muy pequeñas de características.

El comportamiento indeseado del uso de BSS/WSS se debe al hecho que las características son seleccionadas independientemente y pueden pertenecer al mismo bloque. Características en el mismo bloque tienen alta correlación, por lo que el incremento de discriminación es bajo. Este mismo problema se presenta con el uso de técnicas tradicionales como el test t, ampliamente utilizado en bioinformática Quackenbush (2002) para preselección de características.



a) b)

Figura 1: Error de las características seleccionadas para distintos valores de variancia σ^2 (a) y correlación ρ (b) para el algoritmo goloso



a) b)

Figura 2: Error de las características seleccionadas para distintos valores de variancia σ^2 (a) y correlación ρ (b) para el algoritmo BSS/WSS

5. CONCLUSIONES

En este trabajo mostramos como el algoritmo goloso puede realizar una tarea ampliamente superior en la selección de un alto número de características, relativo a seleccion independiente basada en medidas de similitud o discriminación. Los resultados experimentales muestran que las primeras características seleccionadas aportan gran poder discriminatorio, lo cual no sería posible en el caso habitual de selección independiente. Por su naturaleza, el algoritmo goloso puede ser empleado eficientemente en datos con una alta cantidad de características, donde es

necesario seleccionar un número elevado de las mismas. En nuestros experimentos tardo 6 segundos en procesar un conjunto de datos con 30 características. Esto no es posible de realizar eficientemente con algoritmos de selección de características usuales, por la alta dimensionalidad del espacio, ni con otros algoritmos de selección propuestos como PFA debido al elevado costo computacional.

Futuros trabajos incluyen el análisis comparativo de distintas medidas de similitud, y la aplicación a datos públicos de microarreglo de ADN.

REFERENCIAS

Braga-Neto U. and Dougherty E.R. *Classification*, pages 93–128. EURASIP Book Series on Signal Processing and Communications. Hindawi Publishing Corporation, 2005.

Choudhary A., Brun M., Hua J., Lowey J., Suh E., and Dougherty E.R. Genetic test bed for feature selection. *Bioinformatic*, 22(7):837–842, 2006.

Dougherty E.R. and Brun M. On the number of close-to-optimal feature sets. *Cancer Informatics*, 2:189–196, 2006.

Dougherty E.R., Shmulevich I., and Bittner M.L. Genomic signal processing: The salient issues. *Applied Signal Processing*, 4(1):146–153, 2004.

Dudoit S., Shaffer J.P., and Boldrick J.C. Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18(1):71–103, 2003.

Hashimoto R.F., Dougherty E.R., Brun M., Zhou Z.Z., Bittner M.L., and Trent J.M. Efficient selection of feature sets possessing high coefficients of determination based on incremental determinations. *Signal Processing*, 83(4):695–712, 2003.

Hedenfalk I., Duggan D., Chen Y., Radmacher M., Bittner M.L., Simon R., Meltzer P.S., Gusterson B., Esteller M., Raffeld M., Yakhini Z., Ben-Dor A., Dougherty E.R., Kononen J., Bubendorf L., Fehrle W., Pittaluga S., Gruvberger S., Loman N., Johannsson O., Olsson H., Wilfond B., Sauter G., Kallioniemi O.P., Borg A., and Trent J.M. Gene-expression profiles in hereditary breast cancer. *New England Journal of Medicine*, 344(8):539–548, 2001.

Ira C., Qi T., Zhou X.S., and Thomas S.H. Feature selection using principal feature analysis. *ICIP 2002, ochester, New York, USA.*, 2002.

Isabelle G. and Andre E. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

Quackenbush J. Microarray data normalization and transformation. *Nat Genet*, 32:supplement pp496–501, 2002.

Tabus I. and Astola J. *Gene Feature Selection*, pages 67–92. EURASIP Book Series on Signal Processing and Communications. Hindawi Publishing Corporation, 2005.

Tao L., Chengliang Z., and Mitsunori O. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*, 20:2429–2437, 2004.

Yu L. and Liu H. Redundancy based feature selection for microarray data. *KDD '04, August 22-25, 2004, Seattle, Washington, USA*, 2004.