Karol SZOSTEK*

OPTYMALIZACJA MODELI HMM ORAZ ICH ZASTOSOWANIE W ROZPOZNAWANIU MOWY

STRESZCZENIE

Modelowanie sygnału mowy za pomocą niejawnych modeli Markowa HMM (hidden Markov model) stanowi jeden z najefektywniejszych sposobów rozpoznawania mowy. Niniejszy artykuł poświęcony jest podstawom matematycznym teorii niejawnych modeli Markowa. Szczególną uwagę zwrócono w nim na wyprowadzenie zależności pozwalających stosować modele HMM do modelowania sygnałów. W pierwszej części artykułu przedstawiono wyprowadzenie zależności pozwalające dobierać parametry modelu procesu Markowa. W dalszej części artykułu przedstawione są wyprowadzenia zależności pozwalające w sposób krokowy dobierać parametry modelu łańcucha Markowa. Opisane metody oparte są na minimalizacji prawdopodobieństwa wygenerowania losowej w czasie sekwencji obserwacji w funkcji parametrów modelu. W przedstawionych w artykule wyprowadzeniach na zależności pozwalające optymalizować modele HMM wykorzystano metodę mnożników Lagrange'a.

***Słowa kluczowe**: proces losowy, model Markowa, model HMM, optymalizacja*

OPTIMIZATION OF THE MODELS HMM AND THEIR APPLICATION IN SPEECH RECOGNITION

Modeling the speech signal with the use of hidden Markov models HMM constitutes one of the most effective ways of speech recognition. This article is devoted to mathematical bases of the theory of hidden Markov models. Special attention was paid in it to derivation of dependencies allowing applying the models HMM for modeling signals. In the first part of the article there was presented derivation of dependencies allowing selection of parameters of the model of Markov process. In the further part of the article there are presented derivations of dependencies allowing selecting parameters of the model of Markov chain in a stage way. The described methods are based on minimization of the probability of generating random-in-time sequence of observation in the function of parameters of the model. The method of Lagrange's multipliers was used in the derivations for dependencies, presented in the article, allowing optimizing the models HMM.

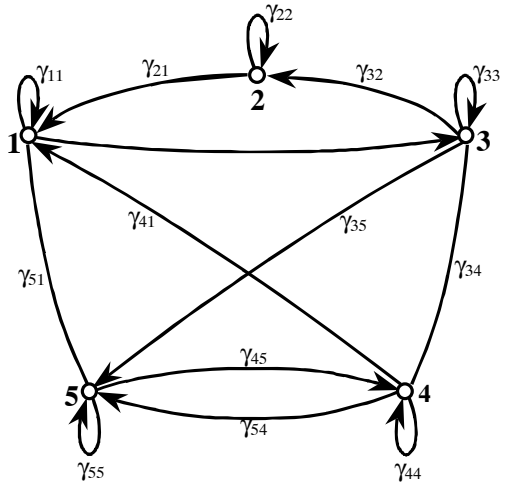
***Keywords**: random process, Markov model, model HMM, optimization*

1. WPROWADZENIE

Sygnały mające charakter niedeterministyczny pochodzące z procesów fizycznych mogą być modelowane za pomocą modeli Markowa. Modele te są często wykorzystywane do automatycznego rozpoznawania mowy. Niniejszy artykuł poświęcony jest podstawom matematycznym teorii niejawnych modeli Markowa. Szczególną uwagę zwrócono na wyprowadzenia zależności pozwalających stosować modele HMM do modelowania sygnałów. W pierwszej części artykułu przedstawiono wyprowadzenie zależności pozwalające dobierać parametry modelu procesu Markowa. Wyprowadzenia dla modelu procesu Markowa przedstawione są w celu pokazania koncepcji postępowania na prostym modelu. W dalszej części artykułu przedstawione są wyprowadzenia zależności pozwalające w sposób krokowy dobierać parametry modelu łańcucha Markowa (**algorytm Bauma–Welcha**). Celem artykułu jest pokazanie czytelnikowi koncepcji modelowania przebiegów losowych w czasie za pomocą niejawnych modeli Markowa, ze szczególnym zwróceniem uwagi na podstawy teoretyczne dotyczące tych modeli.

2. PROCES MARKOWA

Rozważmy model, który może być w jednym z N oddzielnych stanów ponumerowanych jako $S_t = \{1, 2, ..., N\}$ jak pokazano na rysunku 1 (dla $N = 5$). W regularnych odstępach czasu w modelu następują zmiany stanów. Zmiana stanów następuje zgodnie z ustalonymi prawdopodobieństwami przejścia do następnego stanu. Chwile związane ze zmianami stanów oznaczamy jako $t = 1, 2, ...,T$, a stany, w których przebywa model w kolejnych chwilach, oznaczamy przez S_t.



Rys. 1. Model procesu Markowa z pięcioma stanami

* Studia doktoranckie, Wydział EAIiE AGH

Pełny probabilistyczny opis zachowania się tego modelu w czasie t powinien uwzględniać te stany, w jakich model znajdował się we wszystkich poprzednich chwilach czasowych. W tym artykule zajmujemy się modelami HMM pierwszego rzędu, w których probabilistyczna zależność stanów jest skrócona do poprzedniego stanu i jest opisana zależnością

$$P(S_t = j \mid S_{t-1} = i, S_{t-2} = k, ...) = P(S_t = j \mid S_{t-1} = i) \quad (1)$$

Rozważamy tylko procesy stacjonarne takie, dla których prawa strona równania (1) jest niezależna od czasu. Dzięki temu możemy zapisać prawdopodobieństwo zmiany stanów w postaci macierzy prawdopodobieństw przejść Γ, gdzie $\Gamma = (\gamma_{ij})$ jest określone dla wszystkich stanów $i = 1, 2, ..., N$ oraz $j = 1, 2, ..., N$ i wszystkich czasów t

$$\gamma_{ij} = P(S_t = j \mid S_{t-1} = i) \quad (2)$$

Dla współczynników macierzy przejść $\Gamma = (\gamma_{ij})$ zachodzą ograniczenia opisane zależnościami:

$$\gamma_{ij} \geq 0 \quad \forall i, j \quad (3)$$

$$\sum_{j=1}^{N} \gamma_{ij} = 1 \quad \forall i \quad (4)$$

Należy zaznaczyć, że opisany model może pracować jako generator sekwencji losowej, dlatego dodatkowo wprowadzamy rozkład prawdopodobieństwa początkowego $\delta = (\delta_i)$, gdzie δ_i jest opisane zależnością (5), jest to prawdopodobieństwo tego, że model rozpocznie pracę w węźle i:

$$\delta_i = P(S_1 = i), \quad 1 \leq i \leq N \quad (5)$$

Podobnie jak dla macierzy przejść, dla prawdopodobieństw początkowych zachodzą ograniczenia opisane zależnościami:

$$\delta_i \geq 0 \quad \forall i \quad (6)$$

$$\sum_{i=1}^{N} \delta_i = 1 \quad (7)$$

Jeśli HMM spełnia powyższe własności, stanowi standardową probabilistyczną konstrukcję pozwalającą modelować procesy losowe w czasie.

W wyniku pracy N-stanowego modelu przez T chwil dostaniemy pewną obserwację $\mathbf{O} = (O_1, O_2, ..., O_T)$. Obserwowane wartości w kolejnych chwilach odpowiadają stanom, w jakich model przebywał.

Jednym z problemów dotyczących modelu procesu Markowa jest obliczenie prawdopodobieństwa P($\mathbf{O}$|M) wygenerowania przez N-stanowy model pewnej sekwencji stanów $\mathbf{O} = (O_1, O_2, ..., O_T)$. Korzystając z własności (1) prawdopodobieństwo wygenerowania sekwencji można zapisać jak w równaniu

$$\begin{aligned} P(O \mid \mathrm{M}) &= P(S_1 = O_1)P(S_2 = O_2)...P(S_T = O_T) = \\ &= P(O_1, O_2, ..., O_T \mid \mathrm{M}) = \delta_{s_1}\gamma_{s_1 s_2}\gamma_{s_2 s_3}...\gamma_{s_{T-1} s_T} = \\ &= \delta_{s_1}\prod_{i=1}^{N}\prod_{j=1}^{N}\gamma ij^{f_{ij}} \end{aligned} \quad (8)$$

gdzie f_{ij} – jest to liczba przejść ze stanu i do stanu j w rozpatrywanej sekwencji stanów $\mathbf{O}$.

