Wykład VI: Probabilistyczne metody rekonstrukcji drzew.

Anna Gambin
7.11.2002

1 Maksymalizacja wiarygodności

Zastanówmy się jak możemy znaleźć optymalne drzewo (t.j. drzewo o największej wiarygodności) w przestrzeni wszystkich drzew. Rozważając małą liczbę sekwencji (od 2 do 5) możemy pozwolić sobie na przejrzenie wszystkich drzew i obliczenie dla każdego z nich wiarygodności jako funkcji długości krawędzi. Następnie dla każdej topologii możemy znaleźć maksimum tej funkcji, czyli optymalne długości krawędzi. Dla większej liczby sekwencji potrafimy policzyć wiarygodność algorytmem Felsensteina, istnieją też algorytmy pozwalające znajdować dla danego drzewa optymalną długość krawędzi. Trudność obliczeniowa leży w przeszukiwaniu przestrzeni drzew. Stosuje się tutaj często metody próbkowania.

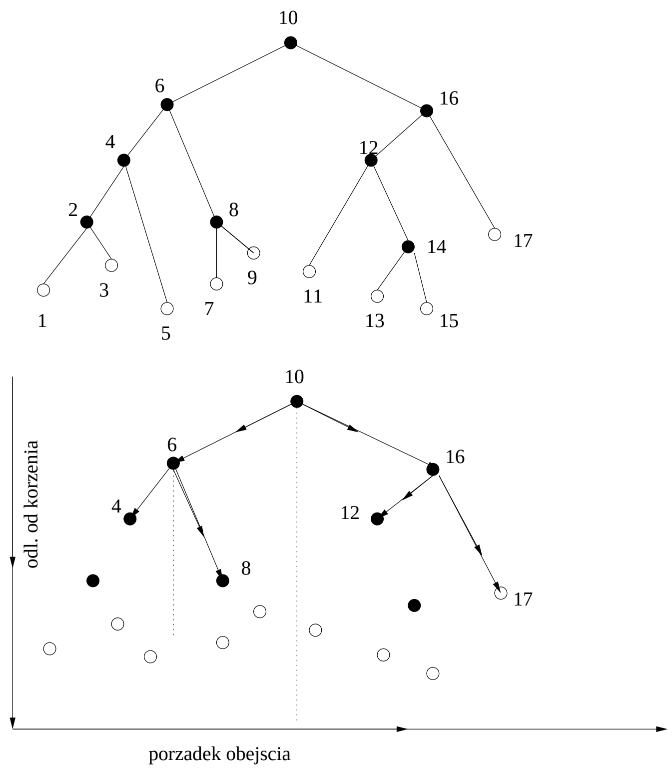
2 Próbkowanie z rozkładu a posteriori

Załóżmy, że znamy rozkład prawdopodobieństwa na ważonych drzewach (prawdopodobieństwo *a priori*): $P(T, \boldsymbol{t})$. Możemy użyć wzoru Bayesa do policzenia prawdopodobieństw *a posteriori* (odwrotnych):

$$P(T, \boldsymbol{t}|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|T, \boldsymbol{t})P(T, \boldsymbol{t})}{P(\boldsymbol{x})}.$$

Prawdopodobieństwo *a posteriori* mówi nam jak prawdopodobny jest dany model (ważona topologia) pod warunkiem danych jakie obserwujemy. Czyli jest wielkością bardzo dla nas interesującą.

Znowu możemy użyć wzoru Bayesa w przypadku niewielkiej liczby topologii. Jeżeli rozważana przestrzeń drzew jest większa można wykorzystać metodę, która losuje drzewo z pewnego rozkładu. Mając dane wyniki wielu prób



Rysunek 1: Przykład drzewa reprezentowanego jako profil obejścia: numery w węzłach spełniają warunek jak dla drzew BST; poniżej rekonstrukcja drzewa z profilu.

