STUDIA INFORMATICA 2002
Volume 23 Number 4 (51)

Piotr FABIAN, Jacek SZEDEL
Politechnika Śląska, Instytut Informatyki

TEORETYCZNE PODSTAWY ZASTOSOWAŃ UKRYTEGO MODELU MARKOWA DO ROZPOZNAWANIA WZORCÓW

Streszczenie. W ostatnim czasie w dziedzinie rozpoznawania wzorców bardzo rozpowszechniły się metody, w których wykorzystuje się tzw. ukryty model Markowa. Mówiąc najogólniej, ukryty model Markowa jest modelem probabilistycznym wykorzystywanym do analizy przebiegów czasowych o losowym charakterze. W niniejszym opracowaniu zaprezentowano podstawy teoretyczne ukrytego modelu Markowa. Rozważania ograniczono do modeli jednowymiarowych z czasem dyskretnym, z dyskretnymi i ciągłymi obserwacjami. Przedstawiona charakterystyka oprócz podstawowych definicji zawiera również obszerne omówienie zagadnień bezpośrednio związanych z zastosowaniami ukrytego modelu Markowa. W końcowej części pracy zawarto krótki opis przykładowych zastosowań ukrytego modelu Markowa w dziedzinie rozpoznawania wzorców.

THEORETICAL BASES OF THE USE OF HIDDEN MARKOV MODELS PATTERN RECOGNITION

Summary. Methods that use Hidden Markov Models (HMM) became recently very popular in the domain of pattern recognition. Generally, a HMM is a probabilistic model used in analysis of random sequences. This article presents theoretical bases of HMMs. Only one-dimensional models with discrete time and continuos observation space have been analyzed. Besides of basic definitions, other problems directly related to applications of HMMs have been discussed. The last part of the article presents some examples of pattern recognition problems, which are solved with HMMs.

1. Wprowadzenie

Prace badawcze z zakresu nauk ścisłych koncentrują się często na opracowaniu modeli matematycznych dla rozważanych zagadnień. Są one bowiem najbardziej wartościowym, z naukowego punktu widzenia, opisem zjawisk i układów fizycznych. Odpowiednio opracowany model matematyczny pozwala przewidzieć przebieg zjawiska lub zachowanie się układu fizycznego w różnych warunkach, bez konieczności przeprowadzania eksperymentów praktycznych. Przydatność modeli matematycznych znacznie wzrosła wraz z rozwojem cyfrowych technik obliczeniowych, ponieważ można je stosunkowo łatwo zapisać w postaci programu komputerowego, a znaczna moc obliczeniowa współczesnych komputerów pozwala na szybkie uzyskanie wyników, nawet w przypadku bardzo złożonych zagadnień.

Zastosowanie modeli matematycznych jest szczególnie pomocne w przypadku analizy sygnałów. Odpowiednie modele przebiegów czasowych i układów generujących te przebiegi stanowią podstawę teoretyczną wielu systemów przetwarzających. Dosyć powszechne są również zastosowania modeli matematycznych w dziedzinie rozpoznawania wzorców. Zastosowania te dotyczą często zagadnień, w których wzorce klasyfikowane są na podstawie przebiegów czasowych o zmiennym i losowym charakterze. W tym zakresie bardzo popularne w ostatniej dekadzie stały się zastosowania tzw. ukrytego modelu Markowa (w skrócie HMM od ang. *Hidden Markov Model*).

Ukryty model Markowa [1,2] jest modelem probabilistycznym stosowanym do modelowania niestacjonarnych przebiegów czasowych. W modelu tym przyjmuje się, że generowaniu sygnałów towarzyszą procesy, mające swe odbicie w zmianach stanu układu, który sygnały te generuje. Zarówno wartości sygnałów, jak i stan układu w danej chwili są losowe. Zakłada się przy tym, że zmiany stanów układu, jak również sygnały przezeń generowane, można opisać za pomocą procesów stochastycznych Markowa.

Zanim przedstawiona zostanie charakterystyka ukrytego modelu Markowa oraz problemy związane z jego zastosowaniami, konieczne jest omówienie, leżących u podstaw modelu, najistotniejszych zagadnień teorii procesów stochastycznych Markowa.

1.1. Łańcuchy Markowa i procesy stochastyczne Markowa

Duża część teorii rachunku prawdopodobieństwa dotyczy analizy zdarzeń niezależnych lub niezależnych zmiennych losowych. W praktyce założenie niezależności przeważnie nie jest spełnione. Z kolei rezygnacja z tego założenia powoduje znaczne skomplikowanie rozważań i obliczeń. Istotne znaczenie w rozwoju metod analizy zależnych zdarzeń losowych mają badania Markowa. W swoich pracach Markow wyodrębnił pewien schemat

doświadczeń, zwany schematem doświadczeń powiązanych w łańcuch Markowa. Rozważania te stały się podstawą teorii procesów stochastycznych Markowa, stanowiącej ważny dział nowoczesnego rachunku prawdopodobieństwa.

Znacznym rozszerzeniem zagadnień analizy schematów zdarzeń losowych lub też ciągów zmiennych losowych jest teoria procesów stochastycznych. Rozszerzenie to polega na rozpatrywaniu nie tylko skończonych lub przeliczalnych zbiorów zmiennych losowych, ale także zbiorów nieprzeliczalnych. Umożliwia to badanie własności statystycznych ciągłych funkcji losowych, takich jak np. przebiegi czasowe o charakterze losowym.

1.1.1. Ciąg zdarzeń powiązanych w łańcuch Markowa

Niech E oznacza skończony lub przeliczalny zbiór zdarzeń losowych, które wykluczają się wzajemnie parami. Rozważmy ciąg N doświadczeń. Załóżmy, że w wyniku każdego doświadczenia może zaistnieć jedno ze zdarzeń należących do zbioru E. Niech $E_j^{(n)}$, $1 \leq n \leq N$ oznacza, że wynikiem n-tego doświadczenia jest zdarzenie E_j. Niech $p_{ij}^{(n)}$ oznacza prawdopodobieństwo warunkowe tego, że wynikiem n-tego doświadczenia jest zdarzenie E_j, pod warunkiem, że w wyniku doświadczenia $(n-1)$-go otrzymano E_i.

$$p_{ij}^{(n)} = P(E_j^{(n)} \mid E_i^{(n-1)}) \tag{1}$$

Ciąg doświadczeń $E_{i_0}^{(0)}, E_{i_1}^{(1)}, \ldots, E_{i_N}^{(N)}$ jest powiązany w łańcuch Markowa, jeżeli dla dowolnych $i, j=1, 2, \ldots, |E|$ oraz dla dowolnego $n=1, 2, \ldots, N$ równości:

$$p_{ij}^{(n)} = P(E_j^{(n)} \mid E_i^{(n-1)}) = P(E_j^{(n)} \mid E_i^{(n-1)}, E_{i_{n-2}}^{(n-2)}, \ldots, E_{i_1}^{(1)}, E_{i_0}^{(0)}) \tag{2}$$

zachodzą przy dowolnych $E_{i_{n-2}}^{(n-2)}, \ldots, E_{i_1}^{(1)}, E_{i_0}^{(0)}$.

Ciąg doświadczeń $E_{i_0}^{(0)}, E_{i_1}^{(1)}, \ldots, E_{i_N}^{(N)}$ jest powiązany w jednorodny łańcuch Markowa, jeżeli dla dowolnych $i, j=1, 2, \ldots, |E|$ prawdopodobieństwo $p_{ij}(n)$ nie zależy od n, tj.:

$$p_{ij}^{(n)} = p_{ij} \qquad (n = 1,2,\ldots,N). \tag{3}$$

W ujęciu bardziej poglądowym zbiór zdarzeń E odpowiadać może zbiorowi stanów fazowych pewnego układu fizycznego, a ciąg zdarzeń $E_{i_0}^{(0)}, E_{i_1}^{(1)}, \ldots, E_{i_N}^{(N)}$ ewolucję tegoż układu w kolejnych chwilach $t=1, 2, \ldots, N$. W takim rozumieniu $p_{ij}^{(n)}$ oznacza prawdopodobieństwo warunkowe tego, że w chwili t układ znajdzie się w stanie E_j pod warunkiem, że w chwili $t-1$ znajdował się w stanie E_i. Prawdopodobieństwo to nazywane jest prawdopodobieństwem przejścia pomiędzy stanami E_i i E_j. Ze wzorów (1) oraz (2) wynika sposób wyznaczenia dla jednorodnego łańcucha Markowa prawdopodobieństwa wystąpienia danej sekwencji stanów $E_{i_0}, E_{i_1}, \ldots, E_{i_n}$ dla n zdarzeń:

$$P(E_{i_0} E_{i_1} \ldots E_{i_n}) = P(E_{i_0}) P(E_{i_1} \mid E_{i_0}) \ldots P(E_{i_n} \mid E_{i_{n-1}}) = P(E_{i_0})\, p_{i_0 i_1} \ldots p_{i_{n-1} i_n}, \tag{4}$$

gdzie $P(E_{i_0})$ oznacza prawdopodobieństwo tego, że w chwili $t=0$ układ znajdował się w stanie E_i.

Macierz, której elementami są prawdopodobieństwa przejścia p_{ij}, określa się mianem macierzy przejścia:

$$A = \begin{bmatrix} p_{11} & p_{12} & p_{13} & \cdots \\ p_{21} & p_{22} & p_{23} & \cdots \\ p_{31} & p_{32} & p_{33} & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}.$$

Rozważania podane dla ciągów zdarzeń losowych przenieść można na ciągi zmiennych losowych.

1.1.2. Proces stochastyczny

Procesem stochastycznym (losowym) $\{X_t, t \in I\}$ jest zbiór zmiennych losowych X zależnych od pewnego parametru t, należącego do pewnego zbioru liczb rzeczywistych I.