Dla f_{ij} zachodzi $\sum_{ij} f_{ij} = T - 1$.

Kolejnym problemem dla omawianego modelu jest estymacja $N^2 + N$ parametrów modelu tak, aby maksymalizować zaobserwowanie sekwencji obserwacji $\mathbf{O} = (O_1, O_2, ..., O_T)$. Rozwiązanie tego zadanie jest konieczne w celu znalezienia modelu odpowiadającego zebranym obserwacjom.

Prawdopodobieństwo sekwencji wyliczymy według zależności (8). Naszym zadaniem jest znalezienie takich parametrów modelu, aby maksymalizować prawdopodobieństwo zaobserwowania sekwencji obserwacji $\mathbf{O} = (O_1, O_2, ..., O_T)$. Do wyznaczenia parametrów wykorzystana zostanie metoda mnożników Lagrange'a. Najpierw wyliczmy logarytm z zależności (8), stąd dostajemy zależność

$$\begin{aligned} \log P &= \log\left(\delta_{s_1}\prod_{i=1}^{N}\prod_{j=1}^{N}\gamma_{ij}^{f_{ij}}\right) = \\ &= \log\delta_{s_1} + \sum_{i=1}^{N}\sum_{j=1}^{N} f_{ij}\log(\gamma_{ij}) \end{aligned} \quad (9)$$

Zamiast maksymalizować prawdopodobieństwo sekwencji maksymalizujemy logarytm prawdopodobieństwa sekwencji jak w zależności (10). Pierwsza część maksymalizowanej sumy związana jest z prawdopodobieństwami początkowymi, natomiast druga z prawdopodobieństwami przejść. Obie części możemy maksymalizować oddzielnie.

$$\begin{aligned} &\max_{\delta_{S_1},\gamma_{ij}} \delta_{s_1}\prod_{i=1}^{N}\prod_{i=1}^{N}\gamma_{ij}^{f_{ij}} \Leftrightarrow \\ &\Leftrightarrow \max_{\delta_{S_1},\gamma_{ij}}\left(\log\delta_{s_1} + \sum_{i=1}^{N}\sum_{j=1}^{N} f_{ij}\log(\gamma_{ij})\right) \end{aligned} \quad (10)$$

W wyniku maksymalizacji pierwszej części dostajemy optymalne wartości prawdopodobieństw początkowych

$$\hat{\delta}_{S_1} = 1 \text{ oraz } \hat{\delta}_i = 0 \text{ dla } i \neq S_1 \quad (11)$$

Do znalezienia optymalnych wartości prawdopodobieństw przejść została zastosowana metoda mnożników Lagrange'a.

Funkcja Lagrange'a dla drugiej części podana jest równaniem

$$L(\sum_{i=1}^{N}\sum_{j=1}^{N} f_{ij}\log(\gamma_{ij})) = \\ = \sum_{i=1}^{N}\sum_{j=1}^{N} f_{ij}\log(\gamma_{ij}) - \sum_{i=1}^{N}\lambda_i\left(\sum_{j=1}^{N}\gamma_{ij} - 1\right) \tag{12}$$

Stąd dostajemy $N{\cdot}N$ równań dla parametrów przejść oraz N równań dla ograniczeń jak w równaniach:

$$\begin{cases} \forall_{i,j} & \dfrac{\partial L(\log P)}{\partial \gamma_{ij}} = f_{ij}\dfrac{1}{\gamma_{ij}} - \lambda_i = 0 \\ \forall_{i} & \dfrac{\partial L(\log P)}{\partial \lambda_i} = \left(\sum_{j=1}^{N}\gamma_{ij} - 1\right) = 0 \end{cases} \tag{13}$$

Mamy do znalezienia $N{\cdot}N + N$ niewiadomych oraz tyle samo równań, dlatego możemy wyznaczyć optymalne wartości prawdopodobieństw przejść dzięki przekształceniom:

$$\forall_{i,j} \quad f_{ij}\frac{1}{\gamma_{ij}} - \lambda_i = 0$$

$$\forall_{i,j} \quad \gamma_{ij} = \frac{f_{ij}}{\lambda_i}, \tag{14}$$

stąd:

$$\sum_{j=1}^{N}\gamma_{ij} = 1 = \sum_{j=1}^{N}\frac{f_{ij}}{\lambda_i} = \frac{1}{\lambda_i}\sum_{j=1}^{N} f_{ij} \quad \Rightarrow \quad \lambda_i = \sum_{j=1}^{N} f_{ij}$$

W rezultacie otrzymujemy estymator modelu, opisany zależnością

$$\hat{\gamma}_{ij} = \frac{f_{ij}}{\sum_{k=1}^{N} f_{ik}} \tag{15}$$

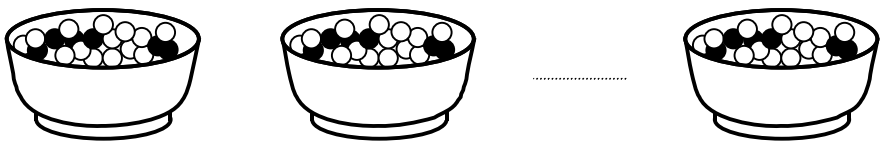
Mając daną sekwencje obserwacji **O**, możemy zastosować zależności (11) i (15) do dobrania parametrów modelu, w wyniku uzyskujemy model M, dla którego P(**O**|M) przyjmuje wartość maksymalną.

3. NIEJAWNY MODEL MARKOWA

W poprzednio rozważanym modelu, każdy stan odpowiadał obserwowanemu zdarzeniu. Wyjście (obserwacja) w takim modelu w danym stanie nie jest wartością losową, dlatego modele takie są zbyt proste, aby zastosować je do wielu praktycznych procesów losowych. W tym rozdziale zostanie przedstawiony niejawny model Markowa, w którym obserwacje są probabilistyczną funkcją stanów, dzięki temu można go zastosować do modelowania bardziej złożonych procesów. Poniżej przedstawiony jest przykład procesu, który jest ukryty (nie jest widoczny). Może być jedynie obserwowany poprzez inny proces stochastyczny, który produkuje sekwencje obserwacji.

3.1 System z urnami i kulami

Na rysunku 2 pokazany jest system, w którym jest N urn. W każdej urnie jest duża liczba kolorowych kul o $\boldsymbol{M}$ różnych kolorach. Proces, dzięki któremu uzyskujemy obserwacje, jest następujący. W sposób losowy wybierana jest pierwsza urna, z wybranej urny wybierana jest w sposób losowy kula, a jej kolor jest obserwacją. Po wylosowaniu kule są zwracane do urny, z której zostały wylosowane, następnie wybierana jest nowa urna, według losowej procedury zależnej od aktualnej urny. Proces losowania kuli jest powtarzany do momentu uzyskania interesującej nas liczby obserwacji. Postępując w powyżej opisany sposób, uzysku-



Urna 1		Urna 2		Urna m	
P(biały)	$=\pi_{11}$	P(biały)	$=\pi_{12}$	P(biały)	$=\pi_{1N}$
P(niebieski)	$=\pi_{21}$	P(niebieski)	$=\pi_{22}$	P(niebieski)	$=\pi_{2N}$
P(zielony)	$=\pi_{31}$	P(zielony)	$=\pi_{32}$	P(zielony)	$=\pi_{3N}$
P(żółty)	$=\pi_{41}$	P(żółty)	$=\pi_{42}$	P(żółty)	$=\pi_{4N}$
:		:		:	
:		:		:	
P(czarny)	$=\pi_{M1}$	P(czarny)	$=\pi_{M2}$	P(czarny)	$=\pi_{MN}$

O={zielony, zielony, niebieski, czerwony, ... , niebieski}

Rys. 2. System z urnami i kulami przedstawiający koncepcje podstawowego modelu HMM

jemy skończone sekwencje kolorów (obserwacji), które chcemy zamodelować za pomocą modelu HMM.

Wydaje się, że najprostszym HMM odpowiadającym systemowi z urnami i kulami jest taki model, w którym stan modelu odpowiada każdej urnie, a rozkład prawdopodobieństwa kolorów jest zdefiniowany dla każdego stanu. Wybór urn jest podyktowany przez macierz przejść modelu HMM.

Należy zaznaczyć, że kolory kul w urnach mogą być te same, urny różnią się jedynie liczbami kul w poszczególnych kolorach. Stąd obserwacja jakiegoś koloru (kuli) nie określa, z której urny ten kolor (kula) pochodzi.