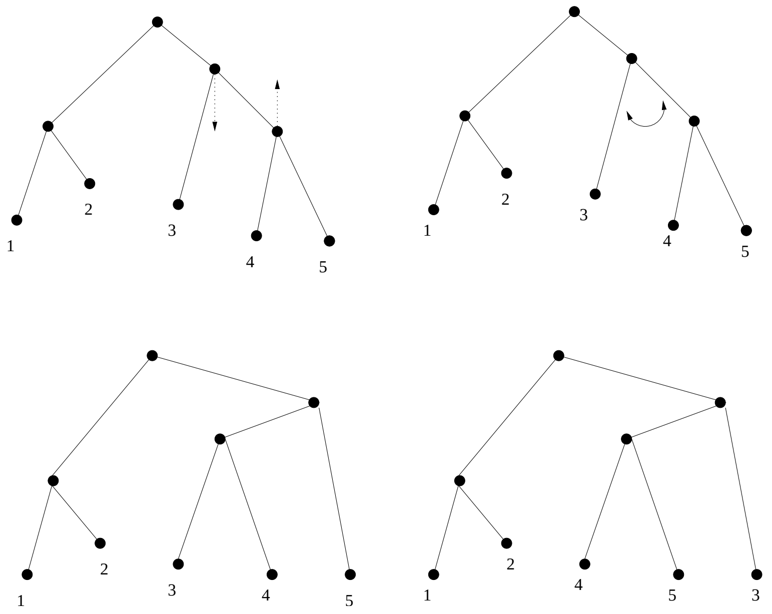
losowych możemy np. estymować prawdopodobieństwo danej topoogii przez frakcję prób, w których wylosowano tę topologię.

Tradycyjnie stosowaną metodą próbkowania jest algorytm Metropolisa, który generuje ciąg drzew, przy czym każde następne drzewo jest dobierane na podstawie poprzednika. Niech $P_1 = P(T, \boldsymbol{t}|\boldsymbol{x})$ oznacza prawdopodobieństwo rozważanego w danej chwili drzewa oraz $P_2 = P(\tilde{T}, \tilde{\boldsymbol{t}}|\boldsymbol{x})$ prawdopodobieństwo nowego drzewa (potencjalnego następnika). Reguła Metropolisa pozwala zaakceptować nowe drzewo jako następne w ciągu o ile $P_2 \geq P_1$. Jeżeli $P_2 < P_1$ nowe drzewo jest akceptowane z prawdopodobieństwem równym P_2/P_1. W przeciwnym przypadku pierwotne drzewo zostaje uznane za kolejny element ciągu. Zauważmy że taka reguła nie wymaga znajomości stałej normującej ze wzoru Bayesa (bo rozważamy ilorazy prawdopodobieństw), której wyliczenie jest trudne, gdyż wymaga całkowania po przestrzeni drzew.

W opisanym podejściu zakłada się że (losowa) procedura proponująca nowe drzewo dla aktualnie rozważanego jest symetryczna, czyli prawdopodobieństwo uzyskania drzewa $\tilde{T}$ z T jest takie samo jak uzyskania drzewa T z $\tilde{T}$.

3 Spacer po przestrzeni drzew

Wybór rozkładu prawdopodobieństwa dla potencjalnych następników danego drzewa jest kluczową decyzją w opisanym algorytmie. Jeżeli następne drzewo bedzie zupełnie losowe, jego prawdopodobieństwo *a posteriori* będzie bardzo małe. Jeśli natomiast kolejny element ciągu będzie bardzo bliski poprzednikowi będziemy potrzebować bardzo długiego spaceru (ciągu drzew) ażeby adekwatnie przepadać przestrzeń drzew. Opiszemy jedną z metod,



Rysunek 2: Dwie fazy modyfikacji drzewa: po lewej zmiana topologi poprzez przesunięcia węzłów z pionie; po prawej zmiana uporządkowania liści.

która składa się z dwóch faz: pierwsza zmienia topologię drzewa i długości krawędzi, druga permutuje sekwencje w liściach. Rysunek 1 wyjaśnia reprezentację drzewa zwaną **profilem obejścia**. Łatwo widać, że z wyższego diagramu możemy w sposób jednoznaczny odtworzyć drzewo. Pierwsza faza modyfikująca drzewo polega na przesunięciu węzłów profilu w góre lub w dół o pewną losową odległość (z rozkładu jednostajnego na ustalonym przedziale). Zauważmy, że zmiana względnej wysokości węzłów pociąga za sobą zmianę topologi (Rys. 2). Druga faza modyfikacji drzewa polega na zamianie porządku gałęzi wychodzących z danego wierzchołka - nie modyfikuje to topologii czyli również nie zmienia prawdopodobieństwa *a posteriori* (zmienia natomiast uporządkowanie liści, co pozwala na modyfikację topologii w kolejnym kroku).