W myśl takiej definicji procesem stochastycznym może być ciąg zmiennych losowych X_k, $k=1, 2, \ldots$, wówczas I jest zbiorem liczb naturalnych, lub też k-wymiarowa zmienna losowa i wówczas $I=\{1,2,\ldots,k\}$. W obu przykładach parametr t jest naturalny, a podane procesy stochastyczne są procesami z czasem dyskretnym. Jednakże zbiór I może być również zbiorem nieprzeliczalnym. Mówi się wówczas o procesach z czasem rzeczywistym. Wartość każdej zmiennej X_t procesu stochastycznego $\{X_t, t \in I\}$ zależy od pewnego zdarzenia elementarnego. Dla ustalonego zdarzenia elementarnego wartość X_t jest funkcją losową argumentu t. Funkcja ta nosi nazwę trajektorii lub realizacji procesu stochastycznego. W większości zagadnień teorii procesów stochastycznych parametr t faktycznie oznacza czas. Jednakże możliwa jest sytuacja, gdy znaczenie parametru t jest inne lub też nie ma on żadnego sensu fizycznego.

1.1.3. Proces stochastyczny Markowa

Podobnie jak w przypadku schematów zależnych zdarzeń losowych oraz ciągów zależnych zmiennych losowych powiązanych w łańcuch Markowa, w procesach stochastycznych Markowa zależność zdarzeń uwzględniana jest w ograniczonym zakresie.

Jeśli dla każdego $n=1,2,\ldots$ i dowolnych wartości parametru $t_m \in I$ $(m=1,2,\ldots,n)$, gdzie $t_1<t_2<\ldots<t_n$ oraz dla dowolnych liczb rzeczywistych a i b równość:

$$P(X_{t_n} < a \mid X_{t_{n-1}} = b, X_{t_{n-2}} = x_{n-2}, \ldots, X_{t_1} = x_1, X_{t_0} = x_0) = P(X_{t_n} < a \mid X_{t_{n-1}} = b), \tag{5}$$

zachodzi dla dowolnych liczb rzeczywistych $x_{n-2}, \ldots, x_0$, wówczas proces $\{X_t, t \in I\}$ jest procesem stochastycznym Markowa.

Oznacza to, że na warunkowy rozkład prawdopodobieństwa zmiennej losowej X_{t_n} pod warunkiem, że zmienna losowa $X_{t_{n-1}}=b$, nie mają wpływu żadne dodatkowe informacje o wartościach zmiennych losowych rozważanego procesu w chwilach poprzedzających moment t_{n-1}. Występującą w równaniu (5) dystrybuantę warunkową dla dowolnych dwóch wartości $t_1 \in I$, $t_2 \in I$ określić można wzorem:

$$F(t_1,a,t_2,b) = P(X_{t_2} < a \mid X_{t_1} = b). \tag{6}$$

Proces stochastyczny Markowa $\{X_t, t \in I\}$ jest jednorodny, jeżeli dla dowolnych $t_1 \in I$, $t_2 \in I$, gdzie $t_1 < t_2$, dystrybuanta (6) zależy jedynie od wartości różnicy $\Delta t = t_2 - t_1$.

$$F(t_1,a,t_2,b) = F(\Delta t,a,b). \tag{7}$$

W zastosowaniach dosyć często spotyka się procesy stochastyczne, w których zmienne losowe mają charakter skokowy, a punktami skokowymi tych zmiennych są liczby całkowite. Taki proces określa się mianem procesu skokowego. Dla takiego przypadku można posłużyć się prawdopodobieństwem warunkowym zamiast warunkową dystrybuantą.

Proces skokowy $\{X_t, t \in I\}$ jest procesem skokowym Markowa, jeśli dla każdego $n=1,2,...$ i dowolnych wartości parametru $t_m \in I$ $(m=1,2, ... ,n)$, gdzie $t_1 < t_2 < ... < t_n$ oraz dla dowolnych liczb rzeczywistych a i b równość:

$$P(X_{t_n} = a \mid X_{t_{n-1}} = b, X_{t_{n-2}} = i_{n-2}, ..., X_{t_1} = i_1, X_{t_0} = i_0) = P(X_{t_n} = a \mid X_{t_{n-1}} = b) \tag{8}$$

zachodzi dla dowolnych liczb całkowitych $i_{n-2}, ... ,i_0$.

Należy zauważyć, że w przypadku skokowego procesu Markowa, w którym parametr t_m przyjmuje wartości całkowite, ciąg zdarzeń polegających na tym, że zmienne losowe X_0, X_1, ..., X_n przyjmują określone wartości, jest ciągiem zdarzeń powiązanych w łańcuch Markowa. Tak więc łańcuch Markowa jest szczególnym przypadkiem procesu stochastycznego Markowa i można go zdefiniować jako skokowy proces Markowa, w którym parametr t_m przyjmuje wartości całkowitoliczbowe $t_m = 0,1,2,...$.

Dla porządku podać należy jeszcze definicję jednorodnego procesu skokowego Markowa: Skokowy proces Markowa $\{X_t, t \in I\}$ jest jednorodny, jeżeli dla dowolnych i, j oraz dowolnych $t_1 \in I$, $t_2 \in I$, gdzie $t_1 < t_2$, prawdopodobieństwa przejścia zależą jedynie od różnicy $\Delta t = t_2 - t_1$.

2. Charakterystyka ukrytego modelu Markowa

2.1. Obszar zastosowań

O modelu Markowa mówić można wówczas, gdy ciąg zmiennych losowych reprezentujący jakieś zagadnienie jest ciągiem zmiennych losowych powiązanych w łańcuch Markowa bądź też tworzy proces stochastyczny Markowa.

Przykładem często przytaczanym w literaturze jest przypadkowe błądzenie cząstki po całkowitoliczbowych punktach prostej rzeczywistej. W każdym kroku cząstka z prawdopodobieństwem p przesuwa się o jednostkę w prawo, a z prawdopodobieństwem $q=1-p$ o jednostkę w lewo.

Niech S_n będzie zmienną losową przyjmującą wartość określającą pozycję cząstki w n-tym kroku. Ciąg $S_1, S_2, S_3, \ldots$ jest ciągiem zmiennych losowych powiązanych w łańcuch Markowa. Równość

$$P(S_n = a \mid S_{n-1} = b, S_{n-2} = s_{n-2}, \ldots, S_1 = s_1) = P(S_n = a \mid S_{n-1} = b)$$

$$= \begin{cases} p & dla \quad a-b=-1 \\ q=1-p & dla \quad a-b=1 \end{cases} \tag{9}$$

zachodzi dla dowolnych liczb całkowitych $s_{n-2}, s_{n-2}, \ldots, s_1$ oraz a i b, gdzie dodatkowo $|a-b|=1$, $a \le n$, $b \le n-1$.

W powyższym przykładzie poruszająca się chaotycznie cząstka jest układem fizycznym, którego stany wyznaczone są przez punkty całkowitoliczbowe prostej. Gdy cząsteczka znajduje się w punkcie n, układ znajduje się w n-tym stanie. Przyjmując, że pozycja cząstki S_n odczytywana jest co pewien czas (np. w chwili, gdy układ zmienia stan), następujące po sobie zmiany stanu generują niejako kolejne wartości pewnego sygnału $S(t)=S_n$ dla $n \le t \le n+1$. W tym rozumieniu przedstawiony układ jest źródłem sygnału, którego wartość związana jest ze stanem układu w danej chwili. Przedstawiony model matematyczny określany jest mianem markowowskiego modelu źródła sygnałów.

Załóżmy teraz, że z pewnych względów nie jest możliwe bezpośrednie określenie pozycji cząstki w przedstawionym powyżej przykładzie. Zamiast tego, stosując odpowiedni przyrząd umieszczony w punkcie x_p prostej, odczytać można pewną wielkość V, która jest proporcjonalna do położenia cząstki. Niech V_n oznacza zmienną losową przyjmującą wartość odczytaną w n-tym kroku. Obserwując zachowanie się cząstki w kolejnych krokach otrzymamy ciąg odczytów $V_1, V_2, V_3, \ldots$. Ciąg zmiennych losowych $V_1, V_2, V_3, \ldots$ przyjmujących wartości kolejnych odczytów powiązany jest w łańcuch Markowa. Równanie (9) zastąpić można równaniem:

$$P(V_n = a \mid V_{n-1} = b, V_{n-2} = v_{n-2}, \ldots, V_1 = v_1) = P(V_n = a \mid V_{n-1} = b) = \tag{10}$$

$$= \begin{cases} p & dla \quad a - b = -1 \\ q = 1 - p & dla \quad a - b = 1 \end{cases}$$

Zakładając liniową zależność wielkości V od pozycji cząstki, pozycję wyznaczyć można ze wzoru:

$$S_n = x_p + \frac{V_n}{c}, \tag{11}$$

gdzie c jest pewnym stałym współczynnikiem.

Opisana sytuacja nie różni się zasadniczo od przypadku, w którym pozycja cząstki odczytywana jest bezpośrednio, chyba że ze względu na niedoskonałość zastosowanego przyrządu odczyty wielkości V_n obarczone będą pewnym losowym błędem ε_n. Wówczas

$$V_n = V_n^{(0)} + \varepsilon_n, \tag{12}$$

gdzie $V_n^{(0)}$ oznacza dokładną wartość odczytywanej wielkości. Podstawiając V_n do równania (11) otrzymujemy:

$$S_n = x_p + \left[\frac{V_n^{(0)} + \varepsilon_n}{c} \right], \tag{13}$$

gdzie nawiasy kwadratowe oznaczają funkcję zaokrąglania. Pozycja cząsteczki S_n jest zatem funkcją losową odczytanej wartości V_n. W przypadku gdy wartość ilorazu ε_n / c przekroczy 0.5, jednoznaczne określenie pozycji cząstki nie będzie możliwe. Pozycja cząsteczki stanie się w pewnym sensie niejawna lub, mówiąc inaczej, ukryta.

