Selección de atributos en clasificación supervisada. Uso de la entropía condicional

Romero, María del Carmen
Di Rienzo, Julio Alejandro
Clausse, Alejandro

Ponencia presentada en el IV Encuentro Iberoamericano de Biometría y XVIII Reunión Científica del Grupo Argentino de Biometría. Mar del Plata, Argentina, 25 al 27 de septiembre de 2013

SELECCIÓN DE ATRIBUTOS EN CLASIFICACIÓN SUPERVISADA. USO DE LA ENTROPÍA CONDICIONAL

MARÍA DEL CARMEN ROMERO[1], JULIO A. DI RIENZO[2] Y ALEJANDRO CLAUSSE[3]

[1]Fac. Cs. Económicas, Universidad Nacional del Centro de la Provincia de Buenos Aires
[2] Estadística y Biometría, Fac. Ciencias Agrarias, Universidad Nacional de Córdoba
[3]PLADEMA, Fac. Cs. Exactas, Universidad Nacional del Centro de la Provincia de Buenos Aires

mariadelc.romero@gmail.com

RESUMEN

Las bases de datos de alta dimensionalidad pueden encontrarse en diferentes áreas de conocimiento. Los datos provenientes de microarreglos de ADN son buenos representantes de estos contextos y tienen, además, la particularidad de poseer mayor cantidad de atributos que de observaciones. Si bien, la clasificación supervisada suele ser una de las técnicas más usadas en estos casos, el "ruido" debido a las particularidades expuestas provocan que los clasificadores convencionales tengan resultados inestables. En este trabajo se propone el uso de la entropía condicional como medida para realizar la selección del subconjunto de atributos que distingan entre tratamientos en contextos de microarreglos de ADN. La entropía mide la cantidad media de información que es necesaria proveer para no tener incertidumbre sobre una variable determinada y tiene la ventaja de poder aplicarse a contextos con variables pertenecientes a cualquier escala de medición. Se desarrolló un algoritmo en R y se simularon diferentes escenarios de microarreglos de ADN. Las conclusiones se obtuvieron considerando el tamaño promedio del subconjunto seleccionado y el porcentaje de atributos seleccionados que efectivamente son diferenciales. Entre los resultados preliminares puede mencionarse que: en la mayoría de los casos, la entropía condicional con el subconjunto de atributos seleccionados es 0; a mayor cantidad de réplicas, mayor es el tamaño del subconjunto y mayor el porcentaje de atributos efectivamente diferenciales; y que, a mayor porcentaje de atributos diferenciales, menor es el tamaño del subconjunto de atributos seleccionados y mayor es el porcentaje de atributos efectivamente diferenciales.

Palabras clave: ***alta dimensionalidad, microarreglos de ADN, incertidumbre, simulación***

Introducción

Los avances tecnológicos de las últimas décadas facilitaron la creación de grandes bases de datos de alta dimensionalidad (gran cantidad de atributos), lo cual involucra no sólo grandes exigencias en la adquisición de los datos sino también en el almacenamiento, el mantenimiento, y sobre todo, el procesamiento de los mismos. En 1996, Fayyad y otros, introdujeron la estrategia de análisis conocida como descubrimiento de conocimiento en bases de datos, la cual involucra métodos y técnicas para extraer conocimiento de alto nivel a partir de grandes conjuntos de datos de bajo nivel. Comprende el proceso completo de descubrir conocimiento a partir de los datos, incluyendo los pasos que van desde la obtención de los mismos hasta la aplicación del conocimiento adquirido. Lo esencial de este proceso es la minería de datos, se define como los métodos para el descubrimiento y la extracción de patrones de relación entre atributos a partir de los datos. Otros autores, como Kantardzic (2011) no la definen como componentes del proceso de extracción de conocimiento sino como el proceso en sí mismo.

La clasificación supervisada suele ser una de las técnicas más usadas en contextos de alta dimensionalidad. Sin embargo, los clasificadores convencionales producen resultados

inestables en aplicaciones (cada vez más frecuentes) en las cuales la cantidad de atributos es superior a la cantidad de observaciones. En estos casos, el "ruido" suele ser un denominador común, provocando la sobreparametrización y la propensión al sobreajuste de los clasificadores convencionales.
El área genómica es un contexto con las particularidades descriptas, en el cual la clasificación supervisada suele ser un objetivo. Ante este requerimiento, se expone un método que usa la entropía condicional y algunos resultados preliminares. El desarrollo se realiza en el contexto de análisis de expresión génica.