3.2. Elementy HMM

Powyższy przykład pokazał, czym jest niejawny model HMM. Teraz formalnie zdefiniujemy elementy HMM. Model HMM dla obserwacji przyjmujących dyskretne wartości w czasie tak jak dla powyższego modelu procesu losowania kul jest scharakteryzowany przez parametry:

1) Liczbę stanów w modelu, którą oznaczamy przez N; numerujemy stany $\{1, 2, ..., N\}$ oraz oznaczamy stan w czasie t jako S_t, mimo że stany są ukryte, często odpowiadają im pewne fizyczne znaczenia. W problemie z urnami i kulami każdemu ze stanów odpowiadała inna urna.
2) Liczbę różnych obserwacji w stanie, którą oznaczamy przez M. Występujące obserwacje numerujemy liczbami $\{1, 2, ..., M\}$ oraz oznaczamy obserwacje w czasie t jako O_t. Dla przykładu z kulami i urnami obserwacjami były kolory kul wyciąganych z urn.
3) Macierz rozkładu prawdopodobieństw przejść (16), którą oznaczamy przez $\Gamma = \{\gamma_{ij}\}$:

   $\gamma_{ij} = P(S_{t+1} = j | S_t = i)$, $1 \le i, j \le N$.

$$\Gamma = \begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1N} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2N} \\ \cdots & \cdots & & \cdots \\ \gamma_{N1} & \gamma_{N2} & \cdots & \gamma_{NN} \end{bmatrix} \tag{16}$$

   Wartość γ_{ij} jest prawdopodobieństwem tego, że w kroku t+1 model znajdzie się w stanie j, przy założeniu, że w kroku t model znajdował się w stanie i.
   Aby każdy stan był osiągalny, z każdego innego stanu w jednym kroku przyjmujemy $\gamma_{ij} > 0$ dla wszystkich i, j. Istnieją modele HMM, np.: model Bakisa, dla których przyjmuje się $\gamma_{ij} = 0$ dla jednej lub kilku par (i, j).
4) Rozkład prawdopodobieństwa obserwacji (17) dla poszczególnych stanów opisany jest przez macierz $\pi = \{\pi_{ki}\}$:

   $\pi_{ki} = P(O_t = k | S_t = i)$, $1 \le k \le M$, $1 \le i \le N$.

$$\pi = \begin{bmatrix} \pi_{11} & \pi_{12} & \cdots & \pi_{1M} \\ \pi_{12} & \pi_{22} & \cdots & \pi_{2M} \\ \cdots & \cdots & & \cdots \\ \pi_{N1} & \pi_{N2} & \cdots & \pi_{NM} \end{bmatrix} \tag{17}$$

   Wartość π_{ki} jest prawdopodobieństwem tego, że model wygeneruje obserwacje k, przy założeniu, że model znajduje się w stanie i.
5) Rozkład początkowy prawdopodobieństwa opisany jest przez wektor $\delta = \{\delta_i\}$, $\delta_i = P(S_1 = i)$, $1 \le i \le N$. Składowa wektora δ_i, $1 \le i \le N$, jest to prawdopodobieństwo rozpoczęcia przez model pracy w stanie i.

W punktach 1–5 w pełni zdefiniowany został model HMM i jego parametry, Tak zdefiniowany model oznaczamy trójką (Γ, π, δ) oraz skrótowo przez M.

Ponadto należy zdefiniować prawdopodobieństwo wygenerowania obserwacji **O** przez model. Prawdopodobieństwo to oznaczamy przez $P(\mathbf{O}|M)$. Sposób wyliczania prawdopodobieństwa wygenerowania obserwacji zostanie omówiony w następnym podrozdziale.

Jeśli mamy dane wartości N, M, $M = (\Gamma, \pi, \delta)$ modelu HMM, może on być użyty jako generator obserwacji. W wyniku pracy modelu HMM przez T chwil czasowych, zostaje wygenerowany niejawny łańcuch Markowa w postaci ciągu obserwacji $\mathbf{O} = (O_1, O_2, ..., O_T)$.

3.3. Trzy podstawowe problemy dotyczące HMM (Wyniki analizy niejawnych łańcuchów Markowa)

Dany jest model HMM jak w poprzednim rozdziale. Trzy zadania muszą zostać rozwiązane, aby model ten mógł być zastosowany w aplikacjach. Zadania są następujące:

Zadanie 1

Dana jest sekwencja obserwacji $\mathbf{O} = (O_1, O_2, ..., O_T)$ oraz model $M = (\Gamma, \pi, \delta)$. Znaleźć prawdopodobieństwo zaobserwowania tej sekwencji $P(\mathbf{O}|M)$ dla danego modelu M.

Zadanie 2

Dana jest sekwencja obserwacji $\mathbf{O} = (O_1, O_2, ..., O_T)$ oraz model $M = (\Gamma, \pi, \delta)$. Znaleźć optymalną sekwencję stanów $S = (S_1, S_2, ..., S_T)$.

Zadanie 3

Dana jest sekwencja obserwacji $\mathbf{O} = (O_1, O_2, ..., O_T)$. Dobrać parametry modelu $M = (\Gamma, \pi, \delta)$ tak, aby maksymalizować prawdopodobieństwo $P(\mathbf{O}|M)$ zaobserwowania danej sekwencji **O**.

Zadanie 1 jest zagadnieniem oceny. Mianowicie mamy dany model i sekwencje obserwacji, należy ocenić, jakie jest prawdopodobieństwo wygenerowania takiej sekwencji przez model M. Możemy również ocenić, w jakim stopniu sekwencja pasuje do danego modelu. Jeżeli rozważymy przypadek wyboru jednego modelu spomiędzy kilku rozpatrywanych, wówczas wybierzemy model, który najlepiej pasuje do tej obserwacji, daje największe prawdopodobieństwo $P(\mathbf{O}|M)$.

Zadanie 2 jest zadaniem, w którym staramy się odkryć ukryte sekwencje stanów. W celu znalezienia najlepszego rozwiązania tego problemu zostaną wykorzystane metody optymalizacji.

Istnieją dwa możliwe podejścia:

1) wybieramy stany, które są indywidualnie najbardziej prawdopodobne;
2) wybieramy stany, które tworzą najbardziej prawdopodobną sekwencje stanów.

Zadanie 3 jest zadaniem, w którym optymalizujemy parametry modelu tak, aby model z jak największym prawdopodobieństwem generował daną sekwencję obserwacji. Sekwencja obserwacji służąca do dobrania parametrów modelu jest nazywana sekwencją treningową, ponieważ służy do dobierania parametrów modelu HMM. Problem ten jest decydujący w większości aplikacji modeli HMM, ponieważ pozwala na adaptację parametrów modelu do obserwowanej sekwencji, w wyniku czego otrzymujemy model rzeczywistego zjawiska.

W celu podsumowania, rozważmy prosty system rozpoznający odizolowane słowa, którego schemat blokowy zamieszczony jest w podrozdziale (3.8). Dla wszystkich słów W ze słownika chcemy zaprojektować oddzielny m-stanowy model HMM. Sygnał mowy reprezentowany jest przez czasową sekwencję próbkowanych wartości. Dodatkowo wartości są z pewnego zakresu od 1 do n. Mamy w ten sposób podaną sekwencję treningową dla każdego słowa ze słownika. Najpierw musimy stworzyć model dla poszczególnych słów. To zadanie jest realizowane za pomocą zadania 3, w celu zoptymalizowania parametrów modelu dla każdego słowa ze słownika. Zadanie 2 stosujemy w celu poprawy modelu (zwiększenie liczby stanów albo zmiana rozdzielczości n), tj: zwiększenia zdolności modelowania sekwencji obserwacji przez model. Gdy mamy dobrane modele HMM, rozpoznanie nieznanego słowa uzyskujemy, stosując zadanie 1. Wyliczamy prawdopodobieństwo dla każdego modelu, a następnie wybieramy model dający największe prawdopodobieństwo.

W następnych podrozdziałach zostaną przedstawione formalne rozwiązania każdego z fundamentalnych problemów związanych z HMM.

