Selección de compuestos para la tipificación, un problema de selección de características usando GA's y PCA, caso Virus del Papiloma Humano

Mariano Rodríguez-Muro, Pilar Gómez-Gil, Carolina Castañeda-Roldán,
Javier Garcés-Eisele, Mauricio Osorio-Galindo
Universidad de las Américas-Puebla
Sta. Catarina Mártir, Cholula, Puebla.
72820 México
mariano.muro@gmail.com, mariap.gomez@udlap.mx,
{ccastane, jgarces, josorio}@mail.udlap.mx

Abstract

Se presenta el problema de la selección de compuestos para la tipificación de muestras como un problema de selección de características. Se propone el uso del Análisis de Componentes Principales y de Algoritmos Genéticos como técnicas de selección de características para ser aplicados sobre el problema de selección de compuestos. Se presentan resultados comparativos de las dos técnicas aplicadas sobre el caso de la selección de enzimas para la genotipificación del Virus del Papiloma Humano (VPH).

1. Introducción

Describimos el problema de la selección de reactivos para la identificación de muestras como sigue: dado un conjunto A de reactivos, donde cada reactivo solamente permite la identificación parcial de las muestras m, encontrar un subconjunto $a \in A$ de tamaño mínimo que permita una identificación total de las muestras. Éste es un problema general que encuentra ejemplificación en muchos subproblemas, como la *selección de enzimas para la genotipificación del Virus del Papiloma Humano* o la *selección de sondas para la tipificación de donadores* para el área de bioinformática.

Completando la definición del problema decimos que, si tenemos un conjunto de clases $T = (t_1, t_2, \ldots, t_n)$ a los que puede pertenecer una muestra, y una función $clasificar(a, m) = t$, donde a es un subconjunto de reactivos $a \in A$, m es una muestra que se desea clasificar y t es el conjunto de clases $t \in T$ a los que puede pertenecer la muestra m basándose en la información aportada por los reactivos contenidos en a, entonces $|clasificar(a, m)| > 1$ y $|clasificar(a, m)| = 1$ corresponden respectivamente a una clasificación parcial y a una total, para toda posible m.

Atacamos la selección de compuestos como un problema de *selección de características* debido a la similitud de sus definiciones, como veremos a continuación. Aplicamos dos algoritmos de selección de características diferentes. El primero, el análisis de componentes principales (PCA), que no se usa frecuentemente como técnica de selección de características, pero deseamos investigar su comportamiento en este problema para explorar las diferentes capacidades de este método. El segundo, los algoritmos genéticos, son comúnmente usados para resolver problemas de optimización combinatoria similares al problema de selección de características y también para problemas específicos de selección de características por lo que aplicarlos al problema de la selección de compuestos resulta natural.

Como dijimos anteriormente, la selección de compuestos es un problema general que tiene diferentes áreas de aplicación. Como caso de trabajo tenemos el problema de la *selección de enzimas para la genotipificación del Virus del Papiloma Humano (VPH)*, un problema del área de bioinformática del cual conocemos la solución óptima y que nos sirve como guía para evaluar nuestros algoritmos.

2. Trabajo relacionado

Los trabajos anteriores en nuestro grupo de investigación se centraron en resolver el problema de la optimización de la genotipificación del VPH sin formalizarlo como un problema de selección de características. En el contexto de la selección de enzimas para la tipificación del VPH, se ha demostrado que encontrar el número mínimo de enzimas necesarias para identificar un grupo de virus, como el del VPH, es un problema NP-completo [7]. Sin embargo, puede atacarse heurísticamente de diferentes maneras. En [3] se utiliza la Teoría de la Información de Shannon para crear

un algoritmo de selección basándose en la cantidad de información contenida por enzima. En [8] se utiliza un algoritmo ávido y un algoritmo ramifica-limita para realizar la selección.

3. Selección de características

Definición 1 [11]. Un problema de selección de características o variables es una tupla $\langle X, \Phi, T, A, M \rangle$, donde X es una muestra de patrones de entrada definidos sobre un conjunto de características Φ, $T \in \Phi$ es una variable objetivo, A es un algoritmo de clasificación que produce un modelo de predicción para T dado T y X; y M es una métrica de desempeño del modelo del clasificador y de las características seleccionadas. Una solución al problema es un subconjunto de características $\phi \subseteq \Phi$ que maximiza $M(\phi, A(T, X \downarrow \phi))$, donde $X \downarrow \phi$ es la proyección de los datos X sobre las características pertenecientes a ϕ. En esta métrica M, generalmente deseamos minimizar $|\phi|$ y maximizar el desempeño de T para el problema específico.

