Modele pamięci immunologicznej narzędziem eksploracyjnej analizy danych

Algorytmy immunologiczne

Prof. dr hab. inż. Sławomir T. Wierzchoń jest profesorem i kierownikiem Zakładu Sztucznej Inteligencji w IPI PAN

Dr hab. inż. Mieczysław A. Kłopotek jest docentem w IPI PAN. Zajmuje się badaniami naukowymi w dziedzinie sztucznej inteligencji

Dr inż. Krzysztof Ciesielski, adiunkt w IPI PAN, zajmuje się tematyką grupowania danych tekstowych w wyszukiwarkach internetowych

SŁAWOMIR WIERZCHOŃ
KRZYSZTOF CIESIELSKI
MIECZYSŁAW KŁOPOTEK
Instytut Podstaw Informatyki, Warszawa
Polska Akademia Nauk
Slawomir.Wierzchon@ipipan.waw.pl
Krzysztof.Ciesielski@ipipan.waw.pl
Mieczyslaw.Klopotek@ipipan.waw.pl

Jedną z metod wykorzystywanych w informatyce jest naśladowanie natury. Do zadania wyodrębniania jednorodnych grup tematycznych w dużych zbiorach dokumentów tekstowych można użyć algorytmu inspirowanego mechanizmem formowania pamięci immunologicznej

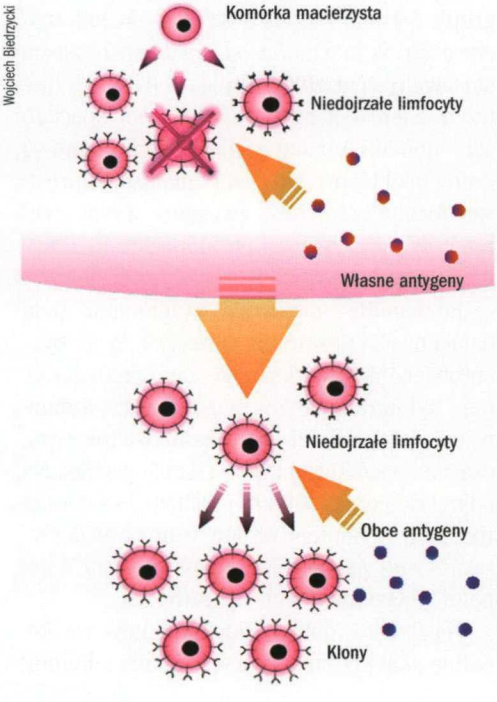
Trwające od kilkunastu lat prace w zakresie sztucznych systemów immunologicznych (SSI) przynależą do nurtu badań nad biologicznie inspirowanymi metodami obliczeniowymi. Obejmują one obliczenia neuronowe, inspirowane mechanizmami ewolucji naturalnej algorytmy ewolucyjne czy też odwołujące się do zasad kolektywnego zachowania i samoorganizacji niezależnych jednostek (np. mrówek czy roju ptaków) algorytmy rojowe. SSI oferują algorytmy, których sposób działania inspirowany jest mechanizmami wyodrębnionymi przez teoretyczną immunologię. Jak można wykorzystać mechanizmy działające w układzie odpornościowym do przeszukiwania danych? By to zrozumieć, prześledźmy pracę układu odpornościowego.

Układ obronny

Głównymi aktorami uczestniczącymi w akcji obronnej układu odpornościowego są limfocyty, czyli białe ciałka krwi. Dzielą się one (w zależności od miejsca powstawania) na dwie zasadnicze grupy: limfocyty typu B i typu T; w skrócie będziemy je nazywać B-komórkami (ewentualnie T-komórkami).