3.4. Rozwiązanie problemu 1 (Obliczenie prawdopodobieństwa obserwacji)

Chcemy obliczyć prawdopodobieństwo P($\mathbf{O}$|M) zaobserwowania sekwencji $\mathbf{O} = (O_1, O_2, ..., O_T)$ przez model M = (Γ, π, δ). Najprostsza metoda polega na wyliczeniu wszystkich możliwych sekwencji stanów (ścieżek) o długości T (długość obserwacji), jest ich N^T, następnie sumujemy prawdopodobieństwa wygenerowania obserwacji na znanej ścieżce po wszystkich możliwych ścieżkach. Jedna sekwencja stanów (ścieżka) pokazana jest w zależności

$$S = (S_1, S_2, ..., S_T) \tag{18}$$

gdzie S_1 jest stanem początkowym.

Prawdopodobieństwo zaobserwowania sekwencji obserwacji $\mathbf{O}$ pod warunkiem sekwencji stanów S można zapisać jak w zależności (19), ponieważ obserwacje są zależne tylko od stanów

$$P(\mathbf{O} \mid S) = \prod_{t=1}^{T} P(O_t \mid S_t) = \pi_{O_1 S_1} \pi_{O_2 S_2} ... \pi_{O_T S_T} \tag{19}$$

A prawdopodobieństwo sekwencji stanów podobnie jak dla procesu Markowa opisane jest zależnością

$$P(S) = \delta_{S_1} \gamma_{S_1 S_2} \gamma_{S_2 S_3} ... \gamma_{S_{T-1} S_T} \tag{20}$$

Korzystając z (19) oraz (20), prawdopodobieństwo dołączone można zapisać w postaci równania

$$P(\mathbf{O}, S) = P(\mathbf{O} \mid S) P(S) \tag{21}$$

Korzystając z prawdopodobieństwa warunkowego, poszukiwane prawdopodobieństwo możemy zapisać

$$\begin{aligned} P(\mathbf{O} \mid \mathrm{M}) &= \sum_{\text{wszystkich } S} P(\mathbf{O} \mid S) P(S) = \\ &= \sum_{S_1, S_2, ..., S_T} (\pi_{O_1 S_1} \pi_{O_2 S_2} ... \pi_{O_T S_T}) x (\delta_{S_1} \gamma_{S_1 S_2} \gamma_{S_2 S_3} ... \gamma_{S_{T-1} S_T}) = \\ &= \sum_{S_1=1}^{N} \sum_{S_2=1}^{N} ... \sum_{S_T=1}^{N} (\pi_{O_1 S_1} \pi_{O_2 S_2} ... \pi_{O_T S_T}) x (\delta_{S_1} \gamma_{S_1 S_2} \gamma_{S_2 S_3} ... \gamma_{S_{T-1} S_T}) \end{aligned} \tag{22}$$

Obliczenie powyższego wyrażenia wymaga $2T \cdot N^T$ działań. Mamy N^T sekwencji stanów, a dla każdej z nich trzeba wykonać $2T$–1 działań mnożenia oraz zsumować wyniki dla każdej sekwencji stanów. Dokładnie musimy wykonać $(2T–1)N^T$ mnożeń oraz N^T–1 dodawań. To zadanie jest niewykonalne nawet dla małych wartości np: $N = 5$ $T = 100$ jest $2 \cdot 100 \cdot 5^{100} \approx 10^{72}$ obliczeń. Ale istnieje lepsza procedura rozwiązująca to zagadnienie (nazwana procedurą *forward-backward*).

Definiujemy parametry $\alpha_t(i)$, $\beta_t(i)$ jak w równaniach (23) i (24) dla wszystkich stanów i oraz wszystkich t od 1 do T:

$$\alpha_t(i) := P(O_1, ..., O_t, S_t = i) \tag{23}$$

$$\beta_t(i) := P(O_{t+1}, O_{t+2}, ..., O_T / S_t = i) \tag{24}$$

Korzystając z własności (62) z podrozdziału 3.9, iloczyn $\alpha_t(i)$ $\beta_t(i)$ można zapisać równaniem

$$\begin{aligned} &\alpha_t(i)\beta_t(i) = \\ &= P(S_t = i) P(O_1, O_2, ..., O_t \mid S_t = i) P(O_{t+1}, O_{t+2}, ..., O_T \mid S_t = i) = \\ &= P(S_t = i) P(O_1, O_2, ..., O_T \mid S_t = i) \\ &= P(O_1, O_2, ..., O_T \wedge S_t = i) \end{aligned} \tag{25}$$

Następnie korzystając z prawdopodobieństwa całkowitego, poszukiwane prawdopodobieństwo można zapisać równaniem

$$\sum_{i=1}^{N} \alpha_t(i) \beta_t(i) = P(O_1, O_2, ..., O_T) = P(\mathbf{O} \mid M) \tag{26}$$

Poszukiwane prawdopodobieństwo można obliczyć dla każdego t, stąd mamy T różnych metod obliczenia tego prawdopodobieństwa. Na przykład dla $t = T$ dostajemy zależność

$$P(\mathbf{O} \mid M) = \sum_{i=1}^{N} \alpha_T(i) \tag{27}$$

W celu znalezienia $\alpha_t(i)$ oraz $\beta_t(i)$ przyjmujemy, że: $\beta_T(i) = 1$, $\alpha_1(i)$ ma postać

$$\alpha_1(i) = P(S_1 = i)P(O_1 \mid S_1 = i) = \delta_i \pi_{O_1 i} \tag{28}$$

Wartości $\beta_T(i)$, $\alpha_1(i)$ stanowią dane startowe przy wyliczaniu rekurencyjnych wzorów na $\alpha_t(i)$, $\beta_t(i)$. Wykorzystując własność (63) oraz własności prawdopodobieństwa warunkowego dostajemy rekurencyjny wzór na $\alpha_t(i)$

$$\begin{aligned}
&\alpha_{t+1}(j) = \\
&= \sum_{i=1}^{N} P(O_1,...,O_{t+1} \wedge S_t = i, S_{t+1} = j) = \\
&= \sum P(S_t = i, S_{t+1} = j) P(O_1,...,O_{t+1} \mid S_t = i, S_{t+1} = j) = \\
&= \sum P(S_t = i)\gamma_{ij} P(O_1,...,O_t \mid S_t = i) P(O_{t+1} \mid S_{t+1} = j) = \\
&= \sum P(O_1,...,O_t \wedge S_t = i)\gamma_{ij} \pi_{O_{t+1} j} = \\
&= \left(\sum_{i=1}^{N} \alpha_t(i)\gamma_{ij} \right) \pi_{O_{t+1} j}
\end{aligned} \tag{29}$$

Podobnie po zastosowaniu własności (64) oraz (65), gdzie ($l = t+1$), dostajemy rekurencyjny wzór na $\beta_T(i)$

$$\begin{aligned}
&\beta_t(i) = \\
&= \sum_{j=1}^{N} P(O_{t+1},...O_T \wedge S_t = i, S_{t+1} = j) / P(S_t = i) = \\
&= \sum P(O_{t+1},...,O_T \mid S_t = i, S_{t+1} = j) P(S_t = i, S_{t+1} = j) / P(S_t = i) = \\
&= \sum P(O_{t+1},...,O_T \mid S_{t+1} = j)\gamma_{ij} = \\
&= \sum P(O_{t+1} \mid S_{t+1} = j) P(O_{t+2},..,O_T \mid S_{t+1} = j)\gamma_{ij} = \\
&= \sum_{j=1}^{N} \pi_{O_{t+1} j} \beta_{t+1}(j)\gamma_{ij}
\end{aligned} \tag{30}$$

Wyrażenia na $\alpha_t(i)$, $\beta_t(i)$ mogą zostać zapisane w notacji macierzowej. Definiujemy wektory α_t, β_t dla wszystkich t do 1 do T jak w równaniach

$$\begin{aligned}
\alpha_t &= (\alpha_t(1), \alpha_t(2), ..., \alpha_t(m)) \\
\beta_t &= (\beta_t(1), \beta_t(2), ..., \beta_t(m))
\end{aligned} \tag{31}$$

Wówczas P($\mathbf{O}$|M) może zostać zapisane

$$P(\mathbf{O} \mid \mathrm{M}) = \alpha_t \beta_t' \quad \forall\, t \tag{32}$$

Rekurencyjne wzory na $\alpha_t(i)$, $\beta_t(i)$ mają postać

$$\begin{aligned}
\alpha_{t+1} &= \alpha_t B_{t+1} \\
\beta_t' &= B_{t+1} \beta_{t+1}'
\end{aligned} \tag{33}$$

