Sieci Bayesa w rozpoznawaniu mowy

Anna Mermon

AGH Akademia Górniczo-Hutnicza, Wydział EAIiE, Katedra Automatyki

Streszczenie: Problematyka rozpoznawania mowy nie doczekała się, jak dotąd, kompleksowego rozwiązania. Współczesne efektywne systemy rozpoznawania mowy korzystają najczęściej z metod stochastycznych opartych na ukrytych modelach Markowa. Alternatywą dla nich mogą być sieci Bayesa, będące odpowiednią strukturą do formułowania modeli probabilistycznych, które cechują się jednocześnie precyzją oraz zwartością. Sieci Bayesa mogą reprezentować rozkład prawdopodobieństwa dowolnego zbioru zmiennych losowych. Mnogość dostępnych obecnie algorytmów i narzędzi obliczeniowych sprawia, że testowanie i wdrażanie nowych rozwiązań staje się mniej pracochłonne. Zalety te determinują duże możliwości wykorzystania sieci Bayesa do rozwiązywania praktycznych problemów również w zakresie rozpoznawania mowy.

Słowa kluczowe: sieci Bayesa, sygnał mowy, cyfrowe przetwarzanie sygnałów, rozpoznawanie sygnału mowy, DBN

1. Wprowadzenie

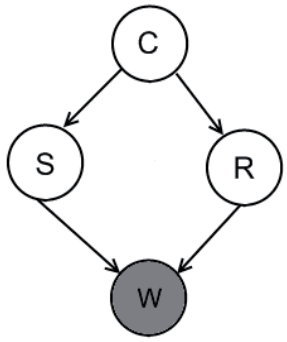
Dynamiczne sieci Bayesa są potężnym i elastycznym narzędziem mogącym służyć do reprezentacji probabilistycznych modeli dla procesów stochastycznych. Można zaobserwować coraz większe zainteresowanie wykorzystaniem tego narzędzia do rozwiązywania problemów praktycznych, w tym także do rozpoznawania elementów sygnału mowy.

Główną cechą dynamicznych sieci Bayesa (DBN) jest to, że z ich pomocą można zamodelować dowolny zestaw zmiennych wraz z ich zmianami w danym czasie. Co więcej, można wyszczególnić dowolny zbiór wzajemnie zależnych założeń, co umożliwia reprezentację rozkładu łącznego w postaci opisanej przez współczynniki. Rozkład na czynniki pierwsze umożliwia tworzenie modeli wydajnych obliczeniowo oraz opisanych za pomocą niewielkiej liczby parametrów. Mnogość modeli probabilistycznych oraz standardowych procedur obliczeniowych sprawia, że testowanie nowych rozwiązań staje się względnie proste. Dodatkowymi zaletami są łatwość zastosowania DBN w przypadku modelu, którego wejście stanowią sygnały w różnych skalach czasu.

Celem pracy jest wykazanie możliwości zastosowania dynamicznych sieci Bayesa do rozwiązania problemu rozpoznawania sygnału mowy.

2. Sieci Bayesa

Ze względu na to, że sieci Bayesa są bardziej ogólną klasą modeli niż DBN, należy najpierw przybliżyć ich tematykę. Sieć Bayesa (BN) jest modelem graficznym reprezentującym połączenia między zbiorem losowych zmiennych – innymi słowy jest ona rozkładem łącznym prawdopodobieństwa. Składa się z kierowanego, acyklicznego grafu, na którym przedstawione są zależności między zmiennymi oraz zbioru rozkładów prawdopodobieństwa, który je wartościuje. Z reguły reprezentacja zbioru rozkładów jest mniejsza objętościowo niż jeden pełny rozkład łączny, co stanowi kolejną zaletę tej metody. Liczba wartości prawdopodobieństw w rozkładzie jest wykładniczo zależna od liczby zmiennych (dla n zmiennych dwójkowych istnieje 2^n prawdopodobieństw). Ponieważ w wielu przypadkach zmienne są względem siebie niezależne, poszczególne rozkłady zmiennych będą zawierały mniej wartości prawdopodobieństw. Jest to przykładowo przedstawione na rys. 1.



Rys. 1. Przykładowa sieć Bayesa modelująca system przewidujący zachmurzenie na podstawie stanu trawy

Fig. 1. Sample of a Bayesian network modeling system that predicts if it's cloudy or not, based of grass state

Pokazana sieć Bayesa (rys. 1) modeluje układ, którego przeznaczeniem jest przewidywanie pogody definiowane jako przewidywanie zachmurzenia na podstawie aktualnego stanu trawy. Sieć składa się z czterech zmiennych logicznych:
C – określa, czy występuje zachmurzenie (tak lub nie),
S – określa, czy włączony jest zraszacz trawy (tak lub nie),
R – określa, czy w danym momencie pada (tak lub nie),
W – określa, czy trawa jest mokra (tak lub nie).