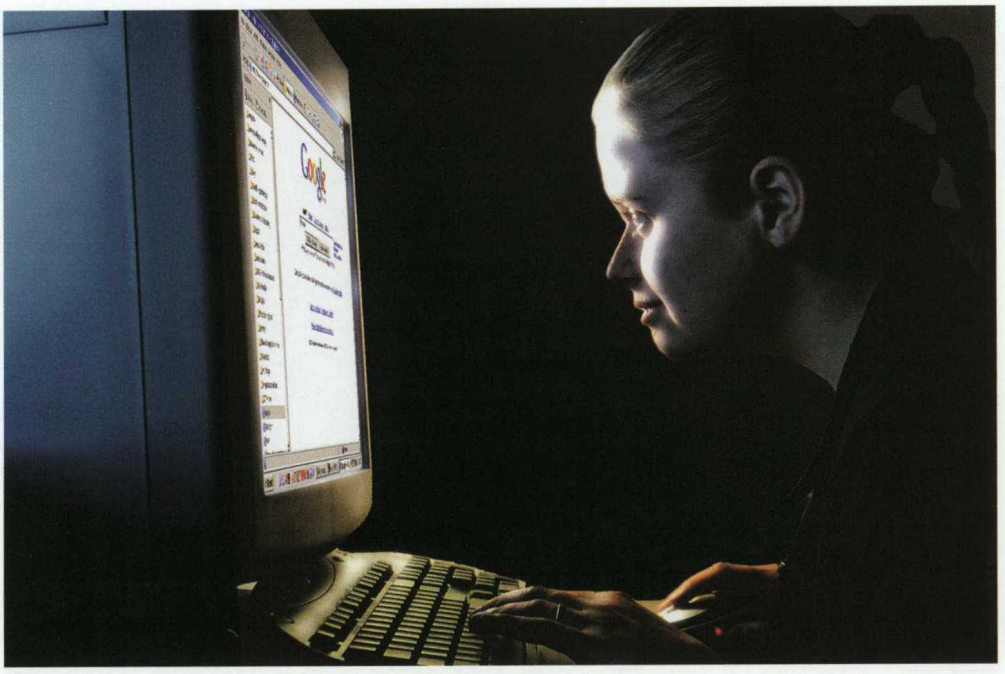
Na powierzchni każdej B-komórki znajdują się receptory nazywane przeciwciałami. Są to proteiny zdolne wiązać antygeny, czyli obce ciała (bakterie, wirusy, grzyby itp.) stanowiące zagrożenie dla funkcjonowania organizmu. Charakterystyczny fragment antygenu, który jest rozpoznawany i wiązany przez przeciwciało, nosi nazwę epitopu lub determinanty antygenowej; podstawową determinantę określa się terminem idiotypu. Podobnie fragment przeciwciała aktywnie wiążący epitop danego antygenu nazywany jest paratopem. Antygeny posiadają wyłącznie epitopy, przeciwciała natomiast wyposażone są zarówno w epitopy, jak i paratopy.

Rzeczywiste paratopy i epitopy są 3-wymiarowymi strukturami. Jeżeli są one komplementarne ze względu na własności geometryczne i fizykochemiczne, mówimy, że paratop rozpoznaje albo wiąże prezentowany epitop. Aby badać interakcje między epitopami a paratopami, wprowadzono pojęcie przestrzeni kształtów, czyli wielowymiarowej przestrzeni, której poszczególne wymiary odpowiadają charakterystykom analizowanych molekuł. W tym kontekście specyficzność wiązania epitop-paratop można trak-



Wojciech Biedrzycki

Mechanizm selekcji klonalnej – procesu optymalizującego odpowiedź odpornościową organizmu. W obecności groźnego antygenu aktywacji ulegają tylko limfocyty, mające receptory rozpoznające dany antygen. Liczba limfocytów, które związą się z właściwymi obcymi antygenami, gwałtownie się zwiększa i szybko dochodzi do wytworzenia swoistej odporności na patogen



Krzysztof Kaliński

Kolosalne zbiory tekstowe to coś, z czym pośrednio styka się każdy, chociażby korzystając z wyszukiwarek internetowych. By szybko przeszukiwać duże zbiory danych, opracowuje się wciąż nowe, coraz doskonalsze algorytmy

tować jako stopień podobieństwa obu molekuł, określany najczęściej za pomocą funkcji odwrotnie proporcjonalnej do odległości między punktami reprezentującymi te molekuły w przestrzeni kształtów.