Jeśli przyjmiemy, że $B_t = \Gamma\lambda(O_t)$, gdzie $\lambda(O)$ jest diagonalną macierzą $N \times N$ o elementach na przekątnej równych π_{Oi}, wzory rekurencyjne startują od $\alpha_1 = \delta\lambda(O_1)$, a $\beta_T = \mathbf{1}$. Wyrażenie na (*forward*) oraz (*backward*) algorytm przyjmuje postać (34), równania te są prawdziwe dla każdego t od 1 do T włącznie:

$$\begin{aligned}
\alpha_t &= \delta\lambda(O_1) B_2 B_3 ... B_t \\
\beta_t' &= B_{t+1} B_{t+2} ... B_T \mathbf{1}'
\end{aligned} \tag{34}$$

Teraz zostanie pokazane, jak wyprowadzić postać macierzową na P($\mathbf{O}$|M). Prawdopodobieństwo to można zapisać jak w równaniu (22). Wyrażenie to jest niewygodne do obliczeń, dlatego wykorzystując zależności macierzowe na α_t, β_t', dostajemy równanie

$$P(\mathbf{O} \mid \mathrm{M}) = \delta\lambda(O_1)\Gamma\lambda(O_2)\Gamma ... \Gamma\lambda(O_T)\mathbf{1}' \tag{35}$$

Korzystając z zależności (35), szukane prawdopodobieństwo możemy zapisać

$$P(\mathbf{O} \mid \mathrm{M}) = a \left(\prod_{t=2}^{T} B_t \right) \mathbf{1}' \tag{36}$$

gdzie: $a = \delta\lambda(O_1)$, ($a_j = \delta_j \pi_{O_1 j}$), natomiast macierz B jest zdefiniowana

$$B_t = \Gamma\lambda(O_t) = (\gamma_{ij} \pi_{O_t j}) \tag{37}$$

Jeśli przyjmiemy, że $\delta\Gamma = \delta$ (tj: gdy δ jest prawdopodobieństwem w stanie stacjonarnym), wówczas prawdopodobieństwo P($\mathbf{O}$|M) może być zapisane

$$P(\mathbf{O} \mid \mathrm{M}) = \delta \left(\prod_{t=1}^{T} B_t \right) \mathbf{1}' \tag{38}$$

Algorytm z równania (38) jest liniowy ze względu na T oraz kwadratowy ze względu na N, dlatego jest to w zasadzie ta sama procedura jak $P(\mathbf{O} \mid \mathrm{M}) = \sum_i \alpha_t(i)\beta_t(i)$, w przypadku gdy $t = T$ (algorytm *forward-backward*).

3.5. Rozwiązanie problemu 2 (Optymalna sekwencja stanów)

W odróżnieniu od zadania 1 w tym przypadku mamy wiele możliwych rozwiązań zadania 2. Jeżeli chcemy znaleźć optymalną sekwencję stanów, problem tkwi w określeniu

kryterium optymalności. Jednym ze sposobów jest wybranie stanów S_t, które są indywidualnie najbardziej prawdopodobne w kolejnych chwilach czasowych, w wyniku tego dostaniemy oczekiwaną liczbę indywidualnie najbardziej prawdopodobnych stanów. W celu przeprowadzenia optymalizacji należy wprowadzić parametr $\chi_t(i)$ jak w równaniu (39), jest to prawdopodobieństwo tego, że model był w stanie i, w czasie t i wygenerował obserwację **O**

$$\begin{aligned}\chi_t(i) &= P(S_t = i \mid \mathbf{O}) = \\ &= \frac{P(\mathbf{O} \wedge S_t = i)}{P(\mathbf{O})} = \frac{P(\mathbf{O} \wedge S_t = i)}{\sum_{i=1}^{m} P(\mathbf{O} \wedge S_t = i)} = \\ &= \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)}\end{aligned} \tag{39}$$

gdzie: $\alpha_t(i)$ dotyczy obserwacji O_1, O_2, ..., O_t, natomiast $\beta_t(i)$ obserwacji O_{t+1}, ..., O_T. Korzystając z $\chi_t(i)$ możemy znaleźć najbardziej prawdopodobny stan S_t^* w chwili t, zgodnie z równaniem

$$O_t^* = \arg\max_{1 \le i \le N} [\chi_t(i)] \tag{40}$$

Stosując zależność (40), pomimo że dostajemy pewną sekwencję stanów, może się zdarzyć, że taka sekwencja stanów jest niemożliwa, ponieważ w modelu HMM pewne prawdopodobieństwa przejścia między stanami mogą być równe zero, a takie przejścia mogą występować w znalezionej sekwencji stanów. Jedynym ze sposobów rozwiązania tego problemu jest zmodyfikowanie kryterium optymalności tak, aby maksymalizować prawdopodobieństwo obserwacji dla pewnej sekwencji stanów. Optymalizację taką można przeprowadzić dla dwóch stanów (S_t, S_{t+1}), trzech stanów (S_t, S_{t+1},S_{t+2}) itd. Pomimo że takie kryteria mogą zostać zastosowane w aplikacjach, najpowszechniej stosowane jest kryterium poszukiwania sekwencji stanów, dla której $P(\mathrm{S} \wedge \mathbf{O})$ przyjmuje wartość największą. Metoda znajdowania takiej sekwencji bazuje na metodzie programowania dynamicznego i jest nazywana algorytmem Viterbiego.

Algorytm Viterbiego

Definiujemy wyrażenie na parametr $\eta_t(i)$ jak w równaniu

$$\eta_t(i) = \max_{S_1, S_2, \ldots, S_T} P(S_1, S_2, \ldots, S_t = i \wedge O_1, O_2, \ldots, O_t) \tag{41}$$

$\eta_t(i)$ jest największym prawdopodobieństwem wzdłuż pojedynczej ścieżki o długości t, która przechodzi przez t stanów i kończy się w stanie i. Wzór rekurencyjny na $\eta_t(i)$ jest podany zależnością

$$\eta_{t+1}(j) := [\max_i \eta_t(i)\gamma_{ij}]\pi_{O_{t+1}j} \tag{42}$$

Aby uzyskać optymalną ścieżkę, musimy zapamiętywać argument maksymalny w kolejnych krokach algorytmu. Zrobimy to, wprowadzając wektor $\varphi_1(j)$ jak w równaniach (43). Aby znaleźć najlepszą sekwencje stanów musimy postępować według zależności

$$\begin{cases} \eta_1(i) = \delta_1 \cdot \pi_{O_1 i}, & \varphi_1(i) = 0 \qquad 1 \le i \le N \\ \eta_t(j) = \left[\max\limits_{1 \le i \le N} [\eta_{t-1}(i) \cdot \gamma_{ij}]\right] \cdot \pi_{O_t j}, & \varphi_t(j) = \arg\max\limits_{1 \le i \le N} [\eta_{t-1}(i) \cdot \gamma_{ij}] \\ & 1 \le j \le N, \quad 2 \le t \le T \\ \mathrm{P}^* = \max\limits_{1 \le i \le N} [\eta_T(i)] \\ \iota_T^* = \arg\max\limits_{1 \le i \le N} [\eta_T(i)], & \iota_t^* = \varphi_{t+1}(\iota_{t+1}^*) \\ & t = T-1, T-2, \ldots, 1 \end{cases} \tag{43}$$

Główna różnica pomiędzy algorytmem pozwalającym wyznaczać współczynniki $\alpha_t(i)$, oraz współczynniki $\eta_t(j)$ polega na tym, że w algorytmie (43) w kolejnych krokach wybierane są wartości największe, natomiast w algorytmie wyznaczania $\alpha_t(i)$ liczona jest suma tych wartości. P^* oznacza prawdopodobieństwo na optymalnej ścieżce $i^*=(i_1^*, i_2^*, \ldots, i_T^*)$.

Ścieżka nie musi być zapamiętywana w postaci $\varphi_1(j)$, można ją również wyznaczyć we wstecznym poszukiwaniu według wzoru rekurencyjnego:

$$\begin{aligned} \iota_T^* &= \arg\max_{1 \le i \le N} \eta_T(i) \\ \iota_t^* &= \arg\max_{1 \le i \le N} (\eta_i \gamma_{i\iota_{t+1}^*}) \qquad \text{dla } t = T-1, T-2, \ldots, 1 \end{aligned} \tag{44}$$

Ponieważ przedstawiony algorytm wymaga wykonywania dużej liczby mnożeń, niekiedy stosuje się alternatywny algorytm o mniejszej złożoności obliczeniowej. Zmniejszenie złożoności obliczeniowej uzyskujemy dzięki zastosowaniu logarytmów współczynników modelu HMM.