En palabras más simples, selección de características es el problema de encontrar, dado un conjunto de características, un subconjunto de éstas ϕ tal que éste maximice cierta(s) propiedad(es) deseadas. En el caso de nuestra investigación estas propiedades son: que el subconjunto sea de tamaño mínimo y que éste permita una clasificación total.

4. Caso de trabajo: selección de enzimas para la clasificación del VPH

Contamos con la información arrojada por 205 enzimas de restricción para 48 tipos del virus el papiloma humano. El objetivo es obtener una clasificación total para los 48 tipos virales.

La información está en forma de las distancias, en milímetros, recorridas por los fragmentos de los patrones de restricción generados por las enzimas. Esta información fue generada por el Dr. Garcés-Eisele a partir de los datos de las cadenas genéticas de los 48 tipos de HPV y de los datos sobre los puntos de corte de las 205 enzimas de restricción con las que trabajamos.

Si tenemos dos tipos de HPV, V_{t1}, V_{t2}, y la información de dos patrones de restricción para estos tipos virales P_{t1}, P_{t2}, se considera que P_{t1} y P_{t2} permiten diferenciar entre V_{t1} y V_{t2} si:

1. El número de fragmentos en P_{t1} es diferente del número de fragmentos en P_{t2}
2. Siendo el número de fragmentos iguales y estos estando ordenados con base en su aparición en el patrón de restricción, al menos uno de los fragmentos en P_{t1} difiere de su correspondiente en P_{t2} en al menos un milímetro.

5. PCA para la selección de características

El Análisis de Componentes Principales (Principal Components Analysis, PCA) es una de las técnicas de análisis multivariante más sencillas [9],[1]. Su objetivo es el de, dado un conjunto de variables $Y_1, Y_2, Y_3, \ldots, Y_p$, encontrar combinaciones lineales de éstas que produzcan nuevos índices $Z_1, Z_2, Z_3, \ldots, Z_p$, llamados componentes principales, tal que éstos no estén correlacionados y la varianza esté maximizada en cada uno de ellos. Es una técnica comúnmente usada para la reducción de la dimensionalidad de datos de entrada.

Aunque no es su función principal, en [9], [4], [6] se menciona el uso de esta técnica para evaluar la importancia de las variables en un conjunto de datos. También, en [10] y [2] se hace referencia al PCA explícitamente como técnica de selección de características.

Para utilizar PCA como técnica de selección de características es necesario analizar como se calculan las nuevas variables. Cuando se aplica el PCA sobre un conjunto de datos Y_{np}, se obtiene una matriz de transformación W_{mp} donde $m = p$ y cada renglón m contiene los coeficientes de la ecuación que permite calcular el m-ésimo componente principal. El cálculo de cada componente principal se hace con un polinomio lineal como el siguiente: $Z_m = W_{m1}Y_1 + W_{m2}Y_2 + \ldots + W_{mp}Y_p$ donde Z_m es el m-ésimo componente principal y Y_p es la p-ésima característica o variable. Los coeficientes W_{mp} equivalen a la participación que tiene cada característica Y_p para el cálculo de cada componente Z_m. Si sabemos que ciertos componentes $\zeta = Z_{p1}, Z_{p2} \ldots, Z_{pn}$ contienen la varianza suficiente para permitir una clasificación total entonces podemos asegurar que la selección de las características Y que participan en el cálculo de éstos también permitirá una clasificación total. Para asegurar que $|\phi| < |\Phi|$ agregaremos características Y_p a ϕ una a una, agregando primero las que más participación tengan en el cálculo de los componentes en ζ y verificando la capacidad de clasificación del conjunto ϕ con cada adición..

6. Algoritmos Genéticos para la selección de características

Para la codificación del problema se utiliza un cromosoma $\overline{X}$ en forma de cadena binaria de tamaño n, donde n es el número de características en el conjunto total. Cada bit de la cadena binaria está relacionado con una característica. Si el correspondiente bit es igual a 1, la característica relacionada estará seleccionada para pertenecer a ϕ y viceversa.