W sytuacjach podobnych do tej, którą przedstawiono, zastosowanie znajduje ukryty model Markowa. W modelu tym wartości zmiennych losowych tworzących proces stochastyczny Markowa $\{X_t, t \in I\}$, reprezentujący stany układu, nie są znane. Znane są jedynie wartości zmiennych losowych innego procesu stochastycznego Markowa $\{Y_t, t \in I\}$. Wartości zmiennych losowych tworzących proces $\{Y_t, t \in I\}$ są funkcją losową wartości zmiennych losowych procesu $\{X_t, t \in I\}$. Postulat Markowa dla takiej sytuacji można zapisać w formie:

$$P(Y_k < b \mid (X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_1 = x_1) \wedge \tag{14}$$

$$\wedge (Y_{k-1} = y_{k-1}, Y_{k-2} = y_{k-2}, \ldots, Y_1 = y_1)) =$$

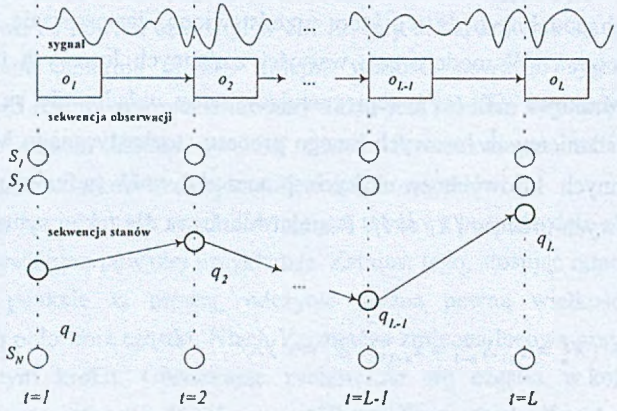
$$= P(Y_{t_k} < b \mid X_{t_k} = x_k)$$

dla każdego n i k, gdzie $n \geq k$ oraz dla dowolnej liczby rzeczywistej x i dowolnych liczb rzeczywistych $x_{k-1}, \ldots, x_1$, $y_{k-1}, \ldots, y_1$. Inaczej mówiąc rozkład zmiennej Y_k procesu stochastycznego Markowa $\{Y_t, t \in I\}$ zależy jedynie od wartości zmiennej losowej X_k innego procesu stochastycznego Markowa $\{X_t, t \in I\}$.

Ukryty model Markowa wprowadza rozgraniczenie procesów związanych z obserwowanymi wartościami sygnału a faktycznym stanem układu. Pozwala to na uniezależnienie się od losowych zmian sygnału wynikających z zakłóceń lub też z losowego charakteru jego źródła. Powstają zatem dogodne możliwości zastosowań ukrytego modelu Markowa w rozpoznawaniu wzorców, gdzie podstawową trudnością jest uniezależnienie wyników klasyfikacji od losowych różnic pomiędzy wzorcami należącymi do tej samej klasy.

2.2. Podstawowe założenia

W zastosowaniach ukrytego modelu Markowa rozważane zagadnienia najwygodniej rozpatruje się w kontekście układu generującego sygnały. Zakłada się, że w każdej chwili układ znajduje się w określonym s t a n i e (rys. 1). Stany definiuje się tak, aby odpowiadały faktycznym właściwościom układu. Zbiór stanów układu oznacza się zwykle symbolem S, gdzie $S=\{S_1, S_2, \ldots, S_N\}$, a N jest liczbą możliwych stanów układu. W modelach z czasem dyskretnym wartości sygnałów odczytywane są w kolejnych momentach czasu $t_1<t_2<\ldots<t_L$, gdzie $t_i=1, 2, 3, \ldots$. Zakłada się, że w chwili dokonywania odczytu następuje zmiana stanu układu bądź też ponowne przejście do tego samego stanu. Na podstawie odczytanych wartości sygnału wyznaczane są wielkości charakteryzujące sygnał w kolejnych chwilach $t_1, t_2, \ldots, t_L$. Wartości te noszą nazwę o b s e r w a c j i .



Rys. 1. Interpretacja działania układu przyjęta w ukrytym modelu Markowa
Fig. 1. An interpretation of the hidden Markov process

2.3. Sposób reprezentacji procesów losowych

Niech $\{Q_t, t \in I\}$ oznacza proces stochastyczny, który w ukrytym modelu Markowa związany jest ze zmianami stanów układu – proces ten to proces stanów. Zmienne losowe $Q_1, Q_2, ..., Q_L$ tego procesu przyjmują wartości dyskretne z zakresu $<1,N>$, gdzie N jest liczbą możliwych stanów układu. Ponadto parametr t przyjmuje wartości całkowitoliczbowe. Proces stochastyczny $\{Q_t, t \in I\}$ jest zatem procesem skokowym Markowa z czasem całkowitym, a ciąg zdarzeń polegających na tym, że zmienne losowe $Q_1, Q_2, ..., Q_L$ przyjmują wartości $q_1, q_2, ..., q_L$, jest ciągiem zdarzeń powiązanych w łańcuch Markowa [1]. Właściwości procesu stanów można zatem scharakteryzować za pomocą macierzy o N wierszach i N kolumnach, zawierającej prawdopodobieństwa przejścia pomiędzy dwoma stanami w kolejnych chwilach czasu. Macierz tę oznacza się najczęściej symbolem A, a jej elementy symbolem a_{ij}, gdzie

$$a_{ij} = P(Q_t = S_i \mid Q_{t-1} = S_j) \tag{15}$$

jest prawdopodobieństwem warunkowym tego, że układ znajdzie się w chwili t w stanie S_j pod warunkiem, że w chwili $t+1$ znajdował się w stanie S_i.

Uzupełnieniem reprezentacji procesu stanów jest, oznaczany zwykle symbolem Π, wektor prawdopodobieństw początkowych, którego elementami są prawdopodobieństwa π_i tego, że w chwili $t=1$ układ znajduje się w stanie S_i. Prawdopodobieństwa te określa się mianem prawdopodobieństw początkowych.

Inna jest reprezentacja procesu stochastycznego, którego realizacją w ukrytym modelu Markowa jest ciąg obserwacji. Dla procesu tego przyjęto oznaczenie $\{O_t, t \in I\}$. Proces ten określany jest mianem procesu obserwacji. W zależności od rozpatrywanego zagadnienia zmienne losowe procesu obserwacji mogą mieć charakter ciągły lub dyskretny. Uwzględniając fakt, że proces $\{O_t, t \in I\}$ jest procesem stochastycznym Markowa, dla pełnego scharakteryzowania jego właściwości konieczna byłaby znajomość dystrybuant warunkowych

$$F_n(x,y) = P(O_n < x \mid O_{n-1} = y) \quad \text{dla} \quad n = 1,2,...,L. \tag{16}$$

Korzystając jednak z faktu, że obserwacja jest funkcją losową stanu układu, skorzystać można z dystrybuanty warunkowej występującej w równaniu (14). Dystrybuanta ta dla stanu S_i dana jest wzorem:

$$F_{S_i}(x) = P(O_k < x \mid Q_k = S_i). \tag{17}$$

---

[1] Dla przejrzystości dalszych rozważań sekwencja stanów rozpatrywana będzie w dziedzinie procesów stochastycznych.

Dystrybuanty $F_{s_1}(x)$, $F_{s_2}(x)$, ..., $F_{s_N}(x)$ wyznaczają warunkowe rozkłady prawdopodobieństwa zmiennych losowych procesu $\{O_t, t \in I\}$. Warunkiem jest pozostawanie układu w określonym stanie. Rozkłady te określa się mianem rozkładów obserwacji. Rozkład obserwacji dla stanu S_i oznacza się symbolem $b_i(O)$. Całościową reprezentację procesu obserwacji w ukrytym modelu Markowa stanowi tzw. wektor rozkładów obserwacji $B=\{b_1(O), b_2(O),...,b_N(O)\}$.

2.4. Definicja modelu

Wymienione w poprzednim rozdziale elementy reprezentujące procesy stochastyczne stanów i obserwacji w ukrytym modelu Markowa uwzględnia przyjęta w literaturze definicja modelu.

Ukrytym modelem Markowa λ o N stanach $\{S_1, ... ,S_N\}$ nazywamy trójkę:

$$\lambda=(\Pi,A,B),$$

gdzie:

- $\Pi=\{\pi_1, \pi_2, ..., \pi_N\}$ jest wektorem prawdopodobieństw wejściowych, a π_i jest prawdopodobieństwem, że początkowym stanem układu będzie stan S_i.
- $A=\{a_{i,j}\}_{0<i<=N,\ 0<j<=N}$ jest macierzą przejść zawierającą prawdopodobieństwa $a_{i,j}$ przejścia ze stanu S_i do stanu S_j.
- $B=(b_i(O))_{0<j<=N}$ jest wektorem rozkładów obserwacji, gdzie $b_i(O)$ jest rozkładem prawdopodobieństwa zmiennej losowej związanej z wartością obserwacji w stanie S_i.

3. Najważniejsze zagadnienia i problemy związane z ukrytym modelem Markowa

Do najważniejszych zagadnień związanych z ukrytym modelem Markowa i jego zastosowaniami należą: wyznaczanie prawdopodobieństwa wygenerowania określonej sekwencji obserwacji przez dany układ, określenie najbardziej prawdopodobnej sekwencji stanów układu dla danej sekwencji obserwacji oraz reestymacja parametrów modelu na podstawie wzorców uczących. Innym problemem, istotnym z punktu widzenia implementacji, jest tzw. skalowanie obliczeń. Umożliwia ono ominięcie ograniczeń wynikających ze zbyt małej precyzji obliczeniowej. Omówienia wymaga również ogólna koncepcja klasyfikacji wzorców z zastosowaniem ukrytego modelu Markowa.