Alternatywny algorytm Viterbiego

Mamy dane logarytmy parametrów modelu:

$$\begin{aligned} \bar{\delta}_i &= \log(\delta_i) && 1 \le i \le N \\ \bar{\pi}_{ki} &= \log(\pi_{ki}) && 1 \le i \le N, 1 \le k \le M \\ \bar{\gamma}_{ij} &= \log(\gamma_{ij}) && 1 \le i, j \le N \end{aligned} \tag{45}$$

Po zastosowania logarytmów współczynników modelu optymalną ścieżkę można uzyskać stosując zależności:

$$\begin{cases} \bar{\eta}_1(i) = \log(\eta_1(i)) = \bar{\delta}_1 + \bar{\pi}_{O_1 i}, & \varphi_1(i) = 0 \qquad 1 \le i \le N \\ \bar{\eta}_t(j) = \left[\max\limits_{1 \le i \le N} [\bar{\eta}_{t-1}(i) + \bar{\gamma}_{ij}]\right] + \bar{\pi}_{O_t j}, & \varphi_t(j) = \arg\max\limits_{1 \le i \le N} [\bar{\eta}_{t-1}(i) + \bar{\gamma}_{ij}] \\ & 1 \le j \le N, \quad 2 \le t \le T \\ \bar{\mathrm{P}}^* = \max\limits_{1 \le i \le N} [\bar{\eta}_T(i)] \\ \iota_T^* = \arg\max\limits_{1 \le i \le N} [\bar{\eta}_T(i)], & \iota_t^* = \varphi_{t+1}(\iota_{t+1}^*) \\ & t = T-1, T-2, \ldots, 1 \end{cases} \tag{46}$$

W wyniku dostajemy taką samą ścieżką $i^* = (i_1^*, i_2^*, ..., i_T^*)$ oraz logarytm prawdopodobieństwa na optymalnej ścieżce $\hat{P}^* = \log(\hat{P}^*)$. Algorytm (46) wymaga N^2T dodawań oraz przeliczenia logarytmów, które wykonuje się raz, dlatego koszt takiego przeliczenie jest nieistotny.

Algorytm Viterbiego estymacji parametrów modelu HMM

Dzięki algorytmowi Viterbiego uzyskujemy optymalną ścieżkę, którą można wykorzystać do estymacji parametrów modelu HMM.

Mamy daną sekwencję obserwacji $\mathbf{O} = (O_1, O_2, ..., O_T)$ oraz optymalną ścieżkę $i^*=(i_1^*, i_2^*, ..., i_T^*)$ dla modelu M. Możemy stworzyć nowy model $\hat{\mathbf{M}}$, w którym liczba stanów będzie taka jak liczba różnych stanów w uzyskanej dzięki algorytmowi Viterbiego optymalnej ścieżce. W wyniku zastosowania takiej metody może nastąpić redukcja ilości stanów w nowym modelu $\hat{\mathbf{M}}$, jeżeli pewne stany nie występują w rozpatrywanej ścieżce. Natomiast parametry nowego modelu są dobierane według zależności:

$$\begin{aligned}
&\hat{\delta}_{\iota_1^*} = 1, \quad \delta_j = 0 \text{ dla } j \neq \iota_1^* \\
&\hat{\gamma}_{ij} = \text{liczba przejść ze stanu } i \text{ do } j \text{ w ścieżce} \\
&\iota^* = (\iota_1^*, \iota_2^*, ..., \iota_T^*) \\
&\hat{\pi}_{kj} = \text{liczba obserwacji } k \text{ w } \mathbf{O} = (O_1, ..., O_T) \\
&\text{dla ustalonego stanu } j \text{ ze ścieżki } \iota^* = (\iota_1^*, ..., \iota_T^*)
\end{aligned} \tag{47}$$

Parametry nowego modelu $\hat{\mathrm{M}}=\left(\hat{\Gamma} = \{\hat{\gamma}_{ij}\}, \hat{\pi} = \{\hat{\pi}_{kj}\}, \hat{\delta} = \{\hat{\delta}_j\}\right)$ po zastosowaniu zależności (47) są następnie normowane.

3.6. Rozwiązanie problemu 3 (Estymacja parametrów)

Trzecim najtrudniejszym problemem w modelu HMM jest stworzenie metod pozwalających na dobór parametrów (Γ, π, δ) modelu tak, aby zoptymalizować prawdopodobieństwo wygenerowania przez model danej obserwacji. Nie istnieją metody analityczne w zamkniętej formie pozwalające na dobór parametrów modelu w taki sposób, aby uzyskać maksymalną wartość prawdopodobieństwa zaobserwowania sekwencji. Jednak możemy dobrać tak parametry M = (Γ, π, δ), że prawdopodobieństwo P(**O**|M) osiąga lokalne maksimum dzięki zastosowaniu metod iteracyjnych, takich jak algorytm Bauma–Welcha znany jako metoda RM (*Expectation-maximization*), metoda jest opisana w pracy [3], lub dzięki metodzie gradientowej, opisanej w pracy [4]. W tym rozdziale zostanie przedyskutowany algorytm Bauma–Welcha pozwalający poprawiać krokowo parametry modelu HMM.

W celu przeprowadzenia procedury estymacji parametrów HMM najpierw zostanie zdefiniowany parametr $\xi_t(i,j)$ opisany równaniem (48), jest to prawdopodobieństwo, że model był w stanie i w czasie t oraz stanie j w czasie t+1 i wygenerował obserwacje **O**

$$\begin{aligned}
\xi_t(i,j) &:= P(S_t = i, S_{t+1} = j / \mathbf{O}) = \\
&= \frac{P(S_t = i, S_{t+1} = j \wedge \mathbf{O})}{P(\mathbf{O})} = \\
&= \frac{\alpha_t(i)\gamma_{ij}\pi_{O_{t+1}j}\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)\gamma_{ij}\pi_{O_{t+1}j}\beta_{t+1}(j)}
\end{aligned} \tag{48}$$

W równaniu (39) został zdefiniowany parametr $\chi_t(i)$. W zależności (49) pokazane są powiązania pomiędzy parametrami $\xi_t(i,j)$ oraz $\chi_t(i)$

$$\begin{aligned}
\chi_t(i) &= \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N}\alpha_{\hat{t}}(i)\beta_{\hat{t}}(i)} = \frac{\alpha_t(i)\cdot\sum_{j=1}^{N}\left[\gamma_{ij}\pi_{O_{t+1}j}\beta_{t+1}(j)\right]}{\sum_{i=1}^{N}\alpha_{\hat{t}}(i)\beta_{\hat{t}}(i)} = \\
&= \sum_{j=1}^{N}\frac{\alpha_t(i)\gamma_{ij}\pi_{O_{t+1}j}\beta_{t+1}(j)}{\sum_{i=1}^{N}\alpha_{\hat{t}}(i)\beta_{\hat{t}}(i)} = \sum_{j=1}^{N}\xi_t(i,j)
\end{aligned} \tag{49}$$

Korzystając z zależności na $\xi_t(i,j)$ oraz $\chi_t(i)$, można podać metodę na estymację parametrów modelu równania (50):

$$\begin{aligned}
\hat{\delta}_i &= \chi_1(i) \\
\hat{\gamma}_{ij} &= \frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\chi_t(i)} \\
\hat{\pi}_{O_k j} &= \frac{\sum_{(po\ \ t\ \ t.\dot{z}.\ S_t=O_k)}^{T-1}\chi_t(j)}{\sum_{t=1}^{T-1}\chi_t(j)}
\end{aligned} \tag{50}$$

Jeżeli do modelu z parametrami M = (Γ, π, δ) zastosujemy zależności (50), otrzymamy model $\hat{\mathbf{M}} = (\hat{\Gamma}, \hat{\pi}, \hat{\delta})$. W zależności od parametrów modelu pierwotnego uzyskamy nowy model, który posiada własność $P(\mathbf{O}|\hat{\mathbf{M}}) \geq P(\mathbf{O}|M)$. Zależności (50) stosuje się iteracyjnie w celu uzyskania odpowiednio dobrych parametrów modelu.