Desarrollo

Datos

Datos continuos, provenientes de experimentos con matrices de expresión génica. Se caracterizan por tener gran cantidad de atributos (genes), baja cantidad de observaciones (cantidad de tratamientos * cantidad de réplicas) y una fracción pequeña de genes diferenciales. El objetivo es identificar genes con respuesta diferencial entre tratamientos.

		Clase 1			Clase 2		
		R_{11}	R_{12}	R_{13}	R_{21}	R_{22}	R_{23}
Gen	1	Y_{111}	Y_{121}	Y_{131}	Y_{211}	Y_{221}	Y_{231}
	2	Y_{112}	Y_{122}	Y_{132}	Y_{212}	Y_{222}	Y_{232}
	…	…	…	…	…	…	…
	N	Y_{11N}	Y_{12N}	Y_{13N}	Y_{21N}	Y_{22N}	Y_{23N}

Las "clases" hacen referencia a los tratamientos. Las "réplicas" a la cantidad de individuos (perfiles genéticos) que hay dentro de cada grupo. R_{ij} donde *i* indexa la clase y *j* la repetición dentro de la clase. Las expresiones génicas están denotadas por Y_{ijk}, i indexa la clase, j la réplica y *k* indexa al gen.

Medida de evaluación: Entropía condicional

La optimalidad de un subconjunto siempre es relativa a una medida de evaluación. La entropía puede definirse como la cantidad media de información que es necesaria proveer para no tener incertidumbre sobre una fuente determinada: $H(X) = -\sum_i p(x_i).log_2\, p(x_i)$, X representa la variable aleatoria, p(xi) la probabilidad de ocurrencia de cada uno de sus valores y el logaritmo en base 2 considera que la información se representará mediante código binario (bits). La entropía condicional referencia la entropía que tiene una determinada variable Y (en el caso de la aplicación, el tratamiento) conociendo la información que aporta otra variable X (genes).

$$H(X/Y) = -\sum_y p(y) \sum_x p(x/y)\,.log_2 p(x/y)$$

En el contexto de la aplicación, si la entropía (incertidumbre) del tratamiento disminuye por el conocimiento de algún gen, dicho gen es relevante para la distinción entre tratamientos.

Algoritmo

El objetivo es generar un subconjunto de atributos que permita la clasificación supervisada (basado en la entropía condicional). Se implementó un algoritmo que consiste en agregar incrementalmente al subconjunto de atributos seleccionados aquéllos que minimicen la entropía condicional de la clase. Esta evaluación se realiza considerando los atributos ya ingresados en dicho subconjunto y el ciclo termina cuando se obtiene una entropía igual a 0 o cuando no logra mejorarse la entropía del ciclo anterior.
La generación de datos se realizó a través de un algoritmo desarrollado por Di Rienzo (2006), que permite especificar la cantidad de genes, tratamientos, réplicas, y el porcentaje de genes diferenciales y evaluar, en una instancia posterior, el rendimiento de los métodos.
Se generaron 1000 repeticiones para 6 escenarios dados por los siguientes parámetros: 1000 genes, 2 tratamientos, 5 y 10 réplicas biológicas y 5, 10 y 30% de genes diferenciales.
Debido a la continuidad de los datos y a que es necesaria la discretización de los mismos para la aplicación de la entropía condicional se consideraron distintas alternativas de discretización: en 2, 5, 10 y 20 intervalos de igual amplitud.
Se desarrolló un programa en R y se utilizó el package entropy de R (Hausser y Strimmer, 2013). Se trabajó con la función "entropy" la cual estima la entropía Shannon de una variable aleatoria a partir de frecuencias observadas.

