Spis rzeczy

5 Uliniowienie wielu sekwencji **58**
5.4 Miara typu 'konsens' 58
5.5 Uliniowienie z drzewem fiologenetycznym 60

5 Uliniowienie wielu sekwencji

5.4 Miara typu 'konsens'

Niech s będzie odległością w alfabecie $\Sigma \cup \{-\}$ i niech $S_1, \ldots, S_k$ będzie skończonym ciągiem słów nad Σ. Dla słowa $\bar{S} \in \Sigma^*$, *błędem konsensu* słowa $\bar{S}$ względem $S_1, \ldots, S_k$ nazwiemy

$$E(\bar{S}) = \sum_{i=1}^{k} \overline{sim}(\bar{S}, S_i).$$

Słowo $\bar{S}$ nazwiemy *słowem Steinera* dla $S_1, \ldots, S_k$, jeśli minimalizuje ono błąd konsensu $E(\bar{S})$.

Twierdzenie 5.4.1 (de la Higuera, Casacuberta'2000)
Niech s będzie odległościa edycyjną (tzn. $s(x,y) = 0$, gdy $x = y$; $s(x,y) = 1$, gdy $x \neq y$). Następujący problem jest NP-zupełny. Dany alfabet Σ, ciąg słów $S_1, \ldots, S_k$ oraz liczba $m \in \mathbb{N}$. Czy istnieje $\bar{S} \in \Sigma^*$, takie że $E(\bar{S}) \leq m$, gdzie $E(\bar{S})$ jest liczone względem $S_1, \ldots, S_k$ w sensie odległości s?

Dowód Twierdzenia 5.4.1 jest poprzez redukcję z problemu najdłuższego wspólnego podciągu (NWP). Słabą stroną powyższego twierdzenia jest to, że dla każdej instancji problemu NWP dobiera się alfabet Σ, natomiast ciekawe jest to, że wynik dotyczy odległości edycyjnej. Interesującym problem jest poprawienie powyższego twierdzenia tak, żeby zachodziło ono dla jednego ustalonego alfabetu.

Lemat 5.4.2 *Niech s będzie odległością. Dla każdego ciągu słów $S_1, \ldots, S_k$ istnieje i takie, że*

$$\frac{E(S_i)}{E(\bar{S})} \leq 2 - 2/k,$$

gdzie $\bar{S}$ jest słowem Steinera dla $S_1, \ldots, S_k$.

Dowód: Niech $1 \leq i \leq k$ będzie dowolne. Wtedy

$$E(S_i) = \sum_{j=1}^{k} \overline{sim}(S_i, S_j) \leq \sum_{j \neq i} [\overline{sim}(S_i, \bar{S}) + \overline{sim}(\bar{S}, S_j)] = (k-2) \cdot \overline{sim}(S_i, \bar{S}) + E(\bar{S}).$$

Teraz niech i będzie takie, że $\overline{sim}(S_i, \bar{S})$ jest minimalne. Wówczas

$$E(\bar{S}) = \sum_{j=1}^{k} \overline{sim}(\bar{S}, S_j) \geq k \cdot \overline{sim}(\bar{S}, S_i).$$

Zatem

$$\frac{E(S_i)}{E(\bar{S})} \leq \frac{(k-2) \cdot \overline{sim}(S_i, \bar{S})}{E(\bar{S})} + 1 \leq \frac{(k-2) \cdot \overline{sim}(S_i, \bar{S})}{k \cdot \overline{sim}(S_i, \bar{S})} + 1 = \frac{k-2}{k} + 1 = 2 - 2/k.$$

To kończy dowód lematu. ∎

Poniższe twierdzenie jest bezpośrednim wnioskiem z Lematu 5.4.2 oraz z definicji centrum (por. Sekcja 5.3).

Twierdzenie 5.4.3 *Niech s będzie odległoscią. Dla każdego ciągu słów $S_1, \ldots, S_k$, jeśli S_i jest centrum dla $S_1, \ldots, S_k$ oraz $\bar{S}$ jest słowem Steinera, to*

$$\frac{E(S_i)}{E(\bar{S})} \leq 2 - 2/k.$$

Teraz opiszemy trochę inne podejście. Nadal zakładamy, że s jest odległością. Niech $S_1^{\#}, \ldots, S_k^{\#}$ będzie uliniowieniem dla $S_1, \ldots, S_k$. *Konsensowa litera* dla i-tej kolumny jest to taka litera x, która minimalizuje

$$d(i) = \sum_{j=1}^{k} s(x, S_j(i)).$$

Konsensowe słowo dla uliniowienia $S_1^{\#}, \ldots, S_k^{\#}$ jest to słowo składające się z konsensowych liter (na i-tej pozycji stoi konsensowa litaera dla i-tej kolumny uliniowienia). *Błąd uliniowienia* $S_1^{\#}, \ldots, S_k^{\#}$ jest to $\sum_{i=1}^{n} d(i)$, gdzie $n = |S_1^{\#}| = \ldots = S_k^{\#}$.