3.1. Prawdopodobieństwo wygenerowania sekwencji obserwacji

Aby wyznaczyć prawdopodobieństwo $P(O|\lambda)$ tego, że sekwencja obserwacji $O=o_1,o_2,...,o_L$ zostanie wygenerowana przez układ dany modelem λ, należałoby skorzystać ze wzoru na prawdopodobieństwo wystąpienia sekwencji określonych wartości w jednorodnym procesie stochastycznym Markowa. W tym przypadku wzór ten zapisać można w następujący sposób [2]

$$P(O|\lambda) = P(o_1)P(o_2|o_1)...P(o_L|o_{L-1}). \tag{18}$$

Należy jednak pamiętać, że w ukrytym modelu Markowa zależności pomiędzy kolejnymi zmiennymi losowymi procesu obserwacji nie są znane. Informacje o właściwościach procesu obserwacji wyrażone są pośrednio: z jednej strony poprzez zależności pomiędzy stanami układu, z drugiej zaś przez rozkłady obserwacji określone dla poszczególnych stanów.

Rozważmy następujące zdarzenia losowe. Zdarzenie polegające na wystąpieniu określonej sekwencji stanów $Q=q_1, q_2, ..., q_L$ w modelu λ, zdarzenie polegające na wygenerowaniu przez model λ określonej sekwencji obserwacji $O=o_1, o_2,..., o_L$, dla ustalonej trajektorii stanów Q oraz zdarzenie polegające na tym, że określona sekwencja stanów i określona sekwencja obserwacji pojawią się jednocześnie. Prawdopodobieństwo zajścia pierwszego zdarzenia dane jest wzorem:

$$P(Q|\lambda) = P(q_1)P(q_2|q_1)...P(q_L|q_{L-1}), \tag{19}$$

zaś prawdopodobieństwo drugiego zdarzenia wzorem:

$$P(O|Q,\lambda) = P(o_1|q_1)P(o_2|q_2)...P(o_L|q_L). \tag{20}$$

Prawdopodobieństwo jednoczesnego wystąpienia określonej sekwencji stanów i określonej sekwencji obserwacji w modelu λ dane jest iloczynem

$$P(O,Q|\lambda) = P((o_1,o_2,...,o_L)\wedge(s_1,s_2,...,s_L)) = P(O|Q,\lambda)P(Q|\lambda) \tag{21}$$

Uwzględniając zależności (20) i (21) otrzymamy:

$$P(O,Q|\lambda) = P(o_1|q_1)P(o_2|q_2)...P(o_L|q_L)\cdot P(q_1)P(q_2|q_1)...P(q_L|q_{L-1}) =$$

$$= P(q_1)P(o_1|q_1)P(q_2|q_1)P(o_2|q_2)\cdot...P(q_L|q_{L-1})P(o_L|q_L) =$$

$$= P(q_1)P(o_1|q_1)\prod_{t=2}^{L}P(q_t|q_{t-1})P(o_t|q_t). \tag{22}$$

---

[2] Dla przejrzystości, w dalszych rozważaniach dla oznaczenia wartości procesów zmiennych tworzących procesy losowe przyjęto nieco uproszczony zapis: na przykład zamiast zapisać $P(Q_1=q_1,Q_2=q_2,...Q_L=q_L)$ zapisano po prostu $P(q_1,q_2,...,q_L)$ lub zamiast $P(O_1=o_1|Q_1=q_1)$ zapisano $P(o_1|q_1)$.

Przyjmując oznaczenia podane w definicji 10 wzór (22) można zapisać w postaci:

$$P(O,Q\,|\,\lambda)=\pi_{s_1}b_{s_1}(o_1)\prod_{t=2}^{L}a_{s_{t-1}s_t}b_{s_t}(o_t)\,. \tag{23}$$

$P(O,Q|,\lambda)$ jest zatem iloczynem prawdopodobieństw przejść pomiędzy kolejnymi stanami $q_1, q_2, ..., q_L$ oraz prawdopodobieństw tego, że obserwacja przyjmować będzie dla tych stanów wartości $o_1, o_2,..., o_L$.

Należy zauważyć, że dana sekwencja obserwacji może pojawić się w efekcie wystąpienia różnych sekwencji stanów. Zatem, aby obliczyć prawdopodobieństwo $P(O|\lambda)$ wystąpienia ciągu obserwacji O w modelu λ, należy uwzględnić sumę wszystkich możliwych trajektorii stanów w chwilach $t=1,2,...,L$. Zdarzenia polegające na wystąpieniu sekwencji obserwacji O w efekcie przejścia układu przez dwie różne trajektorie stanów są rozłączne, zatem

$$P(O\,|\,\lambda)=\sum_{Q\in\Gamma_\lambda}P(O,Q\,|\,\lambda)\,, \tag{24}$$

gdzie Γ_λ oznacza zbiór wszystkich możliwych sekwencji stanów w modelu λ.

Nietrudno zauważyć, że wyznaczenie prawdopodobieństwa $P(O|\lambda)$ jest zadaniem o dużej złożoności obliczeniowej. W najgorszym przypadku należy bowiem uwzględnić N^L możliwych sekwencji stanów, natomiast każdorazowe obliczenie prawdopodobieństwa $P(O,Q|\lambda)$ wymaga wykonania $2L$ mnożeń. Złożoność obliczeniowa uwzględniająca operacje mnożenia i dodawania jest więc $O(N^L)$. Biorąc pod uwagę fakt, że w większości zastosowań spotyka się modele o liczbie stanów od kilku do kilkudziesięciu, a rozpatrywany ciąg obserwacji może składać się kilkudziesięciu elementów, wyznaczenie prawdopodobieństwa $P(O|\lambda)$ na podstawie wzoru (26) jest praktycznie niemożliwe.

Istnieje jednak algorytm pozwalający w inny sposób wyznaczyć prawdopodobieństwo wystąpienia ciągu obserwacji w danym modelu. Jego złożoność obliczeniowa jest znacznie mniejsza. Algorytm ten, znany w literaturze pod nazwą „w przód – w tył" (ang. *forward-backward*), to w rzeczywistości dwie procedury o podobnym działaniu. Ich istota sprowadza się do sekwencyjnego wyznaczania prawdopodobieństw wystąpienia częściowych sekwencji obserwacji dla kolejnych chwil i zastosowania otrzymanego wyniku w kolejnej iteracji. Pierwsza ze wspomnianych procedur, procedura „w przód", korzysta z prawdopodobieństwa $\alpha_t(i)$ tego, że przejściu układu do stanu S_i w chwili t towarzyszyła sekwencja obserwacji $o_1, o_2, ..., o_t$:

$$\alpha_t(i)=P(o_1,o_2,...,o_t,q_t=S_i\,|\,\lambda)\,. \tag{25}$$

Prawdopodobieństwa $\alpha_t(i)$ wyznaczane są dla kolejnych chwil czasu, dla wszystkich stanów układu, w następujący sposób:

//inicjalizacja
for i = 1 to N do
$\alpha_1(i) = \pi_i b_i(o_1)$
end for
//indukcja
for t = 1 to L-1 do
for j = 1 to N do
$\alpha_{t+1}(j) = [\sum_{i=1}^{N} \alpha_t(i) a_{ij}] b_j(o_{t+1})$
end for
end for
//zakończenie
$P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_L(i)$

Prawdopodobieństwo $P(O|\lambda)$ jest sumą wyników uzyskanych dla $t=L$.

W procedurze „w tył" wyznaczane są prawdopodobieństwa $\beta_t(i)$ tego, że po wejściu do stanu S_i w chwili t układ wygeneruje sekwencję obserwacji $o_{t+1}, o_{t+2}, ..., o_L$:

$$\beta_t(i) = P(o_{t+1}, o_{t+2}, ..., o_L \mid q_t = S_i, \lambda). \tag{26}$$

Zapis algorytmu wyznaczania prawdopodobieństwa $P(O|\lambda)$ z użyciem częściowego prawdopodobieństwa $\beta_t(i)$ podano poniżej. Złożoność obu przedstawionych procedur jest wielomianowa.

3.2. Najbardziej prawdopodobna sekwencja stanów

Kolejnym ważnym zagadnieniem dotyczącym ukrytego modelu Markowa jest zadanie wyznaczenia dla danej sekwencji obserwacji O i danego modelu λ najbardziej prawdopodobnej sekwencji stanów układu. Zagadnienie to jest typowym problemem optymalizacji dyskretnej. Dla modelu o N stanach i sekwencji obserwacji o długości L przestrzeń rozwiązań zawiera N^L możliwych ciągów. Rozwiązanie w dużej mierze zależy od zastosowanych kryteriów optymalizacyjnych. Można na przykład przyjąć kryterium polegające na wyborze stanów, które są indywidualnie najbardziej prawdopodobne dla kolejnych elementów ciągu obserwacji. Mówiąc ściślej, dla każdego $t=1,2,...L$ należy wskazać stan układu, dla którego wartość prawdopodobieństwa

$$\gamma_t(i) = P(q_t = S_i \mid O, \lambda) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)} \tag{27}$$

jest największa ($\alpha_t(i)$ i $\beta_t(i)$, są prawdopodobieństwami dla częściowych sekwencji zastosowanymi w procedurze „w przód - w tył"). Prawdopodobieństwo $\gamma_t(i)$ jest prawdopodobieństwem zdarzenia polegającego na tym, że układ dany modelem λ znajdzie się w chwili t w stanie S_i dla sekwencji obserwacji O. Rozwiązanie oparte na przedstawionym kryterium będzie miało jednak bardzo istotną wadę. Kryterium to nie uwzględnia bowiem prawdopodobieństw przejść pomiędzy stanami. Może zatem zaistnieć sytuacja, w której otrzymana optymalna sekwencja stanów $q_1, q_2, ..., q_L$ nigdy w rozpatrywanym modelu nie wystąpi. Będzie tak wówczas, gdy w sekwencji tej pojawi się przynajmniej jedna para stanów q_t, q_{t+1}, dla których prawdopodobieństwo przejścia $a_{q_t q_{t+1}}$ jest zerowe. Należałoby zatem zmodyfikować kryterium optymalizacyjne na przykład tak, aby uwzględniało ono maksymalne wartości prawdopodobieństw dla par stanów (q_t, q_{t+1}), dla trójek (q_t, q_{t+1}, q_{t+2}) lub, co jest najlepszym rozwiązaniem, dla całej sekwencji $q_1, q_2, ..., q_L$. Rozwiązanie oparte na kryterium optymalizacyjnym uwzględniającym całe sekwencje stanów podał J.A.Viterbi 345. Rozwiązanie to znane jest w literaturze jako a l g o r y t m V i t e r b i e g o.