Korzystając z zasady łańcuchowej prawdopodobieństwa, rozkład łączny dla tego układu dany jest za pomocą zależności:

$$P(C,S,R,W) = P(C)P(S \mid C)P(R \mid C,S)P(W \mid C,S,R) \quad (1)$$

Model ten może zostać uproszczony ze względu na zależności warunkowo niezależne do postaci:

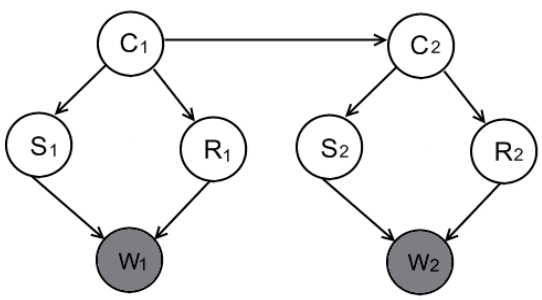
$$P(C,S,R,W) = P(C)P(S \mid C)P(R \mid C)P(W \mid S,R) \quad (2)$$

W tym przypadku reprezentacja zbioru niezależnych od siebie rozkładów jest mniejsza niż byłaby w przypadku całościowego rozkładu łącznego.

3. Dynamiczne sieci Bayesa

Dynamiczne sieci Bayesa (DBN) są rozszerzoną wersją sieci Bayesa uzupełnioną o stochastyczną reprezentację procesów zmieniających się w czasie. Człon „dynamiczne" w nazwie DBN jest nieco mylący ze względu na to, że założeniem dynamicznych sieci Bayesa nie jest zmiana w czasie ich struktury (występują jednakże przypadki kiedy jest to możliwe).

Ponieważ dynamiczna sieć Bayesa ewoluuje w czasie, zatem reprezentowana jest przez dwa modele: poprzedni oraz przejściowy. Przykładowe rozszerzenie sieci Bayesa (rys. 1) do postaci DBN jest przedstawione na rys. 2. Tylko zmienna C jest ze sobą połączona w czasie, co oznacza, że to czy w czasie t występuje zachmurzenie zależy od tego czy zachmurzenie występowało również w czasie t-1.



Rys. 2. Przykładowa dynamiczna sieć Bayesa, powstała z rozwinięcia sieci z rys. 1

Fig. 2. Sample of a dynamic Bayesian network, created by extending network from fig. 1

4. Rozpoznawanie sygnału mowy z wykorzystaniem dynamicznych sieci Bayesa

W systemach rozpoznawania mowy zakłada się, że mowa jest sekwencją pewnych elementarnych dyskretnych jednostek. Takimi jednostkami w zależności od systemu mogą być przykładowo sylaby, wyrazy czy fonemy. Niestety jednostki te nie mają rozłącznych opisów akustycznych, co uniemożliwia jednoznaczne przypisanie sekwencji jednostek akustycznych do rejestrowanych (obserwowanych) dyskretnych obrazów akustycznych. Standardowo stosuje się wówczas w systemach automatycznego rozpoznawania mowy niejawne modele Markowa [9]. Innym, proponowanym podejściem jest wykorzystanie dynamicznych sieci Bayesa.

Proces rozpoznawania elementów mowy może być podsumowany wzorem:

$$\hat{W} = \arg\max_{W \in L} P(W \mid O) \quad (3)$$

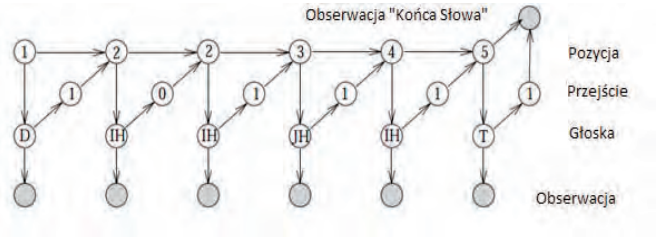
W tym przypadku $\hat{W}$ określa rozpoznawaną wypowiedź ze zbioru dopuszczalnych wypowiedzi oznaczonego jako L, natomiast O jest zarejestrowaną obserwacją. W tej formie rozkład jest trudny do wyznaczenia, ponieważ zmienne losowe występujące we wzorze mogą mieć nieskończenie wiele wartości. W związku z powyższym z wykorzystaniem twierdzenia Bayesa dokonuje się przekształcenia:

$$\hat{W} = \arg\max_{W \in L} \frac{P(W \mid O)P(W)}{P(O)} \quad (4)$$

Ponieważ P(O) ma taką samą wartość dla każdego W, zależność można uprościć:

$$\hat{W} = \arg\max_{W \in L} P(O \mid W)P(W) \quad (5)$$

W wyrażeniu (5) występują dwie nowe wartości, które mogą być obliczone: prawdopodobieństwo, że dana obserwacja jest instancją wypowiedzi (może być wyznaczone z wykorzystaniem modelu akustycznego) oraz prawdopodobieństwo, że dana wypowiedź już się pojawiła, np. po innej wypowiedzi w zdaniu (może być wyznaczone z wykorzystaniem modelu językowego). Aby możliwe było zastosowanie dynamicznych sieci Bayesa w rozpoznawaniu sygnału mowy należy opracować technikę pozwalającą połączyć modele fonetyczne – na przykład sylaby w całe słowo, a dalej w modele składające się z wielu słów. Tworzenie modelu składa się z dwóch niezależnych zagadnień, po pierwsze, określenia dozwolonych przejść między podmodelami. Po drugie, wykorzystanie sieci Bayesa do określenia zachowania każdego modelu składowego. Rys. 3 przedstawia przykład dynamicznej sieci Bayesa.



Rys. 3. Prosty przykład dynamicznej sieci Bayesa

Fig. 3. Simple example of a dynamic Bayesian network

5. Podsumowanie

Dynamiczne sieci Bayesa mogą być zaadaptowane tak, aby spełniać wymagania systemów rozpoznawania mowy oraz mogą modelować wiele istotnych czynników wpływających na proces rozpoznawania. Najważniejszą ich zaletą jest fakt, że dowolny zbiór zmiennych może być połączony z osią czasu. Dodatkowo współdzielenie zmiennych między modelami podrzędnymi prowadzi do naturalnego sposobu opisu zachowań przejściowych, co jest bardzo ważne w przypadku modelowania koartykulacji. Istnieje wiele wydajnych algorytmów o szerokim zastosowaniu, które wspomagają proces implementacji. Dodatkowo rozwiązania z wykorzystaniem DBN są wydajne pod względem statystycznym. Dynamiczna sieć Bayesa jest reprezentacją rozkładu prawdopodobieństwa i może posiadać wykładniczo mniej parametrów niż reprezentacje w innych standardach (np z wykorzystaniem modeli Markowa). W związku z tym parametry te mogą być oszacowane z większą dokładnością z wykorzystaniem ustalonej ilości danych. Wzrost wydajności statystycznej wiąże się dodatkowo ze wzrostem wydajności obliczeniowej.

Bibliografia

1. Flasiński M.: *Wstęp do sztucznej inteligencji*, Wydawnictwo Naukowe PWN, Warszawa 2011.
2. Tadeusiewicz R.: *Speech in Human System Interaction*, [w:] Pardela T., Wilamowski B.M. (red.): 3rd International Conference on Human System Interaction, Rzeszów, IEEE-Press, 2010, 2–13.
3. Martin J.H., Jurafsky D.: *Speech and Language Processing, An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognitio,* Prentice Hall, 2000.
4. Jensen F.V.: *Bayesian Networks and Decision Graphs*, „Information Science and Statistics", Springer, 2011.
5. Murphy K.P.: *Dynamic Bayesian Networks: Representation, Inference and Learning*, PhD thesis, University of California, Berkeley, 2002.
6. Wiggers P.: *Modelling Context in Automatic Speech Recognition.* PhD thesis, TU Delft, Netherlands, 2008.
7. Bourlard H., Morgan N.: *Connectionist Speech Recognition: A Hybrid Approach*, Dordrecht 2002.
8. Jelinek F.: *Statistical Methods for Speech Recognition,* MIT Press, Cambridge, Massachussets, 1997.
9. Grocholewski S.: *Statystyczne podstawy systemu ARM dla języka polskiego*, Wydawnictwo Politechniki Poznańskiej, Poznań 2001.
10. Tadeusiewicz R.: *Sygnał mowy*, Wydawnictwa Komunikacji i Łączności, Warszawa 1988. ■

Bayes networks used in application to speech signal recognition

Abstract: Speech recognition problem hasn't been fully-scaled solved till nowadays. Contemporary effective speech recognition systems mostly use stochastic methods based on Hidden Markov Models. Bayes networks can be alternative to them. BN are appropriate structures to formulate probabilistic models, which are simultaneously precise and compact. They can represent a probability distribution of arbitrary set of random variables. Variety of algorithms and computational tools which are available to use makes testing and implementing new solutions less demanding. Those advantages determine that Bayes networks have potential to be used in solving practical problems also in the area of speech recognition.

Keywords: Bayes networks, speech signal, digital signal processing, speech signal recognition, DBN

mgr inż. Anna Mermon

Ukończyła kierunek Automatyka i Robotyka na Wydziale Elektrotechniki, Automatyki, Informatyki i Elektroniki Akademii Górniczo-Hutniczej w Krakowie. Obecnie asystent w Laboratorium Biocybernetyki Katedry Automatyki. Główne obszary zainteresowań oraz pracy naukowej to sieci Bayesa oraz ich zastosowanie w cyfrowym przetwarzaniu sygnałów, a także akustyka i sieci neuronowe.

e-mail: anna.mermon@gmail.com