Ze wzoru Bayesa widać, że do zdefiniowania prawdopodobieństw *a posteriori* niezbędny jest nam rozkład prawdopodobieństwa $P(T, \boldsymbol{t})$ na przestrzeni drzew. Często zakłada się rozkład jednorodny przypisujący jednakowe prawdopodobieństwo wszystkim przypisaniom dlugości krawędzi dla każdej topologii. Później poznamy inne rozkłady na przestrzeni drzew, które są indukowane przez model ewolucji (wzrostu) drzewa, np. model Yule.

4 Bootstrapping

Poznaliśmy wiele metod pozwalających na zbudowanie drzewa dla interesujących nas danych. Chcielibyśmy umieć ocenić istotność klasyfikacji według drzewa, na ile badana przez nas cecha filogenetyczna (np. podział zbioru gatunków na gałęzie) jest sensowna.

Pomysł metody znanej jako *ang. bootstrapping* jest następujący: załóżmy, że mamy dane multiuliniowienie pewnej rodziny sekwencji; tworzymy teraz sztuczne dane wybierając losowo (ze zwracaniem) kolumny z tego uliniowienia. Do tak stworzonego uliniowienia stosujemy nasz pierwotny algorytm budowania/poszukiwania drzewa. Cała procedura powtarzana jest wiele razy (rzedu 1000). Częstotliwość z jaką napotykamy na badaną cechę jest uznana za miarę jej istotności.

Bootstrapping może być zastosowany do każdej metody rekonstrukcji drzewa, jednak metody probabilistyczne pozwalają na matematyczną interpretację uzyskanej miary zaufania/istotności danej cechy jako jej przybliżonego prawdopodobieństwa *a posteriori* przy założeniu rozkładu jednorodnego na przestrzeni drzew.

4.1 Bootstrapping w rzutach monetą

Rozważmy bardzo prosty przykład: w N rzutach monetą uzyskaliśmy m razy orła i n razy reszkę. Rozkład *a posteriori* prawdopodobieństwa p wyrzucenia orła (przy jednostajnym rozkładzie *a priori*) jest zadany rozkładem Dirichleta:

$$P(p|m\text{x}O, n\text{x}R) = p^m(1-p)^n \frac{(N+1)!}{m!n!}. \tag{1}$$

Wykorzystując dane z przeprowadzonego raz eksperymentu (zazwyczaj zakładamy, że jest to trudny i kosztowny eksperyment) generujemy dodatkowe sztuczne dane losując kolejne ciągi N rzutów monetą z prawdopodobieństwem orła równym m/N i reszki n/N. Jeżeli w tak wygenerowanym ciągu N rzutów otrzymaliśmy k orłow, estymator największej wiarygodności dla prawdopodobinśtwa p wynosi: $p^* = k/N$. Mamy:

$$P(p^* = \frac{k}{N}) = \left(\frac{m}{N}\right)^k \left(\frac{n}{N}\right)^{N-k} \binom{N}{k}.$$

Dla dużych N możemy przybliżać to prawdopodobieństwo rozkładem:

$$P(p^* = p) = \left(\frac{m}{N}\right)_p^N \left(\frac{n}{N}\right)^{N-N_p} (N+1) \binom{N}{N_p}.$$

gdzie czynnik $(N+1)$ odpowiada za gęstość p na odcinku $(0,1)$ (zakładamy rozkład jednostajny).

Rozkład 1 może być dla dużych N przybliżany rozkładem normalnym:

$$P(p|m\text{x}O, n\text{x}R) \approx \frac{(N+1)}{\sqrt{2\pi Np(1-p)}} \exp\left(-\frac{(m-Np)^2}{2Np(1-p)}\right).$$

Podobnie rozkład otrzymany eksperymentalnie można przybliżyć rozkładem normalnym następująco:

$$P(p^* = p) \approx \frac{(N+1)}{\sqrt{2\pi mn/N}} \exp\left(-\frac{(m-Np)^2}{2mn/N}\right)$$

Można pokazać, że dwa powyższe rozkłady dla odpowiednio dużych N są sobie bardzo bliskie. Oznacza to, że metodą bootstrapingu potrafimy oszacować prawdopodobieństwo odwrotne ze wzoru Bayesa. Powyższe wyniki mogą być łatwo powtórzone dla rozkładu wielomianowego.