Działanie algorytmu Viterbiego sprowadza się do wyznaczenia dla danej sekwencji obserwacji O takiej sekwencji stanów Q, dla której wartość prawdopodobieństwa $P(O|Q,\lambda)$ jest największa. W tym celu zastosowano funkcję $\delta_t(i)$ zdefiniowaną następująco:

$$\delta_t(i) = \max_{q_1, q_2, ..., q_{t-1}} P(q_1, q_2, ..., q_t = S_i, o_1, o_2, ..., o_t \mid \lambda). \tag{28}$$

Dla danego t funkcja ta przyjmuje wartość maksymalnego prawdopodobieństwa wygenerowania częściowej sekwencji obserwacji $o_1, o_2, ..., o_t$ dla ciągu stanów $q_1, q_2, ..., q_t$, gdzie $q_t = S_i$. Wartości $\delta_t(i)$ dla $t=1$ wyznaczyć można ze wzoru

$$\delta_1(i) = P(q_1 = S_i) P(o_1 \mid S_i) = \pi_i b_i(o_1). \tag{29}$$

Wartości funkcji dla pozostałych chwil czasu $t=2, 3, ..., L$ można wyznaczyć indukcyjnie w następujący sposób: dla każdego stanu S_j, $j=1, 2, ..., N$ należy wybrać stan S_i, przy którym iloczyn wartości funkcji $\delta_t(i)$ i prawdopodobieństwa przejścia pomiędzy stanem S_j i stanem S_i jest największy, a następnie maksymalną wartość tego iloczynu pomnożyć przez prawdopodobieństwo wystąpienia w stanie S_j określonej obserwacji. Opisane postępowanie wyrazić można następującym wzorem indukcyjnym:

$$\delta_{t+1}(j) = \max_{1 \le i \le N} [\delta_t(i) P(q_{t+1} = S_j \mid q_t = S_i)] P(o_{t+1} \mid S_j) = \max_{1 \le i \le N} [\delta_t(i) a_{ij}] b_j(o_{t+1}). \tag{30}$$

Niech Ψ oznacza macierz o N wierszach i L kolumnach, której elementami ψ_{ij} są indeksy i stanów układu, dla których wartość iloczynu $\delta_{t-1}(i) a_{ij}$ była największa

$$\Psi = \psi_{ij} = \begin{cases} \arg\max_{1 \le i \le L} [\delta_t(i) a_{ij}] & dla \quad t = 2,3,...,L, \\ 0 & dla \quad t = 1. \end{cases} \tag{31}$$

Najbardziej prawdopodobny dla sekwencji obserwacji $o_1, o_2, ..., o_L$ ciąg stanów $q^*_1, q^*_2, ..., q^*_L$ wyznaczany jest na podstawie wartości elementów macierzy Ψ dla $t=L-1, L-2, ..., 1$ zgodnie ze wzorem

$$q^*_t = \psi_{t+1,k}, \tag{32}$$

gdzie $k=q^*_{t+1}$. Zapis algorytmu Viterbiego podano poniżej.

```
//inicjalizacja
for i = 1 to N do
        δ1(i)=πibi(o1);
        ψ1i=0;
end for
//indukcja
for t = 2 to L do
        for j = 1 to N do
                δt(j) = max[δt-1(i)aij]bj(ot);
                1≤i≤n
                ψtj = arg max[δt-1(i)aij];
                1≤i≤N
        end for
end for
//zakończenie
P*(O|λ) = max[δL(i)];
          1≤i≤N
q*L = arg max[δL(i)];
      1≤i≤N
//wyznaczenie wynikowej sekwencji stanów
for t = L-1 downto 1 do
        k = q*t+1;
        q*t = ψt+1,k;
end for
```

W wyniku zastosowania algorytmu Viterbiego, oprócz sekwencji stanów Q^*, otrzymuje się również wartość prawdopodobieństwa $P^*(O|\lambda)$ wystąpienia w modelu λ rozpatrywanego ciągu obserwacji O. Złożoność obliczeniowa algorytmu Viterbiego, podobnie jak i procedur „w przód" i „w tył", jest $O(N^2)$.

Należy zauważyć, że działanie algorytmu Viterbiego jest podobne do działania opisanej wcześniej procedury „w przód". Istnieje jednak pewna różnica, która polega na tym, że w algorytmie Viterbiego sumowanie prawdopodobieństw $\alpha_t(i)$ zastąpiono wyborem maksymalnego iloczynu $\delta_{t-1}(i)a_{ij}$.

3.3. Zagadnienie reestymacji parametrów modelu

Celem reestymacji parametrów modelu $\lambda=(\Pi,A,B)$ jest wyznaczenie dla zadanej sekwencji obserwacji O modelu $\overline{\lambda}=(\overline{\Pi},\overline{A},\overline{B})$ o tak zmienionych parametrach, aby prawdopodobieństwo $P(O|\overline{\lambda})$ było jak największe. Inaczej mówiąc, celem reestymacji jest rozszerzenie zbioru mogących pojawić się w danym modelu ciągów obserwacji o nową sekwencję O. Analityczne rozwiązanie tak sformułowanego problemu jest rzeczą niezmiernie trudną. Można jednak poprzestać na wyznaczeniu parametrów modelu, dla których $P(O|\lambda)$ jest lokalnie jak największe, tzn. tylko dla jednej sekwencji stanów. Najszerzej znanym rozwiązaniem problemu reestymacji parametrów w ukrytym modelu Markowa jest metoda Bauma-Welsha [6,7,8].

Niech $\xi_t(i,j)$ oznacza prawdopodobieństwo tego, że układ dany modelem λ dla danej sekwencji obserwacji O znajdował się w chwili t w stanie S_i, a w chwili $t+1$ w stanie S_j:

$$\xi_t(i,j)=P(q_t=S_i,q_{t+1}=S_j\,|\,O,\lambda). \tag{33}$$

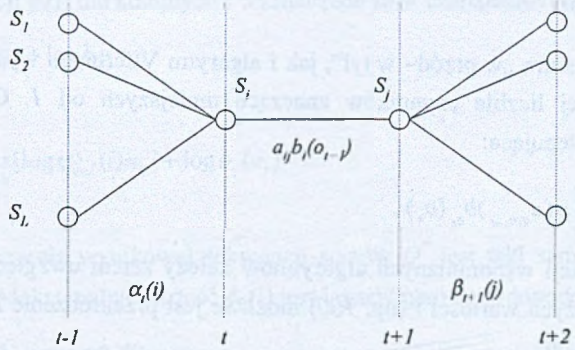
Tak jak pokazano na rys. 5, prawdopodobieństwo $\xi_t(i,j)$ jest prawdopodobieństwem kolejnego zajścia trzech zdarzeń: przejścia układu do stanu S_i w chwilach $1, 2, ..., t$ dla sekwencji obserwacji $o_1, o_2, ..., o_t$, zmiany stanu z S_i na S_j, że w chwilach $t, t+1$, której towarzyszy wygenerowanie obserwacji o_{t+1}, wygenerowaniu sekwencji obserwacji $o_{t+2}, o_{t+3}, ..., o_L$ po wejściu do stanu S_j w chwili $t+1$. Zgodnie z tym można zapisać, że

$$\xi_t(i,j)=P(q_t=S_i,q_{t+1}=S_j\,|\,O,\lambda)=\frac{P(q_t=S_i,q_{t+1}=S_j,O\,|\,\lambda)}{P(O\,|\,\lambda)}=$$

$$=\frac{P(o_1,o_2,...,o_t,q_t=S_i\,|\,\lambda)P(o_{t+1},q_{t+1}=S_{ij}\,|\,q_t=S_i,\lambda)P(o_{t+2},o_{t+3},...,o_L\,|\,q_{t+1}=S_j,\lambda)}{P(O\,|\,\lambda)}=$$

$$=\frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{P(O\,|\,\lambda)}=\frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}, \tag{34}$$

gdzie $\alpha_t(i)$ i $\beta_{t+1}(j)$ są prawdopodobieństwami wystąpienia odpowiednich podciągów obserwacji. Prawdopodobieństwa te zastosowano wcześniej w algorytmie „w przód - w tył".



Rys. 2. Ilustracja sposobu wyznaczania prawdopodobieństwa $\xi_t(i,j)$
Fig. 2. The illustration of calculation the probability $\xi_t(i,j)$

We wcześniejszych rozważaniach zdefiniowano również prawdopodobieństwo $\gamma_t(i)$ jako prawdopodobieństwo zdarzenia polegającego na tym, że układ dany modelem λ znajdzie się w chwili t w stanie S_i dla sekwencji obserwacji O. Wartość tego prawdopodobieństwa można wyznaczyć sumując $\xi_t(i,j)$ po wszystkich $j=1, 2, ..., N$. Ponieważ zdarzenia losowe polegające na tym, że układ znajdzie się w jakimś stanie w chwili t, są zdarzeniami rozłącznymi, zachodzi równość

$$\alpha_t(i)\beta_t(i) = \sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j) . \tag{35}$$

Podstawiając za $\alpha_t(i)$ i $\beta_t(j)$ do wzoru (27) otrzymamy:

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)} = \sum_{j=1}^{N} \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)} = \sum_{j=1}^{N} \xi_t(i,j) . \tag{36}$$

W metodzie Bauma-Welsha do wyznaczania reestymowanych parametrów ukrytego modelu Markowa posłużono się opisanymi prawdopodobieństwami $\xi_t(i,j)$ i $\gamma_t(i)$. Zmodyfikowane parametry modelu dane są następującymi wzorami:

$$\bar{\pi}_i = \gamma_1(i) , \ \bar{a}_{ij} = \frac{\sum_{t=1}^{L-1} \xi_t(i,j)}{\sum_{t=1}^{L-1} \gamma_t(i)} , \ \bar{b}_i(x) = \frac{\sum_{\forall_{1 \le t \le L \wedge o_t = x}} \gamma_t(i)}{\sum_{t=1}^{L} \gamma_t(i)} . \tag{37}$$

3.4. Skalowanie

Zarówno procedura „w przód - w tył", jak i algorytm Viterbiego wymagają wyznaczenia iloczynów o dużej liczbie czynników znacząco mniejszych od *1*. Ogólna postać tych iloczynów jest następująca:

$$\sigma_t(Q) = \prod_{s=1}^{t-1} \rho(a_{q_s q_{s+1}}) b_{q_s}(o_s). \tag{38}$$

W implementacji wspomnianych algorytmów należy zatem uwzględnić fakt, że już dla stosunkowo niedużych wartości t (np. *100*) możliwe jest przekroczenie zakresu zapisu liczb zmiennopozycyjnych.