3.7. Wyprowadzenie zależności na estymator parametrów

Chcemy wyznaczyć takie parametry modelu HMM, dla których prawdopodobieństwo obserwacji $\mathbf{O} = (S_1, S_2, ..., S_T)$ przyjmuje największą wartość. Parametry te mogą zostać wyliczone bezpośrednio poprzez minimalizację funkcji pomocniczej $Q(M|\hat{\mathbf{M}})$.

W celu przeprowadzenia dowodu należy wprowadzić funkcję pomocniczą podaną równaniem

$$Q(\mathrm{M}, \hat{\mathrm{M}}) = \sum_{\text{po wszystkich } S} P(\mathbf{O}, S \mid \mathrm{M}) \log P(\mathbf{O}, S \mid \hat{\mathrm{M}}) \quad (51)$$

gdzie suma po stanach $S = (S_1, S, ..., S_T)$ oznacza sumę po wszystkich sekwencjach stanów. Dla wprowadzonej funkcji Q jest prawdziwa zależność

$$Q(\mathrm{M}, \hat{\mathrm{M}}) \geq Q(\mathrm{M}, \mathrm{M}) \Rightarrow P(\mathbf{O} \mid \hat{\mathrm{M}}) \geq P(\mathbf{O} \mid \mathrm{M}) \quad (52)$$

Będziemy maksymalizować funkcję $Q(\mathrm{M}|\hat{\mathrm{M}})$, dobierając parametry modelu, w wyniku czego uzyskamy poprawę prawdopodobieństwa obserwacji $P(\mathbf{O}|\hat{\mathrm{M}})$. Ostatecznie wartość tego prawdopodobieństwa dojdzie do punktu krytycznego podczas procedury iteracyjnej.

Wyrażenia występujące pod sumą (51) można zapisać jak w równaniach:

$$\begin{aligned} P(\mathbf{O}, S \mid \mathrm{M}) &= \delta_{S_1} \left(\prod_{t=1}^{T-1} \gamma_{S_t S_{t+1}} \pi_{O_t S_t} \right) \pi_{O_T S_T} \\ \log P(\mathbf{O}, S \mid \mathrm{M}) &= \log \delta_{S_1} + \sum_{t=1}^{T-1} \log \gamma_{S_t S_{t+1}} + \sum_{t=1}^{T} \log \pi_{O_t S_t} \end{aligned} \quad (53)$$

Korzystając z (53), możemy zapisać $Q(\mathrm{M}|\hat{\mathrm{M}})$

$$\begin{aligned} &Q(\mathrm{M}|\hat{\mathrm{M}}) = \\ &= \sum_{\text{po wszystkich } S} \left\{ P(\mathbf{O}, S \mid M) \log \hat{\delta}_{S_1} + P(\mathbf{O}, S \mid M) \sum_{t=1}^{T-1} \log \hat{\gamma}_{S_t S_{t+1}} + P(\mathbf{O}, S \mid M) \sum_{t=1}^{T} \log \hat{\pi}_{O_t S_t} \right\} = \\ &= \sum_{i=1}^{m} P(\mathbf{O}, S_1 = i \mid M) \log \hat{\delta}_i + \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=1}^{T-1} P(\mathbf{O}, S_t = i, S_{t+1} = j \mid M) \log \hat{\gamma}_{ij} + \\ &+ \sum_{i=1}^{N} \sum_{t=1}^{T} P(\mathbf{O}, S_t = i \mid M) \log \hat{\pi}_{O_t i} \end{aligned} \quad (54)$$

Ponieważ udało nam się przedstawić $Q(\mathrm{M}|\hat{\mathrm{M}})$ w postaci sumy, możemy maksymalizować osobno każdą składową funkcji Q oddzielnie. Na parametry modelu nałożone są ograniczenia opisane zależnościami:

$$\sum_{j=1}^{N} \hat{\delta}_j = 1, \quad \sum_{j=1}^{N} \hat{\gamma}_{ij} = 1 \quad \forall i, \quad \sum_{k=1}^{M} \hat{\pi}_{ki} = 1 \quad \forall i \quad (55)$$

Poszczególne funkcje, które należy maksymalizować, mają postać

$$\sum_{j=1}^{N} w_j \log y_j \quad (56)$$

natomiast ograniczenia mają postać

$$\sum_{j=1}^{N} y_j = 1, \quad y_j \geq 0 \quad (57)$$

Globalne maksimum funkcji (56) jest pojedynczym punktem, co możemy wykazać, korzystając z metody mnożników Lagrange'a, podobnie jak dla modelu procesu Markowa z rozdziału 2. Rozwiązanie przyjmuje postać jak w równaniu

$$y_j = \frac{w_j}{\sum_{j=1}^{N} w_j}, \quad j = 1, 2, ..., N \quad (58)$$

W wyniku maksymalizacji funkcji pomocniczej dostajemy nowy model $\hat{\mathrm{M}} = (\hat{\Gamma}, \hat{\pi}, \hat{\delta})$, jego parametry możemy wyliczyć z zależności:

$$\begin{aligned} \hat{\delta}_i &= \frac{P(\mathbf{O}, S_1 = i)}{P(\mathbf{O})} = \frac{\alpha_1(i)\beta_1(i)}{\sum_{j=1}^{N} \alpha_T(i)} = \chi_1(i) \\ \hat{\gamma}_{ij} &= \frac{\sum_{t=1}^{T-1} P(\mathbf{O}, S_t = i, S_{t+1} = j)}{\sum_{t=1}^{T-1} P(\mathbf{O}, S_t = i)} = \\ &= \frac{\sum_{t=1}^{T-1} \alpha_t(i) \gamma_{ij} \pi_{s_{t+1} j} \beta_{t+1}(i)}{\sum_{t=1}^{T-1} \alpha_t(i) \beta_t(j)} = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \chi_t(i)} \\ \hat{\pi}_{ki} &= \frac{\sum_{t=1}^{T} P(\mathbf{O}, S_t = i) \delta(O_t, k)}{\sum_{t=1}^{T} P(\mathbf{O}, S_t = i)} = \frac{\sum_{t=1}^{T} \alpha_t(i) \beta_t(i) \delta(O_t, k)}{\sum_{t=1}^{T} \alpha_t(i) \beta_t(i)} = \\ &= \frac{\sum_{(\text{po } t \text{ t.ż. } O_t = k)} \chi_t(i)}{\sum_{t=1}^{T} \chi_t(i)} \end{aligned} \quad (59)$$

gdzie $\delta(O_t, k)$ jest opisane równaniem

$$\delta(O_t, k) = \begin{cases} 1 & \text{dla} \quad O_t = k \\ 0 & \text{dla} \quad O_t \neq k \end{cases} \quad (60)$$

Nie istnieją metody wyznaczania globalnego ekstremum.

3.8. System do rozpoznawania izolowanych słów

W niniejszym podrozdziale przedstawiony zostanie system do rozpoznawania izolowanych słów.

Zakładamy, że mamy słownik złożony z V słów, słowa te będą rozpoznawane. Każde ze słów jest zamodelowane przez oddzielny model HMM. Dodatkowo zakładamy, że mamy K wypowiedzi dla każdego słowa, gdzie każda stanowi odrębną sekwencję obserwacji w odpowiedniej reprezentacji.

W celu uzyskania sekwencji obserwacji stosuje się metody przetwarzania dźwięku takie jak LPC (*Linear Predictive Coding*), w wyniku czego uzyskiwane są obserwacje w postaci ciągu wektorów. Dla dyskretnych modeli HMM uzyskane obserwacje przetwarza się, stosując algorytm VQ (*Vector Quantization*) w celu reprezentacji wektorów przez pojedyncze wartości.

1. Dla każdego słowa v ze słownika, musimy zbudować HMM, to znaczy musimy estymować parametry modelu $\mathrm{M} = (\Gamma, \pi, \delta)$, które optymalizują wektor obserwacji odpowiadającej danemu słowu.
2. Dla każdego rozpoznawanego słowa, zostanie zastosowany proces przedstawiony na rysunku 3. Wyliczamy prawdopodobieństwo $\mathrm{P}(\mathbf{O}|\mathrm{M}_v)$ wygenerowania sekwencji dla modeli odpowiadających poszczególnym słowom v ze słownika gdzie $1 \le v \le V$. Następnie wybieramy słowo odpowiadające modelowi, dla którego obliczone prawdopodobieństwo było największe jak w zależności

$$v^* = \arg \max_{1 \le v \le V} [P(\mathbf{O} \mid \mathrm{M}_v)] \tag{61}$$

3.9. Własności wykorzystywane do wyprowadzenia algorytmu Bauma–Welcha

Przedstawione zostaną cztery własności wykorzystywane w podrozdziale 3.4 do wyznaczenia zależności (25), (29) i (30). Wszystkie te własności są prawdziwe dla modelu Markowa opisanego przez zbiór stanów $\{S_t: t \in \mathbf{N}\}$ oraz zbioru obserwacji $\{O_t: t \in \mathbf{N}\}$.

