The most probable annotation problem in HMMs and its application to bioinformatics

Víctor Alfonso Rivera Zúñiga
Pontificia Universidad Javeriana
Seminario BioInformática

November 12, 2008

1. Introducción
2. BackGround
3. Problema de los Múltiples Caminos
   - Modelos Ocultos de Markov (HMM)
4. Cómputo de la anotación más probable
   - Extensión del algoritmo de Viterbi
   - Prueba
   - Condición de la arista crítica
5. Conclusiones

- Los Modelos Ocultos de Markov (HMMs) se utilizan a menudo para anotación de secuencias biológicas.
- Cada característica en la secuencia está representada por una colección de estados con la misma etiqueta.
- Al describir una nueva secuencia, se busca la secuencia de etiquetas que con más alta probabilidad la describe.
- Se demostró que el cálculo de la anotación más probable es NP-hard, por Lyngso y Pedersen

- Con HMMs el objetivo es identificar caractersticas funcionales dada una secuencia biológica.
- Computar la anotación con máxima probabilidad de una secuencia es equivalente a encontrar la máxima probabilidad de camino de estados para esa secuencia. Para esto se utiliza el cásico algoritmo de Viterbi.
- Sin embargo, este no es el caso de muchos HMMs usados para aplicaciones en bioinformática, donde múltiples caminos de estados corresponden a la misma anotación.

Modelo Oculto de Markov (HMM)

- Un modelo oculto de Markov es un modelo probabilístico compuesto de estados y transiciones.
- Es una tupla $M = \{V, \Sigma, a, e, \Pi\}$
  - V es el conjunto de estados.
  - Σ es el conjuto finito de símbolos.
  - a(v,w) Es la probabilidad de transición desde el estado v al estado w.
  - Para cada estado v $\in$ V, $e_v(x)$ es la probabilidad de emisión del símbolo x estando en el estado v.
- Cuando se usa HMM para anotación de secuencias, se etiqueta cada estado con la característica a la cual corresponde.
  - $\lambda(v)$ Denota la etiqueta del estado v
  - $in(u)$ Para cualquier estado u sea el conjunto de estados con transición al estado u

Dada una secuencia $x_i, ..., x_n$, una secuencia de anotación es una secuencia de n etiquetas.

Cada camino de estados $\pi_1, ..., \pi_n$ corresponde naturalmente a su anotacion $\lambda(\pi_1), ..., \lambda(\pi_n)$.

Sin embargo, este mapeo no siempre es 1 a 1 (muchos caminos de estados pueden corresponder a la misma anotación), esto se denomina el problema de caminos múltiples.

La probabilidad de una anotación dada $L = \lambda_1, ..., \lambda_n$ es la suma de las probabilidades de todos los caminos de estados $\pi_1, ..., \pi_n$ con anotación L.
El problema de la anotación más probable es encontrar la anotación L con máxima probabilidad.

- Con un HMM se pueden responder a 3 problemas.
  - Uno de ellos es determinar la secuencia de estados que con mayor probabilidad reconoce una secuencia de observaciones.
- Evidentemente este es el problema a tratar en la anotación más probable dada una secuencia biológica. Se aborda por medio del conocido algoritmo de Viterbi.
- Sin embargo, se ha descubierto por experimentación (...) que la secuencia de estados que con mayor probabilidad reconoce las observaciones dadas no es necesariamente la anotación más probable.
- Para un HMM donde sólo existe un único camino de estados que corresponden a una única anotación, el algoritmo de Viterbi determina la secuencia de estados que con mayor probabilidad reconoce las observaciones y es, sin duda, la anotación más probable. Sin embargo, en la vida real existen modelos con múltiples caminos.

- Se sabe que computar la anotación más probable es un problema NP-Hard para algunos HMMs (Demostrado por Lyngso y Pedersen)
- Sin embargo, para clases especiales de HHMs, la anotación más probable puede ser computado eficientemente.
- Un ejemplo de ello, HMMs sin el problema de los caminos múltiples. Sabemos que hay un camino de estados por cada anotación. El algoritmo de Viterbi encuentra el camino con máxima probablidad en $O(nm^2)$ tiempo, donde n es la longitud de la secuencia y m es el número de estados.

Anotación extendida

Una arista crítica es la transición entre dos estados con diferente etiqueta. La anotación extendida de un camino de estados $\pi_1, ..., \pi_n$ es la pareja (L,C), donde $L = \lambda_1, \lambda_2, ..., \lambda_n$ es la secuencia de etiquetas para cada estado en el camino y C $= c_1, c_2, ..., c_k$ es la secuencia de todas las aristas críticas en el camino.

Fig. An HMM with two critical edges, $A \rightarrow B$ and $B \rightarrow A$.

La figura tiene tres estados, dos etiquetas diferentes y dos aristas críticas: $A \rightarrow B$ y $B \rightarrow A$. Para el camino de estados ABCB la anotación extendida sería: (b-g-g-g,$A \rightarrow B$).

Teorema Extensión del algoritmo de Viterbi (EVA)

Para una secuencia dada $S = x_1, .., x_n$ y un HMM con m estados, es posible computar la extensíon de la anotación más probable de S en tiempo del orden de $O(n^3 m^2)$

La extensión del algoritmo de Viterbi no siempre encuentra la anotación más probable, sin embargo retornará una anotación con probabilidad al menos tan alta como la probabilidad del camino de estados más probable.