W przypadku algorytmu „w przód" najprostszym rozwiązaniem pozwalającym uniknąć przekroczenia precyzji zapisu jest pomnożenie $\alpha_t(i)$ przez niezależny od i (tzn. zależny jedynie od t) współczynnik skalujący c_t

$$\hat{\alpha}_t(i) = c_t \alpha_t(i), \tag{39}$$

gdzie

$$c_t = \frac{1}{\sum_{i=1}^{N} \alpha_t(i)}. \tag{40}$$

Wartości $\alpha_t(i)$ oraz $\hat{\alpha}_t(i)$ dla kolejnych $t=1, 2, ..., L$ wyznaczyć można iteracyjnie w następujący sposób:

$$\alpha_{t+1}(i) = \sum_{j=1}^{N} \hat{\alpha}_t(j) a_{ij} b_j(o_{t+1}), \tag{41}$$

$$\hat{\alpha}_{t+1}(i) = \frac{\alpha_{t+1}(i)}{\sum_{i=1}^{N} \alpha_{t+1}(i)} = \frac{\sum_{j=1}^{N} \hat{\alpha}_t(j) a_{ij} b_j(o_{t+1})}{\sum_{i=1}^{N} \sum_{j=1}^{N} \hat{\alpha}_t(j) a_{ij} b_j(o_{t+1})}. \tag{42}$$

W przypadku konieczności skalowania prawdopodobieństwa $\beta_t(i)$ zastosować również można współczynnik c_t, wyznaczając zamiast prawdopodobieństwa $\beta_t(i)$ wartość

$$\hat{\beta}_t(i) = c_t \beta_t(i). \tag{43}$$

Należy nadmienić, że skalowane wartości prawdopodobieństw $\alpha_t(i)$ i $\beta_t(i)$ mogą być zastosowane również do wyznaczenia reestymowanych parametrów modelu ($\overline{\Pi}, \overline{A}, \overline{B}$).

W przypadku algorytmu Viterbiego w celu uniknięcia przekroczenia zakresu iloczyny zastępuje się sumami logarytmów stosując zamiast funkcji $\delta_t(i)$ (28) funkcję:

$$\phi_t(i) = \log\left\{ \max_{q_1, q_2, ..., q_{t-1}} P(q_1, q_2, ..., q_t = S_i, o_1, o_2, ..., o_t \mid \lambda) \right\}. \tag{44}$$

Wartości funkcji $\phi_t(i)$ dla kolejnych t wyznaczane są w następujący sposób:

$$\phi_t(i) = \log \pi_i + \log b_j(o_t) \tag{45}$$

dla $t=1$, zaś

$$\phi_t(j) = \max_{1 \le i \le N}[\log \varphi_{t-1}(i) a_{ij}] + \log b_j(o_t) \tag{46}$$

dla pozostałych t.

Sposób wyznaczenia wynikowej sekwencji stanów Q^* jest taki sam, jak w pierwotnej wersji algorytmu. Maksymalna wartość $\phi_L(i)$ jest logarytmem prawdopodobieństwa $P^*(O|\lambda)$:

$$\log P^*(O \mid \lambda) = \max_{1 \le i \le N} \phi_L(i). \tag{47}$$

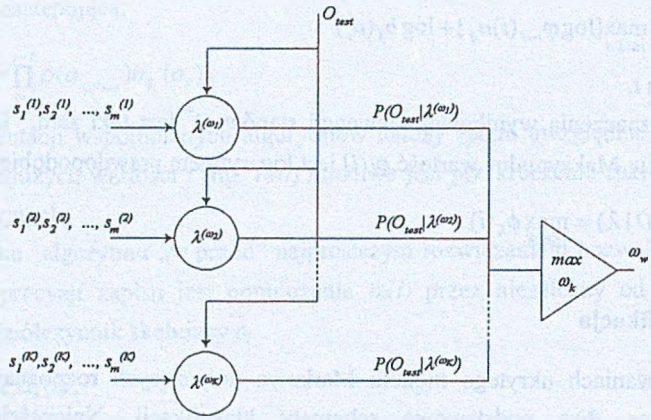
3.5. Klasyfikacja

W zastosowaniach ukrytego modelu Markowa dotyczących rozpoznawania wzorców wskazać można dwa podstawowe schematy klasyfikacji. Najczęściej stosuje się klasyfikatory, w których miarą determinującą wynik jest prawdopodobieństwo $P(O|\lambda)$. Działanie takiego klasyfikatora można najogólniej scharakteryzować w następujący sposób. Niech $\Omega=\{\omega_i: i=1, 2, ..., K\}$ oznacza zbiór rozpatrywanych klas wzorców, niech $S^{(k)}=\{s^{(k)}j : j=1, 2, ..., m\}$ będzie zbiorem próbek uczących należących do klasy ω_k. W fazie uczenia dla każdej klasy $\omega_k \in \Omega$, na podstawie obserwacji uzyskanych z próbek należących do $S^{(k)}$ wyznaczane są parametry osobnego modelu $\lambda^{(\omega_k)}$. Podczas klasyfikacji uwzględnia się prawdopodobieństwo $P(O_{test}|\lambda^{(\omega_k)})$, co w najprostszym przypadku może sprowadzać się do wyboru klasy, dla której wartość tego prawdopodobieństwa jest największa (rys. 3):

$$\omega_w = \arg\max_{\omega_k \in \Omega} P(O_{test} \mid \lambda^{(\omega_k)}). \tag{48}$$

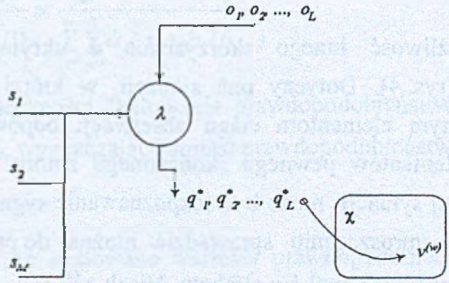
Istnieje również możliwość innego skorzystania z ukrytego modelu Markowa w klasyfikacji wzorców (rys. 4). Dotyczy ona sytuacji, w której zachodzi konieczność przyporządkowania kolejnym elementom ciągu obserwacji odpowiednich etykiet lub - wyrażając się ściślej - elementów pewnego skończonego zbioru. Najlepszym (choć nie jedynym) przykładem takiej sytuacji może być rozpoznawanie sygnału mowy. Zagadnienie to bowiem w znacznym uproszczeniu sprowadzić można do przypisywania kolejnym fragmentom sygnału odpowiednich znaków alfabetu. Niech $v^{(k)} = v_{i_1}, v_{i_2}, ..., v_{i_t}$ będzie ciągiem elementów pewnego zbioru V. Niech $\chi=\{v^{(i)}:i=1,2,...,K\}$ oznacza skończony zbiór takich ciągów. Zadanie opisane powyżej rozumieć można inaczej jako przypisanie ciągowi obserwacji $o_1, o_2, ..., o_L$ odpowiedniego ciągu $v^{(k)} \in \chi$. Rozwiązanie sprowadza się do zdefiniowania modelu, w którym każdemu stanowi przyporządkowuje się jeden z elementów

zbioru V. Prawdopodobieństwa przejść pomiędzy stanami określane są na podstawie ciągów należących do zbioru χ.



Rys. 3. Klasyfikacja z zastosowaniem wartości prawdopodobieństwa $P(O_{test}|\ \lambda^{(\omega_k)})$
Fig. 3. Classification based on the probability $P(O_{test}|\ \lambda^{(\omega_k)})$

Parametry rozkładów $b_i(O)$ ustalane są na podstawie sekwencji obserwacji, które uzyskano ze wzorców uczących, $s_1, s_2, \ldots, s_M$. Klasyfikacja polega na wyznaczeniu dla wejściowego ciągu obserwacji najbardziej prawdopodobnej sekwencji stanów $q^*_1, q^*_2, \ldots, q^*_L$. Rozpatrywanemu ciągowi obserwacji przypisuje się odpowiadający otrzymanej sekwencji stanów ciąg $v^{(w)} = v_{l_1}, v_{l_2}, \ldots, v_{l_L}$, jeżeli $v^{(k)} \in \chi$.



Rys. 4. Klasyfikator wykorzystujący najbardziej prawdopodobną sekwencję stanów
Fig. 4. Classification based on the most probable sequence of states

4. Wybrane przykłady zastosowań ukrytego modelu Markowa w rozpoznawaniu wzorców

Jak już wspomniano wcześniej, najbardziej rozpowszechnionym obszarem zastosowań ukrytego modelu Markowa jest analiza i rozpoznawanie sygnałów mowy. W tym zakresie modele Markowa znajdują zastosowanie w różnych modułach funkcjonalnych opracowywanych systemów. Zastosowania te dotyczą m.in. analizy fonemów, rozpoznawania głosek i słów, a także analizy leksykalnej.