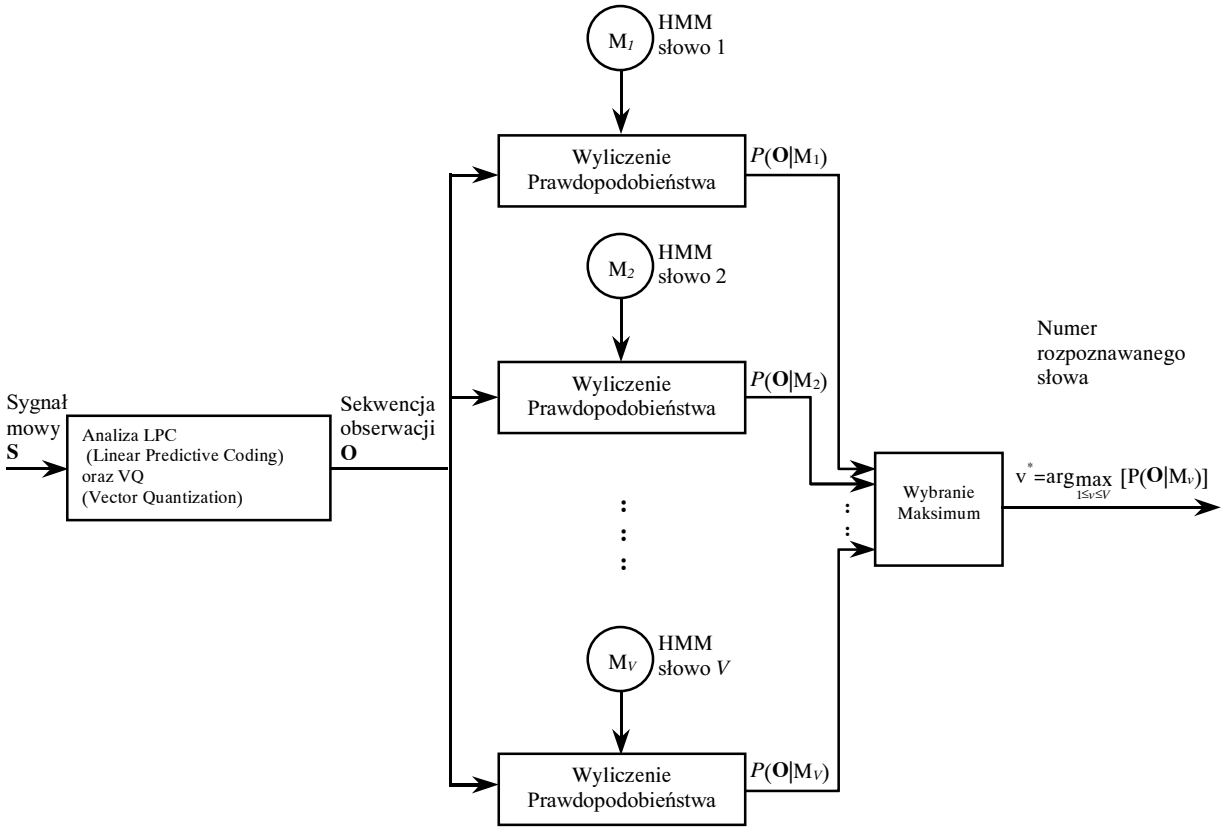
Oba te zbiory są skończoną przestrzenią stanu. Model łańcucha Markowa $\{S_t\}$ oraz sam łańcuch i obserwacje $\{O_t\}$ mają tę własność, że przy określonym ciągu stanów $S^{(T)} = \mathrm{S}_1, \mathrm{S}_2, ..., S_T$, losowe wartości $O_1, O_2, ..., O_T$ są wzajemnie niezależne oraz rozkład prawdopodobieństwa wystąpienia obserwacji O_t jest dany za pomocą zależności $\mathrm{P}(O_t|S_t)$, co oznacza że prawdopodobieństwo to zależy tylko od stanu S_t, natomiast nie zależy od innych stanów.

Własność 1 dla $t = 1, 2, ..., T$ opisana równaniem

$$P(O_1, ..., O_T \mid S_t) = P(O_1, ..., O_t \mid S_t) P(O_{t+1}, ..., O_T \mid S_t) \tag{62}$$

Własność 2 dla $t = 1, 2, ..., T–1$ opisana równaniem

$$\begin{aligned} &P(O_1, ..., O_T \mid S_t, S_{t+1}) = \\ &= P(O_1, ..., O_t \mid S_t) P(O_{t+1}, ..., O_T \mid S_{t+1}) \end{aligned} \tag{63}$$



Rys. 3. Schemat blokowy rozpoznawania izolowanych słów przy użyciu modeli HMM

Własność 3 dla wszystkich l oraz t takich, że $1 \le t \le l \le T$ opisana równaniem

$$P(O_l, ..., O_T \mid S_t, ..., S_l) = P(O_l, ..., O_T \mid S_l) \quad (64)$$

Własność 4 dla $t = 1, 2, ..., T$ opisana równaniem

$$P(O_t, ..., O_T \mid S_t) = P(O_t \mid S_t) P(O_{t+1}, ..., O_T \mid S_t) \quad (65)$$

Dowody własności (62)–(65) można znaleźć w pracy [1].

4. PODSUMOWANIE

W artykule przedstawiono metody pozwalające w praktyce wykorzystywać niejawne modele Markowa. Podstawowym celem artykułu jest pokazanie czytelnikowi koncepcji modelowania przebiegów losowych w czasie za pomocą modeli Markowa, ze szczególnym zwróceniem uwagi na podstawy teoretyczne dotyczące modeli HMM. Zrozumienie dotychczas opracowanych metod dotyczących modeli HMM powinno być dobrym punktem wyjścia do dalszego rozwoju teorii oraz zastosowań modeli Markowa.

Literatura

[1] MacDonald I.L., Zucchini W.: *Hidden Markov and Other Models for Discrete-valued Time Series.* London, Chapman & Hall 1997, ISBN 0-412-55850-5
[2] Juang B.H., Rabiner L.R.: *Hidden Markov models for speach recognition.* Technometrics, 33(3), 1991, 251–272
[3] Dempster A.P, Laird N.M., Rubin D.B.: *Maximum likelihood from incomplete data via the EM algorithm.* J. Roy. Stat. Soc., 39(1), 1977, 1–38
[4] Levison S.E., Rabiner L.R., Sondhi M.M.: *An Introduction to the Application of the Propabilistic Functions if a Markov Process to Automatic Speech Recognition.* Bell System Tech. J., 62, April 1983, 4, 1035–1074
[5] Shomali M., Kapusta M., Gajer M.: *Zastosowanie niejawnych modeli Markowa w systemach automatycznego rozpoznawania mowy.* Kwartalnik AGH, Elektrotechnika i Elektronika, t. 18, z. 3, 1999, 89–98
[6] Gąciarz T.: *Hidden Markov Model (HMM) – opis modelu i algorytmów pod kątem wykorzystania w problemach rozpoznawania mowy i pisma*. Elektrotechnika, t. 17, z. 1, 1998, 17–31
[7] Iosifescu M.: *Skończone procesy Markowa i ich zastosowania.* Warszawa, PWN 1988

Wpłynęło: 30.03.2005

Karol SZOSTEK

Urodził się 22.05.1974 roku Ukończył kierunek automatyka na Wydziale Elektrotechniki, Automatyki Informatyki i Elektroniki Akademii Górniczo-Hutniczej im. Stanisława Staszica w Krakowie. Obecnie jest słuchaczem III roku Studium Doktoranckiego na AGH. Prowadzi badania w dziedzinie modelowania probabilistycznego oraz automatycznego rozpoznawania mowy.

e-mail: kszostek@galaxy.uci.agh.edu.pl