Uliniowienie jest *optymalne w sensie konsensowym*, gdy błąd uliniowienia jest minimalny spośród wszystkich możliwych błędów.

Twierdzenie 5.4.4

1. *Niech S będzie konsensowym słowem uliniowienia optymalnego dla $S_1, \ldots, S_k$ w sensie konsensowym. Wówczas po usunięciu spacji z S otrzymujemy słowo Steinera dla $S_1, \ldots, S_k$.*

2. *Na odwrót, jeśli $\bar{S}$ jest słowem Steinera dla $S_1, \ldots, S_k$ oraz $S_1^{\#}, \ldots, S_k^{\#}, \bar{S}^{\#}$ jest uliniowieniem zgodnym z gwiazdą o środku $\bar{S}$, to $S_1^{\#}, \ldots, S_k^{\#}$ jest optymalnym uliniowieniem w sensie konsensowym.*

Wniosek 5.4.5 *Jeśli $S_1^{\#}, \ldots, S_k^{\#}$ jest uliniowieniem zgodnym z gwiazdą G_c (tak jak w Sekcji 5.3), to błąd tego uliniowienia jest nie większy niż $(2-2/k)$ razy błąd optymalnego uliniowienia w sensie konsensowym.*

5.5 Uliniowienie z drzewem fiologenetycznym

Zaczniemy od następującej definicji. *Drzewem filogenetycznym* dla danego ciągu różnych słów $S_1, \ldots, S_k$ nazwiemy każde drzewo o k liściach, którego liście są etykietowne słowami $S_1, \ldots, S_k$ (każde S_i jest etykietą pewnego liścia).

Filogenetyczne uliniowienie dla drzewa filogenetycznego T jest to dowolne przyporządkowanie pewnych słów wierzchołkom wewnętrznym T. Słowa przyporządkowywane oczywiście mogą być inne niż etykiet liści. Ustalmy funkcję podobieństwa s i załóżmy, że jest ona odległością. Niech $T^{\#}$ będzie filogenetycznym uliniowieniem dla T. *Miarą* (bądź *wagą*) tego uliniowienia nazwiemy

$$\sum_{(v,v')\in E} \overline{sim}(S_v, S_{v'}),$$

gdzie $(v, v') \in E$ oznacza, że wierzchołki v i v' sąsiadują ze sobą (czyli łączy je krawędź) oraz S_v jest etykietą stojącą przy v (podobnie dla $S_{v'}$).

Problem filogenetycznego uliniowienia (Sankoff'75) polega na znajdywaniu filogenetycznego uliniowienia $T^{\#}$ drzewa T, tak że miara $T^{\#}$ jest minimalna. *Uwaga:* problem filogenetycznego uliniowienia z gwiazdą jest równoważny znajdowaniu słowa Steinera dla etykiet liści.

Twierdzenie 5.5.1 (Wang, Jiang'94)
Następujący problem jest NP-zupełny. Dane binarne drzewo filogenetyczne oraz liczba $m \in \mathbb{N}$, stwierdzić czy istnieje filogenetyczne uliniowienie o wadze nie większej niż m.

Używając standardowej metody dynamicznego programowania problem filogenetycznego uliniowienia daje się rozwiązać w czasie wykładniczym. Zajmiemy się algorytmem aproksymacyjnym pochodzącym od Wanga, Jianga i

Lawlera (1996). Uliniowienie T nazwiemy *podniesionym uliniowieniem*, gdy etykieta każdego wewnętrznego wierzchołka drzewa T jest etykietą jednego z jego synów.

Zdefiniujemy *proces podnoszenia uliniowienia*. Niech T będzie filogenetycznym uliniowieniem. Opiszemy podniesione uliniowienie T^L. Etykiety liści w T^L są takie same jak w T. Niech v będzie wierzchołkiem, dla którego jeszcze nie zdefiniowaliśmy etykiety w T^L i załóżmy, że wszyscy synowie $v_1, \ldots, v_m$ wierzchołka v mają określone etykiety $S_1, \ldots, S_k$ w T^L. Niech S będzie etykietą wierzchołka v w T. Wówczas jako etykietę wierzchołka v w T^L wybieramy takie S_i, dla którego zachodzi

$$\overline{sim}(S, S_i) \leq \overline{sim}(S, S_j),$$

dla wszystkich $j = 1, \ldots, m$.

Twierdzenie 5.5.2 (Wang, Jiang, Lawler'96)
Niech s będzie odległością. Jeśli $T^\#$ jest optymalnym filogenetycznym uliniowieniem dla T oraz $(T^\#)^L$ jest podniesieniem $T^\#$, to waga drzewa $(T^\#)^L$ jest co najwyżej dwa razy waga drzewa $T^\#$.