Jednym z obszarów zastosowań ukrytego modelu Markowa jest również analiza pisma odręcznego *on-line*. W analizie *on-line* stosowane są specjalne elektroniczne rejestratory, za pomocą których odczytuje się odpowiednie przebiegi czasowe. Pozwala to na badanie dynamicznych cech pisma, takich jak: kolejność pociągnięć, siła nacisku pisaka. Parametry modeli wyznaczane są na podstawie przebiegów czasowych odczytanych z rejestratora. Zastosowania ukrytego modelu Markowa w tym zakresie dotyczą rozpoznawania pisma, identyfikacji osób oraz weryfikacji podpisów.

Ukryty model Markowa znalazł także zastosowanie w wizji komputerowej i w rozpoznawaniu obrazów. W aplikacjach dotyczących tej dziedziny często pojawia się problem polegający na tym, że obraz nie jest sam w sobie przebiegiem czasowym. Konieczne jest zatem uzyskanie odpowiedniej reprezentacji obrazu, będącej uporządkowanym ciągiem obserwacji. W dalszej części niniejszego opracowania zamieszczono krótki opis wybranych metod rozpoznawania wzorców, w których zastosowano ukryty model Markowa.

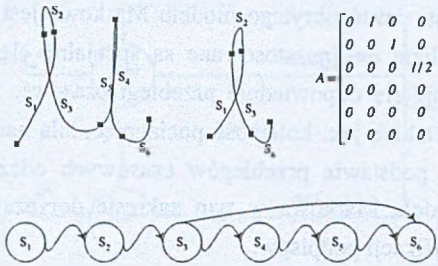
4.1. Rozpoznawanie pisma *on-line*

Prace dotyczące zastosowań ukrytego modelu Markowa do rozpoznawania pisma odręcznego metodami *on-line* są przeważnie bezpośrednim przeniesieniem doświadczeń zdobytych w badaniach nad zastosowaniami z zakresu rozpoznawania sygnałów mowy.

Janying Hu i Michael K. Brown [9], zajmujący się wcześniej systemem analizy mowy, opracowali metodę rozpoznawania pisma odręcznego *on-line* opartą na markowowskich modelach liter alfabetu. W metodzie tej, w fazie uczenia analizowane są przebiegi czasowe określające położenie pisaka na tabliczce rejestratora. Przebiegi pochodzące z odseparowanych liter poddawane są segmentacji. Na podstawie segmentów wyodrębnionych z poszczególnych liter tworzony jest jeden model dla całego alfabetu. Wartości elementów macierzy przejść A określane są na podstawie kolejności występowania segmentów w literach. Wartości parametrów rozkładów obserwacji $b_i(O)$ wyznaczane są na podstawie odpowiednio zdefiniowanych cech charakteryzujących przebiegi odczytywane z rejestratora. Przejście ze stanu do stanu wiąże się z „wygenerowaniem" fragmentu sygnału,

który odpowiada kolejnemu segmentowi pisma. Na rysunku poniżej pokazano przykładowy model uwzględniający dwie litery alfabetu: „U" i „l".

W omawianej metodzie analizuje się jeden po drugim, kolejne fragmenty pisma. Klasyfikator w pierwszej kolejności wyznacza optymalną sekwencję stanów. Na podstawie wynikowej sekwencji stanów odtwarza się ciągi liter, które następnie analizowane są odpowiednio pod względem poprawności gramatycznej. Efektywność rozpoznawania ustalona na podstawie badań przeprowadzonych na ok. 4000 słów zapisanych przez 18 osób wyniosła 94,5%.

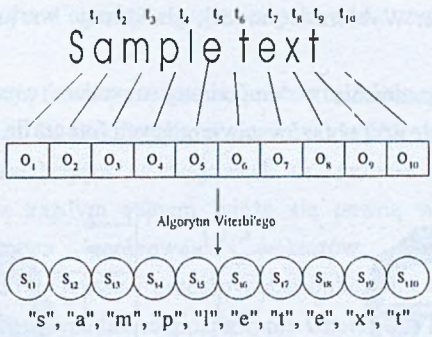


Rys. 5. Przykładowy model dla dwóch liter alfabetu (wynik segmentacji, graf przejść, macierz przejść)

Fig. 5. The example of the model of two letters (the result of segmentation, the transition graph and the transition matrix)

4.2. Rozpoznawanie pisma *off-line*

Kolejny przykład dotyczy zastosowania ukrytego modelu Markowa do rozpoznawania pisma odręcznego na podstawie obrazu [10] (analiza *off-line*). W przykładzie tym każdej literze alfabetu przyporządkowuje się jeden stan modelu. Dla języka angielskiego model ma 26 stanów ($N=26$). Prawdopodobieństwa przejść oraz prawdopodobieństwa początkowe określono na podstawie słownika. Oznacza to, że prawdopodobieństwa początkowe π_i są równe liczbie wyrazów słownika zaczynających się na literę reprezentowaną przez stan S_i podzielonej przez całkowity rozmiar słownika. Prawdopodobieństwa przejść pomiędzy stanami a_{ij} wynikają z częstości występowania w wyrazach słownika określonych par liter reprezentowanych przez stany S_i i S_j.



Rys. 6. Rozpoznawanie znaków pisma odręcznego *off-line*
Fig. 6. *Off-line* handwriting recognition

W celu zdefiniowania obserwacji autorzy określili wektor 15 cech. Wzięto pod uwagę takie cechy, jak: momenty geometryczne, liczbę pętli, liczbę węzłów, liczbę wyizolowanych punktów, stosunek szerokości do wysokości itp. Rozkłady obserwacji określono dla każdego stanu na podstawie cech wyznaczonych dla próbek uczących..

Podczas rozpoznawania (rys. 6) wejściowy ciąg obserwacji stanowią wektory cech opisujące kolejne litery wytupujące w rozpoznawanym fragmencie tekstu $O = \{O_1, O_2,..., O_M\}$. Stosując algorytm Viterbiego ustala się najbardziej prawdopodobną sekwencję stanów dla wejściowej sekwencji obserwacji. Wynikowej sekwencji stanów odpowiada odpowiedni ciąg liter alfabetu.

Skuteczność rozpoznawania wyznaczona w eksperymencie przeprowadzonym na próbkach pisma dwudziestu osób oraz dla słownika zawierającego 90 słów wyniosła 92,5%.

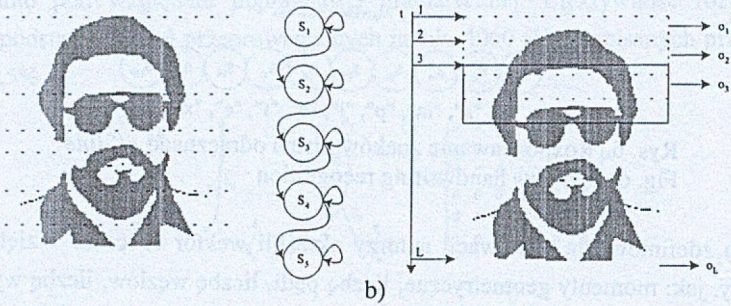
4.3. Identyfikacja twarzy

Ciekawym przykładem zastosowania ukrytego modelu Markowa w wizji komputerowej jest opracowana przez F.Samarię [11] metoda automatycznego rozpoznawania twarzy. Autor zaproponował model o pięciu stanach (rys. 7a). Każdy stan reprezentuje inną część twarzy (czoło, oczy, nos, usta, broda). Obserwacje wyznaczane są na podstawie wartości poziomów szarości punktów zawartych w ramce obejmującej fragment obrazu (rys. 7b). Sekwencję obserwacji uzyskuje się przesuwając ramkę wzdłuż analizowanego obrazu (z góry w dół). Za każdym razem ramka przesuwna jest o stałą liczbę punktów. Zmiana stanu następuje wtedy, gdy ramka przemieści się w region zawierający inną część twarzy.

W systemie klasyfikującym dla każdej osoby, na podstawie kilku próbek uczących, określa się parametry odrębnego modelu. Klasyfikacja polega na wyznaczeniu dla sekwencji obserwacji uzyskanej z próbki testowej prawdopodobieństwa wygenerowania tej sekwencji

przez wszystkie modele. Wybiera się model, dla którego wartość prawdopodobieństwa jest największa.

Skuteczność rozpoznawania w omówionej metodzie oszacowano na 90%. Badania przeprowadzono na bazie 400 obrazów zawierających fotografie twarzy 40 osób.

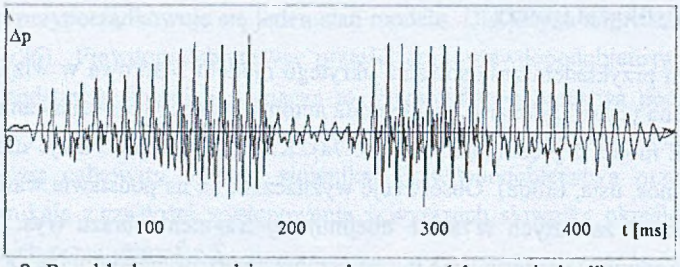


Rys.7. Rozpoznawanie twarzy: a) topologia modelu, b) pozyskiwanie sekwencji obserwacji

Fig. 7. Face identification: a) the proposed topology, b) obtaining the observation sequence

4.4. Rozpoznawanie mowy

Ukryte modele Markowa są szeroko stosowane w rozpoznawaniu mowy. Sygnał mowy dany jest w postaci ciągu próbek opisujących odchylenie chwilowego ciśnienia powietrza od średniej wartości (rys. 8). Celem rozpoznania jest uzyskanie transkrypcji ortograficznej tak zapisanej wypowiedzi.