Por cada "simulación" (se realizaron 1000 en total)

1) Generar la matriz de expresión génica MEG (Di Rienzo, 2006)
   Atributos (genes) = $x_1, x_2, \ldots, x_n$, n = cantidad de atributos. Clases: C={1,2}
2) Selección del subconjunto de atributos diferenciales
   Paso 1: Inicializar el conjunto de atributos seleccionados S como vacío:S={ϕ}
   Paso 2: Ciclo que se repite mientras se tenga una entropía condicional H(clase/xi ∪ S) mayor a 0 o mientras no logre mejorarse (disminuirse) la entropía del ciclo anterior)
   a. Para cada uno de los atributos no incluidos en S, calcular las entropías condicionales de la Clase considerando la inclusión de ellos en el subconjunto resultante S: H(clase/xi ∪ S)
   b. Seleccionar el atributo xi cuya H(Clase/xi) sea la mínima (conociendo al atributo xi, se tiene menor incertidumbre sobre la clase C) y agregarlo al subconjunto S = S ∪ x_i

Resultados y Conclusiones

El comportamiento de la entropía condicional se expone, para todos los escenarios propuestos, considerando el tamaño promedio del subconjunto seleccionado y el porcentaje de atributos seleccionados que efectivamente son diferenciales (Tabla 1). Entre las principales conclusiones:

- En la mayoría de los casos, la entropía condicional con los atributos seleccionados es 0.
- Cuanto mayor es la cantidad de intervalos en los cuales se discretiza, menor es el tamaño del subconjunto seleccionado y mayor es el porcentaje de atributos diferenciales (al discretizar en mayor cantidad de intervalos, mayor precisión en los datos).
- A mayor cantidad de réplicas, mayor es el tamaño del subconjunto y mayor el porcentaje de atributos efectivamente diferenciales (al tener más réplicas, se tiene más información)
- Mayor porcentaje de atributos diferenciales, menor es el tamaño del subconjunto de atributos seleccionados y mayor es el porcentaje de atributos diferenciales.

Tabla 1: Tamaño promedio del subconjunto seleccionado (tamaño) y Porcentaje de atributos seleccionados que efectivamente son diferenciales (%dif)

Discretización en 2 intervalos					**Discretización en 5 intervalos**			
	5 réplicas		10 réplicas		5 réplicas		10 réplicas	
	Tamaño	% dif	Tamaño	% dif	Tamaño	% dif	Tamaño	% dif
5% dif	3.37	24.07	4.85	26.20	2.71	44.45	4.49	51.28
10% dif	3.34	30.52	5.30	38.14	2.66	54.37	4.56	67.51
30% dif	2.93	52.45	5.01	57.68	2.32	74.07	3.90	85.21
Discretización en 10 intervalos					**Discretización en 20 intervalos**			
	5 réplicas		10 réplicas		5 réplicas		10 réplicas	
	Tamaño	% dif	Tamaño	% dif	Tamaño	% dif	Tamaño	% dif
% dif	2.44	64.30	3.71	72.88	1.89	73.00	2.67	77.75
10% dif	2.08	65.75	3.49	82.80	1.62	76.67	2.22	84.08
30% dif	1.88	82.5	2.77	92.58	1.31	93.50	1.99	85.33

Este trabajo muestra resultados preliminares sobre el uso de la entropía condicional para la selección de atributos diferenciales. Se prevé la generación de contextos con mayor cantidad de atributos de diferentes escalas de medición, para evaluar el rendimiento de la medida propuesta.

Bibliografía

[1] Di Rienzo, J. y Romero, M. (2006). "Simulación de matrices de expresión génica de experimentos con micromatrices de ADN". Jornadas Internacionales de Estadística. Rosario, Argentina.

[2] Fayyad, U.; Piatetsky-Shapiro, G. y Smyth, P. (1996). "From Data Mining to Knowledge Discovery in Databases". *Artificial Intelligence Magazine*, 17(3): 37-54.

[3] Hauser, J. y Strimmer, K. (2013). Package Entropy: Estimation of Entropy, Mutual Information and Related Quantities. R package, versión 1.2.0. Versión obtenida en abril de 2013. http://strimmerlab.org/software/entropy

[4] Kantardzic, M. (2011). *Data Mining: Concepts, Models, Methods, and Algorithms.* IEEE Press & John Wiley. Segunda Edición. Agosto, 2011.