Dowód: Weźmy krawędź $e = (v, w)$ w T, taką że w $(T^\#)^L$ waga tej krawędzi jest dodatnia. Niech w będzie synem v (tzn. v leży pomiędzy korzeniem drzewa T a w) i niech S_j będzie etykietą stojącą w $(T^\#)^L$ przy v oraz niech S_i będzie etykietą stojącą w $(T^\#)^L$ przy w. Niech $\Pi_e = v, w, w_1, \ldots, w_m$ będzie drogą w $(T^\#)^L$ prowadzącą do liścia etykietowanego S_i. Wynika natychmiast z konstrukcji podniesienia, że wszystkie wierzchołki na tej drodze (począwszy od w) muszą mieć tę samą etykietę S_i. Mamy dwie oczywiste obserwacje:

- waga drogi Π_e jest równa $\overline{sim}(S_j, S_i)$.
- Dla różnych krawędzi e_1, e_2 o wadze dodatniej, drogi Π_{e_1} oraz Π_{e_2} są rozłączne.

Niech $S_v, S_w, \ldots, S_{w_1}, S_{w_{m-1}}$ będą etykietami w $T^\#$ stojącymi przy wierzchołkach $v, w, w_1, \ldots, w_{m-1}$. Mamy

$$\begin{aligned}\overline{sim}(S_j, S_i) \leq \overline{sim}(S_j, S_v) + \overline{sim}(S_v, S_i) \leq 2 \cdot \overline{sim}(S_v, S_i) \leq \\ 2 \cdot [\overline{sim}(S_v, S_w) + \overline{sim}(S_w, S_{w_1}) + \cdots + \overline{sim}(S_{w_{m-1}}, S_i)]\end{aligned}$$

Zatem waga drogi Π_e w $(T^\#)^L$ jest co najwyżej dwa razy waga tej samej drogi w $T^\#$. Z drugiej z powyższych uwag wynika, że waga całego drzewa $(T^\#)^L$ jest co najwyżej dwa razy waga drzewa $T^\#$. To kończy dowód twierdzenia.

■

Pozostaje do rozstrzygnięcia problem znajdowania minimalnego drzewa podniesionego. Zastosujemy metodę dynamicznego programowania. Niech T będzie filogenetycznym drzewem dla słów $S_1, \ldots, S_k$. Dla dowolnego wierzchołka v w T oraz dla $1 \leq i \leq k$, takiego że etykieta S_i jest widoczna z v, niech $d(v,i)$ będzie najmniejszą wagą podniesienia poddrzewa T_v drzewa T, przy dodatkowym założeniu, że w korzeniu T_v (czyli v) stoi etykieta S_i. Mając wszystkie wartości $d(v,i)$, wagę minimalnego podniesienia znajdujemy biorąc

$$\min_{1 \leq i \leq k} d(v_0, i),$$

gdzie v_0 jest korzeniem T.

Wartości brzegowe definiujemy następująco. Niech v będzie liściem i niech S_i będzie etykietą v. Wówczas

$$d(v,i) = 0.$$

Krok indukcyjny wygląda następująco. Niech $v_1, \ldots, v_m$ będą wszystkimi synami wierzchołka v, $(m > 0)$ i załóżmy, że S_i jest widoczne z v oraz z wierzchołka v_p będącego synem wierzchołka v. Dla $1 \leq j \leq m$ niech $W_j = \{1 \leq l \leq k \mid S_l \text{ jest widoczne z } v_j\}$. Wówczas

$$d(v,i) = d(v_p,i) + \sum_{j=1,\ j \neq p}^{m} \min_{l \in W_j} [\overline{sim}(S_i, S_l) + d(v_j, l)]. \tag{5.5.1}$$

Jak zwykle, używając wskaźników, możemy odtwarzać etykiety minimalnego podniesienia drzewa filogenetycznego T.

Oszacujmy czas potrzebny do całej tablicy d. Niech T ma k liści etykietowanych $S_1, \ldots, S_k$ i niech $M = \sum_{i=1}^{k} |S_i|$.

- Czas potrzebny do obliczenia wszystkich wartości $\overline{sim}(S_i, S_j)$, dla $i < j$ wynosi $O(\sum_{i+1}^{k-1}(|S_i| \cdot \sum_{j=i+1}^{k} |S_j|)) = O(M^2)$.

- Zauważmy, że w (5.5.1) $W_j \cap W_{j'} = \emptyset$, dla $j \neq j'$. Zatem $m \leq k$ i oznacza to, że jeśli mamy już obliczone wartości składowe w sumie (5.5.1), to obliczenie $d(v,i)$ zajmuje $O(k)$ kroków.

- Elementów w tablicy d jest $O(|T| \cdot k)$. Zatem czas potrzebny do wyliczenia całej tablicy d wynosi $O(M^2 + |T| \cdot k^2)$. Jeśli wszystkie wierzcholki wewnętrzne są rzędu większego od 1, to łatwo jest pokazać, że jest ich co najwyżej $k - 1$. Tak więc, w tym przypadku czas potrzebny na wyliczenie d wynosi $O(M^2 + k^3)$.