Rys. 8. Przykładowy przebieg sygnału mowy (słowo „jeden")

Fig. 8. An example of a speech waveform (the Polish word „jeden")

Obserwacjami nie są bezpośrednio wartości próbek, ale wyznaczone na podstawie krótkich fragmentów sygnału mowy parametry, którymi najczęściej są tzw. współczynniki *mel*-cepstrum. Stosowane są też banki filtrów, częstotliwości formantów i inne cechy. Parametry opisujące jeden fragment sygnału, nazywany też okienkiem, tworzą wektor o

liczbie wymiarów zależnej od liczby wyznaczanych cech, zwykle od kilkunastu do kilkudziesięciu.

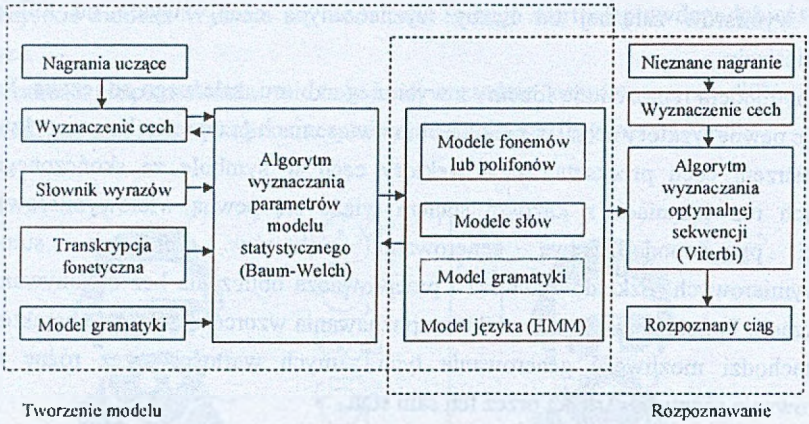
Stany modelu reprezentują fonemy z wybranego zbioru, zależnego od języka. Każdy stan generuje pewne wektory cech. W niektórych rozwiązaniach przeprowadzana jest kwantyzacja w przestrzeni cech przekształcająca wektory cech w symbole ze skończonego zbioru. W innych rozwiązaniach z każdym stanem wiąże się pewną wielowymiarową funkcję gęstości prawdopodobieństwa generowania wektorów cech, np. sumę kilku wielowymiarowych rozkładów Gaussa i przeprowadza obliczenia bez dokonywania jawnej kwantyzacji. Podobnie jak w przypadku rozpoznawania wzorców o innym charakterze, także tutaj zachodzi możliwość generowania tych samych wartości przez różne stany jak i generowania różnych wartości przez ten sam stan.

W bardziej rozbudowanych modelach zbiór fonemów jest uzupełniany o wersje fonemów zależne od kontekstu (sąsiedztwa), w jakim występują. Mówi się o bifonach, trifonach i ogólnie polifonach. Liczba stanów jest wtedy znacznie większa niż dla modelu z fonemami bezkontekstowymi.

Zależnie od przeznaczenia systemu modele mogą opisywać artykulację pojedynczych słów lub sekwencji słów, np. całych zdań. W drugim przypadku jest możliwe włączenie do modelu informacji o składni języka, ograniczającej dopuszczalne sekwencje słów do zgodnych z jego składnią i gramatyką. Ponieważ tutaj również możliwe jest stosowanie modeli Markowa ze stanami będącymi całymi słowami lub klasami słów (np. liczebnik, dzień_tygodnia itp.), można rozważać budowę modelu na wielu poziomach abstrakcji, z najniższym poziomem odpowiadającym fonemom lub ich elementom.

Ze względu na popularność problemu rozpoznawania mowy powstało wiele narzędzi wspomagających tworzenie modeli oraz rozpoznawanie mowy. Jednym z popularniejszych, stosowanych w badaniach, jest zestaw Hidden Markov Model Toolkit (HTK), opracowany na Wydziale Technicznym Uniwersytetu Cambridge [12], potem przejęty przez firmę Entropic i Microsoft. Pozwala on m.in. na definiowanie modelu wypowiedzi na wielu poziomach (subfonemów, słów, zdań) i rozpoznawanie wypowiedzi.

Rysunek 9 przedstawia ogólny schemat blokowy systemu tworzącego modele i rozpoznającego mowę. Współczesne systemy rozpoznawania mowy osiągają 98% trafności rozpoznawania.



Rys. 9. Schemat blokowy systemu tworzącego modele i rozpoznającego mowę
Fig. 9. Block diagram of a speech modeling and recognition system

5. Podsumowanie

Jak już wspomniano, cechą decydującą o możliwości zastosowania ukrytego modelu Markowa w rozpoznawaniu wzorców jest oddzielenie sygnału reprezentującego wzorzec od procesów faktycznie zachodzących w układzie generującym sygnał. Parametry modelu zawierają informacje o zachowaniu się układu i właściwościach statystycznych wzorców uczących. Podczas klasyfikacji nie dokonuje się bezpośredniego porównania wzorców testowych z wzorcami uczącymi, a jedynie wyznacza się prawdopodobieństwo wygenerowania wzorca testowego przez model, którego parametry określono na podstawie wzorców uczących. W przypadku zastosowania algorytmu Viterbiego prowadzi to do odtworzenia procesów losowych związanych z zachowaniem się układu (najbardziej prawdopodobna sekwencja stanów układu dla danej sekwencji obserwacji).

Obszar zastosowań ukrytych modeli Markowa w zakresie rozpoznawania wzorców cały czas się rozszerza - od najbardziej rozpowszechnionych systemów rozpoznawania mowy, do wizji komputerowej i analizy statycznych obrazów komputerowych. Należy jednak podkreślić, że możliwości zastosowania metod opartych na ukrytym modelu Markowa podlegają pewnym istotnym ograniczeniom. Podstawowe ograniczenie związane jest z założeniem, iż poszczególne elementy ciągu obserwacji są od siebie niezależne. Kolejne ograniczenie wynika z założeń Markowa, że prawdopodobieństwo przejścia do danego stanu w chwili t zależy jedynie od stanu układu w chwili $t-1$. Innym istotnym faktem, na który należy zwrócić uwagę, jest ograniczona dokładność opisu cech statystycznych sygnału w przypadku obserwacji o ciągłym charakterze (rozkład obserwacji w postaci funkcji Gaussa,

autoregresja). W przypadku większości zastosowań całkowite spełnienie wymienionych ograniczeń może podlegać kwestii (np. zależności pomiędzy kolejnymi fragmentami sygnału mowy rozciągają się z pewnością na większy fragment tegoż sygnału), jakkolwiek niedokładności wynikające z częściowego tylko spełnienia założeń ukrytego modelu Markowa usprawiedliwić można znacznym uproszczeniem rozważań i obliczeń związanych z analizą procesów losowych, tym bardziej, że opracowane systemy działają poprawnie, a skuteczność klasyfikacji jest zadowalająca.

LITERATURA

1. Elliot R. J., Aggoun L., Moore J. B.: Hidden Markov Models. Springer-Verlag, New York 1997.
2. Rabiner L. R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of IEEE, vol. 77, pp. 257-286, 1989.
3. Viterbi A. J.: Error bounds for convolutional codes and asymptotically optimal decoding algorithm. IEEE Transactions Information Theory, vol. IT-13, pp. 260-269, 1967.
4. Forney G. D.: The Viterbi Algorithm. Proc. IEEE, vol. 61, pp. 268-278, 1973.
5. Rayan M. S., Nudd G. R.: The Viterbi Algorithm. Warwick Research Report, RR 238, University of Warwick, 1993.
6. Baum L. E., Petrie T., Soules G., Weiss N.: A maximisation technique occurring in the statistical analysis of probabilistic functions of Markov Chains. Ann. Math. Stat., vol. 41, no. 1, pp. 164-171, 1970.
7. Baum L. E., Sell G. R.: An inequality and associated maximisation technique in statistical estimation for probabilistic functions of Markov processes. Inequalities, vol. 3, pp. 1-8, 1972.
8. Baum L. E., Sell G. R.: Growth transformations for functions on manifolds. Pacific Journal of Mathematics, vol. 27, pp. 211-227, 1968.
9. Hu J., Brown M. K., Turin W.: HMM based on-ine handwriting recognition. Transactions On Pattern Recognition And Machine Intelligence, vol. 18, nr 10, pp. 39-45, 1996.
10. Kundu A., He Y.: Recognition of handwriten words: first and second order hidden Markov model based approach. Proceedings of IEEE, vol. 4 pp. 457-462, 1988.
11. Samaria F., Fallside F.: Face identification and feature extraction using hidden Markov models. Image Processing: Theory and Applications, Elsevier, 1993.
12. Young S., Kersha D., Odell J., Ollason D., Vatchev V.: The HTK Book. Microsoft Corporation, 2000.

Recenzent: Prof. dr inż. hab. Tadeusz Czachórski

Wpłynęło do redakcji 30 czerwca 2002

Abstract

As mentioned before, the feature making possible recognizing patterns with Hidden Markov Models is the separation of the signal describing a pattern from the process occurring in the system that generates the signal. The parameters of the model contain information about the behavior of the system and statistical properties of the training patterns. The classification process does not compare directly the training data with patterns to be recognized, but only evaluates the probability of generation of such patterns by a model with parameters estimated during a training phase. The Viterbi algorithm reconstructs random processes related to the behavior of the system (the most probable sequence of states for a given sequence of observations).

The scope of HMM applications in pattern recognition is still growing. From the most popular speech recognition to computer vision and analyses of static computer images. However, possible adaptations of methods based on HMMs have some restrictions. The basic limitation is the assumption, that individual elements of the observation sequence are independent. Another limitation is a result of Markov's assumption, that the probability of a transition to a given state at the time t depends only on the state at the time t-1. Finally, statistical properties of the signal are described with a limited precision in case of continuos observations (distribution of observation given as Gaussian distribution, autoregression). In most applications these conditions are not satisfied. For example, dependencies in speech signal cover more than only two adjacent frames of the signal. But all these simplifications and inaccuracies make the models and calculations less complex or even possible, and systems based on them satisfy the classification's accuracy requirements.