Ukryte modele Markowa jako metoda eksploracji danych tekstowych

M. MAZUREK
e-mail: marcin.mazurek@wat.edu.pl

Instytut Systemów Informatycznych
Wydział Cybernetyki WAT
ul. S. Kaliskiego 2, 00-908 Warszawa

W eksploracji danych tekstowych z dużym powodzeniem stosuje się probabilistyczne modele dokumentów. W artykule przedstawiony został jeden z podstawowych, dla tej dziedziny informatyki, sposobów reprezentacji dokumentu za pomocą ukrytych modeli Markowa. Przedstawiono definicję ukrytego modelu Markowa oraz sposób wyznaczenia podstawowych wielkości związanych z wykorzystaniem tego modelu, takich jak prawdopodobieństwo wystąpienia obserwowanej sekwencji symboli (słów), wyszukanie najbardziej prawdopodobnej sekwencji stanów procesu, czy też formuły reestymacji parametrów modelu używane w procesie uczenia modelu.

Słowa kluczowe: eksploracja danych tekstowych, ukryte modele Markowa, ekstrakcja informacji

1. Wprowadzenie

Nieustrukturalizowane bądź semistrukturalne dane tekstowe stanowią część dostępnych zasobów informacji. Dokumenty tekstowe są naturalnym, zrozumiałym dla każdego formatem zapisu komunikatów, stąd mimo istotnego wyzwania, jakie stanowi ich przetwarzanie przez systemy komputerowe, należy spodziewać się dalszego przyrostu danych zapisywanych w takiej postaci w tempie nie wolniejszym niż obserwuje się obecnie.

Dokumenty te mogą stanowić istotne źródło informacji w systemach wspomagania decyzji. Przykładem wykorzystania danych tekstowych w zwiększaniu konkurencyjności organizacji może być identyfikacja tendencji i trendów w prasie. Analiza doniesień prasowych może być istotnym czynnikiem w oszacowaniu wpływu mediów na wiarygodność instytucji zaufania publicznego, w szczególności na runku kapitałowym. Podobnym zagadnieniem jest pozyskiwanie opinii o produkcie [2].

Niezbędnym warunkiem wykorzystania danych tekstowych jest ekstrakcja informacji zawartych w dokumentach. Jest ona zwykle poprzedzona konwersją dokumentu na model wektorowy, wyodrębnieniem rdzeni słów, usunięciem słów nieznaczących, czy też identyfikacją fraz. Ekstrakcja informacji polega na wypełnieniu pewnej zaprojektowanej struktury danych wartościami wyznaczonymi w oparciu o dokument tekstowy.

Obok regułowych systemów ekstrakcji informacji, powszechnie stosowane są modele probabilistyczne dokumentu tekstowego, w których przyjmuje się, że obserwowany ciąg fraz jest realizacją zmiennych losowych.

Artykuł przedstawia założenia jednego z podstawowych dla dziedziny eksploracji dokumentów tekstowych modeli dokumentu tekstowego – ukrytego modelu Markowa.

2. Ukryte modele Markowa

Ukryte modele Markowa różnią się od klasycznych łańcuchów Markowa brakiem możliwości bezpośredniej obserwacji stanu, w jakim przebywa proces. Zamiast tego, obserwujemy realizację probabilistycznej funkcji określonej na zbiorze stanów procesu, której wartościami są symbole pewnego alfabetu. Poniżej przedstawiona została formalna definicja ukrytego modelu Markowa pierwszego rzędu [1].

Definicja

Ukrytym modelem Markowa λ pierwszego rzędu nazywamy piątkę:

$$\lambda = \langle Q, \Sigma, \Pi, \mathrm{A}, \mathrm{B} \rangle \tag{1}$$

gdzie:

$Q = \{q_1, q_2, \ldots, q_N\}$ jest zbiorem stanów procesu,

N jest liczbą stanów procesu,

$\Sigma = \{\delta_1, \delta_2, \ldots, \delta_M\}$ jest zbiorem symboli (alfabet),

M jest liczbą możliwych do zaobserwowania symboli,

$\Pi : Q \to \langle 0,1 \rangle$ oznacza początkowy rozkład prawdopodobieństwa, spełniający warunek:

$$\sum_{q \in Q} \pi(q) = 1 \qquad (2)$$

$A : Q \times Q \to \langle 0,1 \rangle$ oznacza macierz prawdopodobieństw przejść pomiędzy stanami, której elementy $a(q,q')$ spełniają warunek:

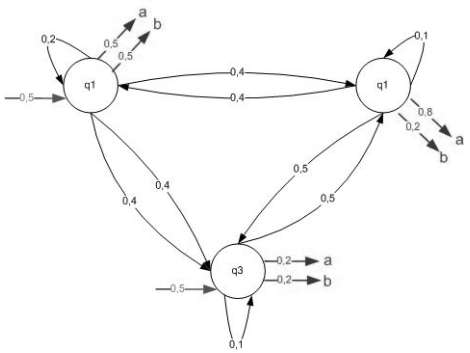
$$\sum_{q' \in Q} a(q,q') = 1 \quad \text{dla} \quad q \in Q \qquad (3)$$

$B : Q \times \Sigma \to \langle 0,1 \rangle$ oznacza macierz prawdopodobieństw emisji następujących po sobie symboli, której elementy $b(q,\sigma)$ spełniają warunek:

$$\sum_{\sigma \in \Sigma} b(q,\sigma) = 1 \quad \text{dla} \quad q \in Q \qquad (4)$$

Przykład

Przykład ukrytego modelu Markowa z trzema stanami, w którym emitowane są dwa symbole „a" oraz „b" przedstawiony został na rysunku 1.



Rys. 1. Przykład ukrytego modelu Markowa z trzema stanami oraz alfabetem składającym się z dwóch symboli

Ukryty model Markowa może zostać wykorzystany jako zarówno generator symboli, jak i narzędzie do odtworzenia sekwencji stanów na podstawie obserwowanych symboli.

W eksploracji danych tekstowych obserwowanymi symbolami są słowa, frazy bądź całe sformułowania. W zależności od postawionego problemu, stanami mogą być elementy struktury dokumentu, jeżeli zadanie polega na rozpoznaniu jego wewnętrznej struktury (rozpoznanie części opisującej dane teleadresowe, streszczenie, część zasadniczą dokumentu). Inny przykład zbioru stanów to zbiór części mowy odpowiadających poszczególnym słowom, jeżeli zadanie polega na rozbiorze gramatycznym zdania.

Dla wykorzystania ukrytych modeli Markowa w analizie danych tekstowych kluczowe jest znalezienie wydajnych procedur obliczeniowych, pozwalających odpowiedzieć na sformułowane trzy podstawowe problemy [1]:

- problem wyznaczenia prawdopodobieństwa zaobserwowania sekwencji symboli, przy założeniu, że znana jest definicja ukrytego modelu Markowa
- problem wyznaczenia sekwencji stanów, w których przebywał proces dla obserwowanej sekwencji symboli, przy założeniu, że znana jest definicja ukrytego modelu Markowa
- problem wyznaczenia elementów definicji ukrytego modelu Markowa, przy założeniu, że dysponujemy sekwencją obserwowanych symboli.

Problemy te mogą być efektywnie (w czasie wielomianowym) rozwiązywane przy zastosowaniu techniki dynamicznego programowania, w której zapamiętywane są pośrednie wyniki obliczeń. Dalej zaprezentowane zostały podstawowe założenia tych algorytmów.

3. Prawdopodobieństwo wystąpienia obserwowanej sekwencji symboli

Problem

Dla zadanego ukrytego modelu Markowa λ wyznaczyć prawdopodobieństwo wygenerowania przez niego *T*-elementowej sekwencji symboli $O = o_1 o_2 \ldots o_T$, gdzie $o_i \in \Sigma$

Zagadnienie to może zostać rozwiązane za pomocą jednego z dwóch, równoważnych modeli obliczeń rekurencyjnych (algorytm prefiksowo-sufiksowy).

Oznaczenia

Załóżmy, że sekwencja stanów procesu w kolejnych krokach miała postać $S = s_1 s_2 \ldots s_T$, gdzie $s_i \in Q$.

Algorytm do przodu

Niech $\alpha_t(q)$ oznacza prawdopodobieństwo wygenerowania sekwencji symboli $O = o_1 o_2 \ldots o_t$, przy założeniu, że w kroku t proces znajduje się w stanie *q*.

$$\alpha_t(q) = P(o_1 o_2 \ldots o_t, s_t = q \mid \lambda) \qquad (5)$$

Wartości $\alpha_t(q)$ mogą zostać wyznaczone rekurencyjnie. Dla pierwszego kroku $t = 1$:

$$\alpha_1(q) = \pi(q) \cdot b(q, o_1) \qquad (6)$$

oraz dla kolejnych kroków $t = \overline{1, T-1}, q \in Q$:

$$\alpha_{t+1}(q) = \left(\sum_{q' \in Q} \alpha_t(q') \cdot a(q', q) \right) \cdot b(q, o_{t+1}) \qquad (7)$$