Tabla 1. Resultados comparativos para el caso de la selección de enzimas para la tipificación del VPH

Tipo de Algoritmo	No. de Enzimas en el conjunto mínimo
Búsqueda exhaustiva	2
PCA	11
Algoritmos Genéticos	4

La función de evaluación o función objetivo está en la forma de una suma ponderada donde $\sum_{obj=1}^{N}(W_{obj}) = 1$ para los N objetivos. Además de los dos objetivos implícitos en el problema de selección de compuestos, agregamos un objetivo más para el caso de nuestro caso de trabajo, la tipificación de VPH, éste es la maximización de la calidad de las enzimas contenidas en ϕ. Todos las funciones que implementan los objetivos mencionados arrojan valores en el mismo rango. La función objetivo queda de la siguiente manera: $F(\overline{x}) = w_{count}f_{count}(\overline{x}) + w_{sep}f_{sep}(\overline{x}) + w_{qlty}f_{qlty}(\overline{x})$ donde $\overline{x}$ es un cromosoma que es una instancia de $\overline{X}$, $f_{count}(\overline{x})$ es el número de características seleccionadas, $f_{sep}(\overline{x})$ es la separabilidad de las clases a distinguir y $f_{qlty}(\overline{x})$ es la calidad de dichas características.

El algoritmo utilizado mantiene la estructura del Algoritmo Genético Simple como se describe en [5]. Para la selección de individuos se probaron los algoritmos de selección por ruleta y el muestreo universal estocástico. Para la reproducción se probó el cruzamiento en uno y dos puntos con y sin remplazo reducido. Para la creación de la población inicial se probó la inicialización al azar uniforme y la inicialización con individuos vacíos.

Debido a la naturaleza estocástica del algoritmo, para poder establecer un promedio de desempeño es necesario que éste sea ejecutado repetidas veces por cada configuración de sus parámetros. Así pues, cada configuración de los parámetros del algoritmo fue repetida 10 veces.

7. Resultados, Conclusiones y Trabajo a Futuro

Como podemos observar en la tabla 1, entre las dos técnicas probadas, la que mejores resultados arrojó fueron los algoritmos genéticos. Si bien el resultado no es el óptimo, no está muy alejado de éste, además de que se generan muchas soluciones a la vez, lo que agrega flexibilidad. El PCA arroja buenos resultados si comparamos con el tamaño del conjunto original, alcanzando una reducción en tamaño del 94 %. En ninguno de los casos se alcanzó el tamaño mínimo reportado por la búsqueda exaustiva.

Como trabajo a futuro, ambas técnicas serán aplicadas sobre el caso de la *selección de sondas para identificación de tipos de donadores de órganos para transplantes*.

Además, se observó para el caso de la selección utilizando PCA, que la información aportada por la matriz de transformación W es insuficiente para alcanzar resultados óptimos. Por esto se propone, para alcanzar soluciones óptimas, combinar esta información con un algoritmo de búsqueda ávido que utilice información adicional sobre el conjunto original de datos.

En el caso de los algoritmos genéticos, se observó que el mayor obstáculo para alcanzar soluciones óptimas es el dominio del espacio de búsqueda por unas cuantas soluciones superiores al promedio durante el ciclo evolutivo. Aunque estas soluciones son buenas en comparación con las demás soluciones actuales, los procesos de selección utilizados provocaban que el espacio de búsqueda se limitara a zonas cercanas a éstas. Es necesario buscar técnicas de selección alternativas, como la selección por torneo, que garanticen la diversidad en el espacio de búsqueda.

Referencias

[1] M. Á. Carreira-Perpiñán. A review of dimension reduction techniques. Technical Report CS–96–09, Dept. of Computer Science, University of Sheffield, January 1997.

[2] I. Cohen, Q. Tian, X. Zhou, and T. Huang. Feature selection using principal feature analysis, 2002.

[3] R. Fernández. Desarrollo de algoritmos para la clasificación de secuencias. Master's thesis, Universidad de las Américas, Puebla, 2002.

[4] B. Flury and H. Riedwyl. *Multivariate Statistics: A practical approach*. Chapmawn and Hall, 1988.

[5] David E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Weley publishing company, Inc, 1989.

[6] W. J. Krzanowski. *Principles of Multivariate Analysis*. Oxford statistical science series. Oxford University Press, 1988.

[7] J. Lozano-Yécora, J. Garcés-Eisele, M. Osorio-Galindo, C. Castañeda-Roldán, and P. Gómez. Genotyping of human papilloma virus (hpv): a combinatorial optimization family of np-hard problems. Documento interno de trabajo. Universidad de las Américas-Puebla, 2004.

[8] J. M. Lozano-Yécora. La genotipificación del virus del papiloma humano: una familia de problemas de optimización combinatoria. Master's thesis, Universidad de las Américas, Puebla, 2004.

[9] B. F. J. Manly. *Multivariate Statistical Methods: A Primer*. Academic Press, 1986.

[10] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Probability and Mathematical Statistics. Chapman and Hall, 1995.

[11] I. Tsamardinos and C. Aliferis. Towards principled feature selection: Relevancy, 2003.