En el algoritmo de Viterbi computa el camino de estados más probable usando programación dinámica. Lo que se hace es computar V[u,i], es la máxima probabilidad de caminos de estados para la secuencia $S = x_1, .., x_n$ sobre todos los caminos que terminen en el estado u.

$maxPr(x_1, .., x_n, \pi_i = s, \pi_2, ..., \pi_{n-1}, \pi_i = u)$

Para computar el valor de V[u,i] el algoritmo de programación dinámica usa la siguiente recurrencia, examina todas las posibles opciones desde el segundo estado:

$$V[u, i] = \max_{v \in in(u)} V[v, i-1] * a(v, u) * e_u(x_i)$$

En el algoritmo extendido de Viterbi (EVA), se computa de la siguiente manera:
$L[u, i] = maxPr(x_1, .., x_n, (L, C), \pi_i = u)$, donde el máximo es tomado de todas las anotaciones extendidas (L,C) de la secuencia $x_1, ..., x_i$ donde el proceso de generación empieza en el estado s y finaliza en el estado u.

Se examina todas las posibles anotaciones del último segmento con la misma etiqueta y en vez de escoger el único camino más probable, se computa la suma de todos los posibles caminos en ese segmento. Esta suma se denotará como P[v,u,j,i]; Si el segmento empieza en la posisción $j \leq i$ de la secuencia.
Esto es la probabilidad de generar la secuencia $x_j, .., x_i$ empezando en el estado v, terminando en el estado j y usando sólo estados con la misma etiqueta $\lambda(y)$

La siguiente es la ecuación de recurrencia.

$$L[u,i] = \max_{j \leq i} \max_{v:\lambda(v)=\lambda(u)} \max_{w \in in(v):\lambda(w)\neq\lambda(v)} (L[w, j-i] * a(w,v) * P[v,u,j,i]$$

Se computa el valor de L en el order de incrementar i. Por cada i, se computa todos los valores relevantes de P[v,u,j,i] en el orden en que j decrezca, usando la siguiente recurrencia:

$$P[v,u,j,i] = \sum_{w:v \in in(w), \lambda(v)=\lambda(w)} (e_v(x_j) * a(v,w) * P[w,u,j+1,i]$$

Al terminar el cómputo de L, la extensión de la anotación más probable se puede reconstruir usando backtracking.

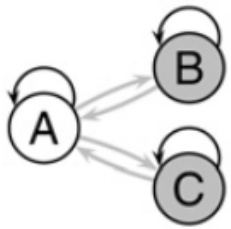
Definición

Un HMM satisface la condición de la arista crítica para una entrada de secuencia s si alguno de dos caminos con la misma anotación tiene la misma secuencia de aristas críticas. Un HMM satisface la condición de la arist crítica, en general, si para todo secuencia de entrada s, la condición de la arista crítica se satisface.

Corolario

Si un HMM satisface la condición de la arista crítica para una secuencia s, entonces la extensión del algoritmo de Viterbi (EVA) computa la anotación más probable de la secuencia s.

Para el camino de estados ABCB la anotación extendida (L,C) sería: (b-g-g-g,$A \rightarrow B$).
Y para el camino de estados ABBB la anotación extendida (L,C) sería: (b-g-g-g,$A \rightarrow B$).



Para el camino de estados ABB la anotación extendida (L,C) sería: (b-g-g,$A \rightarrow B$).
Y para el camino de estados ACC la anotación extendida (L,C) sería: (b-g-g,$A \rightarrow C$).

Algoritmo

- Se calcula S_s. Todas las parejas de estados que son alcanzables desde el estado inicial con la misma anotación.
  - Se empieza con la pareja $(s, s) \in S_s$ y se va adicionando nuevas parejas (u,v) si $\lambda(u) = \lambda(v)$ y si existe $(v', u') \in S_s talqueu' \in in(u)yv' \in in(v)$.
- Se hace lo mismo para S_f. Todas las parejas de estados que el estado final sea alcanzable por la misma anotación.
- Para que la condición crítica sea violada, debe existir una pareja $(u, v) \in S_s y(u', v') \in S_f$ tal que $\lambda(u) \neq \lambda(u')$, y (u,u') y (v,v') sean dos transiciones diferentes.
- El algoritmo toma tiempo del orden de O(m4)

- Determinar la anotación más probable en HMMs es un problema NP-Hard. Sin embargo, se puede computar la anotación más probable para muchos HMMs en tiempo polinomial.
  - Si existe un único camino de estados por cada anotación, el problema puede ser resuelto por el algoritmo de Viterbi.
  - Si en un HMM existe el problema de múltiples caminos y en ese HMM se satisface la condición de la arista crítica, el problema puede ser resuelto por la extensión del algoritmo de Viterbi (EVA)
- Sin embargo, hay casos en los que este tiempo polinomial no es factible. e.g El cromosoma mas pequeño humano tiene aproximadamente una longitud de 50 millones de símbolos.

El problema del múltiple camino está cercanamente relacionado con la ambiguedad estructural in Gramáticas Incontextuales Estacásticas (SCFG siglas en Inglés). En bioinformática las gramáticas son comunmente usadas para la predicción de la segunda estructura del RNA. Dowell y Eddy demostraron que la ambiguedad puede causar un deterioro significante en la precisión de la predicción.