4.2 Bootstrapping dla drzewa o dwóch liściach

Rozważmy przykład drzewa o dwóch liściach x_1 i x_2 zawierających sekwencje nukleotydów, dla którego liczyliśmy największą wiarygodność na poprzednim wykładzie.

Ponieważ wiarygodność drzewa nie zależy od położenia korzenia możemy umieścić korzeń w jednym z liści i połączyć je krawędzią o długości t równej sumie długości krawędzi prowadzących od korzenia do liści.

Jak poprzednio zakładamy, że sekwencje są identyczne na n_1 miejscach i różnią się dla pozostałych n_2 miejsc ($n_1 + n_2 = N$). Policzyliśmy wiarygodność tego drzewa jako funkcję dlugości krawędzi t:

$$P(x,y|T,t) = \frac{1}{16^{n_1+n_2}}(1+3e^{-4pt})^{n_1}(1-e^{-4pt})^{n_2}.$$

Zakładając jednostajne prawdopodobieństwo *a priori*, możemy wyrazić prawdopodobieństwo odwrotne jako:

$$P(e^{-pt}|\text{dane}) = \frac{(1+3e^{-pt})^{n_1}(1-e^{-pt})^{n_2}}{Z},$$

gdzie Z jest czynnikiem normującym we wzorze Bayesa. Załóżmy teraz, że n_{XY} jest liczbą przypisań nukleotydów X i Y w dwóch liściach drzewa w oryginalnym zbiorze danych (uliniowieniu rozważanych sekwencji). Niech m_{XY} oznacza analogiczną wielkość w sztucznym zbiorze danych wykorzystywanych w metodzie bootstrappingu. Prawdopodobieństwo wylosowania ustalonego zbioru danych wynosi:

$$P(\boldsymbol{m}|\boldsymbol{n}) = \left(\frac{n_{AA}}{N}\right)^{m_{AA}} \left(\frac{n_{AC}}{N}\right)^{m_{AC}} \cdots \frac{N!}{m_{AA}!m_{AC}!\ldots}.$$

Optymalna (dająca największa wiarygodność) długość krawędzi wynosiła w naszym przykładzie:

$$t^* = t_1^* + t_2^* = -\frac{3}{4}\ln\frac{3n_1 - n_2}{3n_1 + 3n_2} \tag{2}$$

Możemy ten wynik wykorzystać licząc $t*$, dla którego funkcja wiarygodnośći danych bootstrappingu przyjmuje maksimum:

$$\exp(-pt^*) = \frac{3m_1 - m_2}{3N},$$

gdzie m_1 oznacza liczbę identycznych miejsc m_2 liczbę miejsc, na których sekwencje się różnią. Widzimy, że t^* nie zależy od poszczególnych m_{XY} jedynie od m_1 i m_2. Możemy policzyć stąd m_1 i m_2:

$$\begin{aligned} m_1 =& \frac{N}{4}(1 + 3e^{-pt^*}) \\ m_2 =& \frac{3N}{4}(1 - e^{-pt^*}) \end{aligned}$$

Sumując po możliwych zbiorach danych uzyskamy prawdopodobieństwo dla estymatora największej wiarygodności:

$$P\left(e^{-pt^*} = \frac{3m_1 - m_2}{3N}\right) = \left(\frac{n_1}{N}\right)^{m_1}\left(\frac{n_2}{N}\right)^{m_2}\binom{N}{m_2}.$$

Widzimy, że wyraża się ono rozkładem dwumianowym, podobnie jak w poprzednim przykładzie. Wnioskujemy stąd, że rozkład eksperymentalny uzyskany metodą bootstrappingu pozwoli nam na przybliżenie rozkładu *a posteriori*.

5 Zadania

1. Rozważ bardzo ubogą przestrzeń drzew filogenetycznych, która składa się tylko z dwóch drzew T i $\tilde{T}$ o prawdopodobieństwach $P(T)$ oraz $P(\tilde{T})$. Pokaż, że jeśli zawsze będziemy proponować to drugie drzewo jako następnika, algorytm Metropolisa wygeneruje ciąg drzew, w którym częstości wystąpień będą zbiegać do ich prawdopodobieństw.

2. Uzasadnij wzór (2) na najbardziej wiarygodną długość krawędzi dla dwulistnego drzewa.