Rozwiązaniem problemu jest wartość:

$$P(O \mid \lambda) = \sum_{q \in Q} \alpha_T(q) \tag{8}$$

Algorytm wstecz

Analogicznie do procedury wyliczania wartości $\alpha_t(q)$ można zaproponować kolejną procedurę wyliczania wartości:

$$\beta_t(q) = P(o_t o_{t+1} \ldots o_T, s_t = q \mid \lambda) \tag{9}$$

czyli prawdopodobieństwa wygenerowania sekwencji symboli $O = o_{t+1} o_{t+2} \ldots o_T$, przy założeniu, że w kroku *t* proces przebywa w stanie *q*. Wartości $\beta_t(q)$ mogą zostać wyliczone rekurencyjnie. Dla kroku *T* oraz $q \in Q$ mamy:

$$\beta_T(q) = 1 \tag{10}$$

Dla wcześniejszych kroków zachodzi:

$$\beta_{t-1}(q) = \sum_{q' \in Q} \beta_t(q') \cdot a(q', q) \cdot b(q', o_t) \tag{11}$$

Do rozwiązania sformułowanego problemu wystarczy wykorzystanie jednej z tych procedur (procedury do przodu), jednak wyznaczone wartości $\beta_t(q)$ będą wykorzystywane w rozwiązaniu kolejnych problemów.

4. Wyznaczenie sekwencji stanów na podstawie obserwowanej sekwencji symboli

Problem

Dla zadanego ukrytego modelu Markowa wyznaczyć najbardziej prawdopodobną sekwencję stanów procesu $S^* = s_1^* s_2^* \ldots s_T^*$, $s_i^* \in Q$, przy założeniu, że obserwujemy wygenerowaną przez niego *T*-elementową sekwencję symboli $O = o_1 o_2 \ldots o_T$, gdzie $o_i \in \Sigma$.

Oznaczenia

Niech $\delta_t(q)$ oznacza prawdopodobieństwo przejścia procesu przez najbardziej prawdopodobną sekwencję stanów w krokach od pierwszego do to *t*-1 oraz kończącą się w kroku *t* na stanie *q:*

$$\delta_t(q) = \max_{s_1, s_2, \ldots s_{t-1} \in Q} P(s_1 s_2 \ldots s_{t-1} q, O_t \mid \lambda) \tag{12}$$

Jeżeli przez $\gamma_t(q)$ oznaczymy taką sekwencję:

$$\gamma_t(q) = s_1^* s_2^* \ldots s_{t-1}^* q \,,\; s_i \in Q$$

to:

$$\delta_t(q) = P(\gamma_t(q), O_t \mid \lambda) \tag{13}$$

W kolejnym kroku *t*+1 wartość $\delta_{t+1}(q)$ może zostać wyliczona z zależności:

$$\delta_{t+1}(q) = \left(\max_{q' \in Q} (\delta_t(q') \cdot a(q', q)) \right) \cdot b(q, o_{t+1}) \tag{14}$$

Aby po zakończeniu procedury obliczeniowej odtworzyć sekwencję stanów, dla każdego stanu *q* w kroku *t* należy zapamiętać poprzedni stan, to znaczy ten, dla którego wartość wyrażenia $(\delta_{t-1}(q') \cdot a(q', q))$ była maksymalna:

$$\psi_t(q) = \arg\max_{q \in Q'} (\delta_{t-1}(q') \cdot a(q', q)) \tag{15}$$

Procedura obliczeniowa

Algorytm ma postać następującej procedury obliczeniowej:

Dla $t = 1$ oraz dla każdego $q \in Q$:

$$\gamma_1(q) = q \tag{16}$$

$$\delta_1(q) = \pi(q) \cdot b(q, o_1) \tag{17}$$

$$\psi_1(q) = 0 \tag{18}$$

Dla kolejnych kroków $t = \overline{2, T}$ oraz wszystkich $q \in Q$ mamy:

$$\gamma_{t+1}(q) = \gamma_t(q') q \tag{19}$$

$$\delta_{t+1}(q) = \max_{q' \in Q} (\delta_t(q') \cdot a(q', q) \cdot b(q, o_{t+1})) \tag{20}$$

$$\psi_{t+1}(q) = \arg\max_{q' \in Q} (\delta_t(q') \cdot a(q', q)) \tag{21}$$

Warunkiem stopu algorytmu jest osiągnięcie kroku $t = T$, w którym możemy wyznaczyć prawdopodobieństwo najbardziej wiarygodnej sekwencji oraz ostatni element sekwencji, czyli stan, w jakim zakończył się proces:

$$P(S^* \mid O, \lambda) = \max_{q \in Q} (\delta_T(q)) \tag{22}$$

$$s_t^* = \arg\max_{q \in Q} (\delta_T(q)) \tag{23}$$

Aby odtworzyć wcześniejsze stany procesu, po zakończeniu pełnej iteracji algorytmu dla stanu z kroku *t*+1 odczytujemy zapamiętany poprzednik, z którego przejście było najbardziej prawdopodobne, czyli dla $t = \overline{1, T-1}$ mamy:

$$s_t^* = \psi_{t+1}(s_{t+1}^*) \tag{24}$$

Przedstawiony algorytm wyszukuje sekwencję stanów, która jest najbardziej prawdopodobna. Wyznaczone w ten sposób stany dla kroku *t* procesu mogą się różnić od

takich, które zostały wyznaczone jako najbardziej prawdopodobne dla kroku t bez uwzględniania sekwencji, czyli:

$$\hat{s}_t^* = \arg\max_{q \in Q} P(s_t = q \mid O, \lambda) \quad (25)$$

W przypadku, gdy wyszukiwane są niezależnie dla każdego kroku procesu najbardziej prawdopodobne stany, uzyskane wartości nie muszą tworzyć sekwencji najbardziej prawdopodobnej. W niektórych przypadkach, gdy macierz przejść A zawiera elementy zerowe, tak uzyskana sekwencja może nie być nawet realizowalna [1].

5. Reestymacja parametrów modelu

Problem

Dla zadanego ukrytego modelu Markowa $\lambda = \langle Q, \Sigma, \Pi, \mathrm{A}, \mathrm{B} \rangle$ oraz sekwencji symboli $O = o_1 o_2 \ldots o_T$, gdzie $o_i \in \Sigma$, wyznaczyć parametry ukrytego modelu Markowa $\lambda' = \langle Q, \Sigma, \Pi', \mathrm{A}', \mathrm{B}' \rangle$ takiego, aby $P(O \mid \lambda') \geq P(O \mid \lambda)$.

Postawiony problem jest problemem uczenia ukrytego modelu Markowa w oparciu o dane z ciągu uczącego

Problem wyznaczenia parametrów modelu Markowa rozwiązywany jest za pomocą formuł reestymacji modelu Bauma−Welsha, których zastosowanie prowadzi do znalezienia maksimum lokalnego [1].

Oznaczenia

Niech

$$\mu_t(q) = P(s_t = q \mid O, \lambda) \quad (26)$$

oznacza prawdopodobieństwo tego, że proces jest w stanie q w kroku t pod warunkiem, że zaobserwowano sekwencję symboli O. Uwzględniając zależności (5) oraz (9), można zapisać:

$$\mu_t(q) = \frac{\alpha_t(q) \cdot \beta_t(q)}{P(O \mid \lambda)} \quad (27)$$

Podstawiając za $P(O \setminus \lambda)$, otrzymuje się:

$$\mu_t(q) = \frac{\alpha_t(q) \cdot \beta_t(q)}{\sum_{r \in Q} \alpha_t(r) \cdot \beta_t(r)} \quad (28)$$

Niech

$$\phi_t(q', q) = P(s_t = q', s_{t+1} = q \mid O, \lambda) \quad (29)$$

oznacza prawdopodobieństwo przejścia ze stanu q' do q w kroku t pod warunkiem zaobserwowania sekwencji O.

$$\phi_t(q', q) = \frac{\alpha_t(q') \cdot a(q', q) \cdot b(q, o_{t+1}) \cdot \beta_{t+1}(q)}{P(O \mid \lambda)} \quad (30)$$

Zachodzi:

$$\mu_t(q) = \sum_{q' \in Q} \phi_t(q, q') \quad (31)$$

Suma $\sum_{t=1}^{T} \mu_t(q)$ jest oczekiwaną liczbą kroków procesu, w których przebywa on w stanie q.

Podobnie suma $\sum_{t=1}^{T-1} \mu_t(q', q)$ jest oczekiwaną liczbą przejść ze stanu q' do q.