Jeżeli do organizmu wprowadzono dostatecznie dużą liczbę egzemplarzy pewnego antygenu, z którym ten organizm nigdy wcześniej się nie zetknął, ma miejsce tzw. pierwotna odpowiedź immunologiczna, polegająca na tzw. ekspansji klonów i mutacji hipersomatycznej. Pierwszy z tych terminów oznacza gwałtowne kopiowanie, czy też klonowanie, tych B-komórek, których przeciwciała najsilniej wiążą prezentowane antygeny. Aby zwiększyć skuteczność wynikowych klonów, poddaje się je procesowi bardzo intensywnej mutacji. Zmutowane i efektywne komórki uwalniają przeciwciała do płynów ustrojowych, co pozwala im przemieszczać się w całym organizmie i likwidować zagrożenie. Skuteczne B-komórki podlegają dalszemu klonowaniu i mutacji hipersomatycznej. Równocześnie komórki nieuczestniczące w odpowiedzi immunologicznej są usuwane z organizmu. Proces ten trwa do chwili, gdy koncentracja antygenów spadnie poniżej pewnej wartości progowej. Opisany tu mechanizm nosi nazwę selekcji klonalnej.

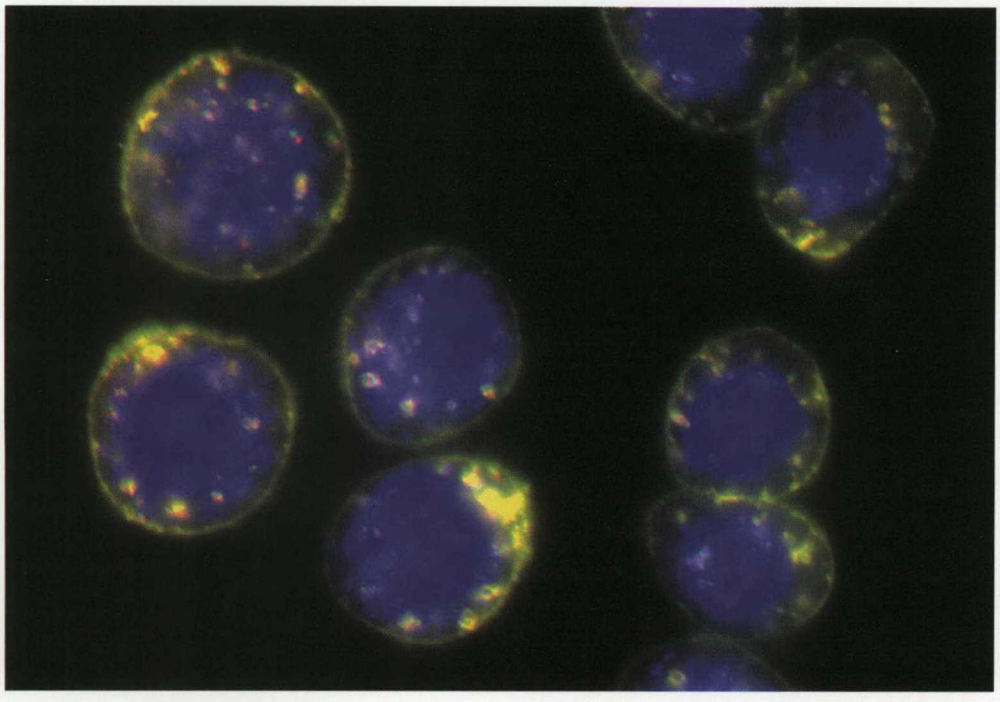
W ramach odpowiedzi układu immunologicznego obserwuje się zjawisko tolerancji: jeżeli stężenie epitopów jest bardzo niskie lub bardzo wysokie, organizm nie reaguje na prezentowany antygen. Jedynie „przeciętne" dawki antygenu wywołują reakcję obronną.

Paratopy przeciwciał są stymulowane przez wszystkie epitopy obecne w organizmie, bez względu na to, czy są one przypisane do antygenów, czy do przeciwciał. Wynika to z faktu, że nowe przeciwciało, powiedzmy Ab_1, stanowi dla organizmu nową proteinę, a jej produkcja w procesie ekspansji klonalnej powoduje odpowiedź organizmu skutkującą produkcją przeciwciała nowego typu, powiedzmy Ab_2. Ogólnie produkcja przeciwciała Ab_i stymuluje produkcję kolejnych typów przeciwciał, przy czym kolejne pokolenia protein tworzą tzw. sieć idiotypową. Jej charakterystyczną cechą jest „samopodtrzymywalność", tzn. sieć ta może istnieć nawet wówczas, gdy inicjujący jej powstanie epitop zostanie całkowicie usunięty z organizmu. Ponowne wprowadzenie antygenu Ag do organizmu skutkuje niemal natychmiastową produkcją efektywnych przeciwciał. Samopodtrzymującą się sieć idiotypową można traktować jako model tzw. pamięci immunologicznej, a zjawisko produkcji „zapamiętanych" przeciwciał skutecznie zwalczających dany typ epitopu nosi nazwę wtórnej odpowiedzi immunologicznej.

Eksploracyjna analiza danych

Ważnym algorytmem inspirowanym zarówno przez teorię selekcji klonalnej, jak i teorię sieci idiotypowych jest aiNET.

Modele pamięci immunologicznej narzędziem eksploracyjnej analizy danych



Invitrogen Molecular Probes, www.probes.invitrogen.com

Choć nie przypominają komputerów, ludzkie komórki układu odpornościowego wypełniają postawione przed nimi zadania w wyjątkowo skuteczny, oparty na efektywnych algorytmach sposób

Zrezygnowano tu z modelowania B-komórek, ograniczając się wyłącznie do przeciwciał. Zarówno epitopy antygenów, jak i paratopy przeciwciał traktowane są tu jako punkty w n-wymiarowej przestrzeni euklidesowej. W procesie generacji przeciwciał wyróżnia się dwa zasadnicze etapy. W pierwszym wykorzystuje się zasady selekcji klonalnej i dojrzewania swoistości. W drugim etapie zachodzą interakcje między komórkami systemu oraz różnicowanie komórek zgodnie z teorią sieci idiotypowych. W szczególności wyznacza się podobieństwo między komórkami i w przypadku, gdy przekracza ono zadaną wartość progową, zbyt podobne komórki są usuwane, aby zredukować redundancję w powstającej pamięci immunologicznej. Ponadto z pamięci usuwane są zbyt specjalizowane komórki, tzn. te przeciwciała, które rozpoznają tylko niewielką liczbę antygenów. Po fazie redukcji następuje wprowadzanie nowych (generowanych losowo) komórek, zapewniające właściwą różnorodność modyfikowanego w kolejnych cyklach repertuaru immunologicznego.

Charakterystyczną cechą omawianego tu rozwiązania jest swoista kompresja danych. Przeciwciała, które można traktować jako prototypy, lokowane są w strategicznych obszarach zajmowanych przez antygeny. Dysponując zredukowanym zbiorem punktów odzwierciedlających najbardziej istotne cechy zbioru danych (antygenów), można przystąpić do ich analizy. Istotne jest to, że analiza prowadzona jest nie na oryginalnym zbiorze antygenów, lecz na liczbowo zredukowanym zbiorze przeciwciał. Ponadto algorytm ten nie wymaga podawania liczby klas, na jakie należy rozdzielić dane, i znakomicie nadaje się do uczenia przyrostowego. Dołączenie nowych danych nie wymaga bowiem ponownej analizy zawartości zagregowanych danych (jak to ma miejsce w przypadku klasycznych algorytmów analizy skupień), a prowadzi jedynie do ewentualnych modyfikacji zbioru komórek pamięciowych.

Analiza dużych kolekcji dokumentów

Zaproponowana przez nas metoda analizy i reprezentacji zawartości dużych zbiorów tekstowych polega na połączeniu zalet omówionego wyżej algorytmu immunologicznego z kontekstową analizą dokumentów.