Procedura obliczeniowa

Formuły aktualizacji parametrów modelu zaproponowane przez Bauma mają postać:

1. Inicjalne prawdopodobieństwo dla każdego stanu może zostać oszacowane jako oczekiwana częstość przebywania procesu w stanie q w kroku $t = 1$.
$$\pi(q) := \mu_1(q) \quad (32)$$
2. Oszacowanie prawdopodobieństwa przejścia ze stanu q' do q może zostać wyznaczone jako iloraz oczekiwanej liczby przejść procesu ze stanu q' do q do oczekiwanej liczby wyjść procesu ze stanu q.
$$a'(q', q) := \frac{\sum_{t=1}^{T-1} \phi_t(q', q)}{\sum_{t=1}^{T-1} \mu_t(q')} \quad (33)$$
3. Prawdopodobieństwo wygenerowania symbolu w stanie q może zostać oszacowane jako iloraz oczekiwanej liczby kroków, w których proces przebywał w stanie q i wygenerował symbol do oczekiwanej liczby kroków, w jakich proces przebywał w tym stanie:
$$b'(q, \sigma) := \frac{\sum_{t \mid o_t = \sigma}^{T-1} \mu_t(q)}{\sum_{t=1}^{T} \mu_t(q)} \quad (34)$$

6. Zastosowanie ukrytych modeli Markowa

Zakres zastosowania ukrytych modeli Markowa jest bardzo szeroki. Pierwsze ich wykorzystanie w przetwarzaniu danych tekstowych związane było z rozpoznawaniem mowy [1].

Ukryte modele Markowa mogą być stosowane do rozpoznawania struktury dokumentu, rozpoznawania części mowy [5].

Najczęściej przyjmuje się, że emitowanymi symbolami są słowa lub frazy (ciąg słów o znaczeniu wynikającym z wszystkich tych słów, np. Wojskowa Akademia Techniczna) [6]. Uczenie ukrytego modelu Markowa polega na wyznaczeniu takich parametrów rozkładów losowych poszczególnych słów w stanach, aby struktura nieznanego dokumentu tekstowego została właściwie odczytana w oparciu o sekwencję stanów. Jeżeli model zostanie zastosowany do rozpoznawania struktury dokumentu, stanom będą odpowiadały części składowe dokumentu (np. dane adresowe, streszczenie, treść itd.).

Otwartym problemem pozostaje dobór topologii modelu – liczba stanów oraz możliwość przejść pomiędzy stanami [7].

7. Podsumowanie

Przedstawione ukryte modele Markowa są jedną z probabilistycznych metod eksploracji danych tekstowych. Innymi metodami są: naiwny klasyfikator bayerowski, stosowany jako metoda referencyjna, stochastyczne gramatyki bez-kontekstowe i inne.

Ukryte modele Markowa zasługują wśród nich na uwagę ze względu na prosty algorytmicznie model obliczeń, a jednocześnie uniwersalny model, który może być stosowany do wielu zagadnień. Praktyczne wykorzystanie ukrytych modeli Markowa wymaga rozszerzenia funkcjonalności oferowanych przez narzędzia dedykowane do text – miningu (WEKA, SAS Text Miner) o procedury obliczeniowe związane z wyznaczaniem parametrów modelu.

8. Bibliografia

[1] L.R. Rabiner, „A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proceedings of the IEEE*, 257-289, 1989.

[2] A. Kao S.R. Poteet, *Natural Language Processing and Text Mining*, Springer--Verlag, Londyn, 2007.

[3] R. Feldman, J. Sanger, *The Text Mining Handbook*, Cambridge University Press, 2007.

[4] Jie Tang, Mingcai Hong, Duo Zhang, Bangyong Liang, Juanzi Li, „Information Extraction: Methodologies and Applications", *Emerging technologies of text mining,* Information Science Reference, Londyn, 2008.

[5] C.D. Manning, H. Schutze, *Foundation of Statistical Natural Language Processing,* MIT Press, 1999.

[6] D.R.H. Miller, T. Leek, R.M. Schwartz, „A Hidden Markov Model Information Retrieval System", *Proceedings of the 22nd ACM SIGIR Conference on Research and development in information retrieval*, 214-221, Nowy Jork, 1999.

[7] K. Seymore, A. McCallum, R. Rosenfeld, „Learning Hidden Markov Model Structure for Information Extraction", *AAAI 99 Workshop on Machine Learning for Information Extraction*, 37-42, 1999.

Hidden Markov Models as a text mining method

M. MAZUREK

In the text mining applications probabilistic models of document are widely used. In this paper the Hidden Markov Models were described as a fundamental method for text processing. Definition of the HMM was presented and the algorithms to find parameters of the model. Some of the possible applications of HMM were suggested.

Keywords: text mining, Hidden Markov Model, information retrieval