Realizując powyższe zadanie, przyjęto, że - jak w większości systemów wyszukiwania i przetwarzania informacji - dokumenty reprezentowane są w postaci wektorów $w_d = (w_{d,1}, ..., w_{d,T})$, gdzie indeks d odnosi się do dokumentu, T oznacza liczbę termów (wyrażeń), natomiast $w_{d,t}$ to waga będąca iloczynem liczby wystąpień termu t w dokumencie d i logarytmu z całkowitej liczby dokumentów podzielonej przez liczbę tych

dokumentów, które zawierają term *t* (tzw. waga *tfidf*).

Nowością naszego rozwiązania jest podejście kontekstowe polegające na zastąpieniu jednolitego schematu oceny istotności słów i fraz w całej kolekcji (wyrażonej wagą *tfidf*) oceną uwzględniającą lokalne rozkłady występowania poszczególnych termów w grupach dokumentów o zbliżonej tematyce. Zbiór dokumentów reprezentowany przy użyciu lokalnego schematu ważenia termów tworzy grupę kontekstową (kontekst). Podejście to jest niezależne od wybranego modelu grupowania. Wśród specyficznych kryteriów, optymalizowanych w trakcie identyfikacji grup kontekstowych, wymienić należy równoważenie liczności poszczególnych grup, homogeniczność rozkładu wag termów wewnątrz grupy (jednolitość tematyczna) oraz wyznaczenie podziału przestrzeni termów (słownika) jednocześnie z podziałem zbioru dokumentów.

Relacje między wyróżnionymi kontekstami, jak również przynależne do nich dokumenty przedstawia się w postaci tzw. mapy dokumentów. Tworzenie map dokumentów ma dwie zalety. Z jednej strony umożliwia użytkownikowi szybkie rozeznanie się w zawartości zbioru dokumentów *D*, a z drugiej w trakcie tworzenia mapy wyróżnia się w zbiorze *D* (niekoniecznie rozłączne) grupy dokumentów „podobnych". Operowanie takimi grupami znakomicie przyspiesza proces wyszukiwania interesujących dokumentów. Tworzenie mapy jest więc procesem wykrywania wewnętrznej struktury zbioru obiektów połączonym z wizualizacją wykrytej struktury.

Skuteczność nowego podejścia potwierdzono, przeprowadzając eksperymenty na zbiorze złożonym z 20 000 wiadomości należących do 20 grup dyskusyjnych. Zbadano dwa alternatywne sposoby inicjalizacji początkowej pamięci immunologicznej (tzn. zbioru przeciwciał, z których każde charakteryzuje pewien zasadniczy temat dyskusji). Pierwszy sposób, nazwany inicjalizacją kontekstową, polegał na świadomym wyborze pewnej, mniejszej od liczby grup dyskusyjnych, liczby przeciwciał opisujących skupienia zawierające dokumenty o zbliżonej tematyce. Drugi sposób polegał na losowej inicjalizacji składowych wektorów reprezentujących przeciwciała.

Okazało się, że właściwa inicjalizacja początkowego zestawu przeciwciał w połączeniu z podejściem kontekstowym istotnie wpływają na czas tworzenia sieci idiotypowych: dla modelu kontekstowego wyniósł on około 10 minut, podczas gdy standardowa metoda aiNet wymagała ponad 20 godzin uczenia. Jedną z przyczyn tego zjawiska jest to, że przeciwciała trafnie dobrane w początkowej fazie uczenia są zdolne do długotrwałego przetrwania w pamięci immunologicznej. Wynik ten oznacza, że poziom dojrzałości przeciwciał zależy częściowo od poprawnej inicjalizacji pamięci i wpływa na zbieżność całego algorytmu.

Obserwowane rezultaty wskazują, że algorytmy immunologiczne będące w istocie fuzją algorytmów rojowych i ewolucyjnych mogą stać się uniwersalnym narzędziem rozwiązywania problemów o różnym stopniu złożoności. ■

Chcesz wiedzieć więcej?

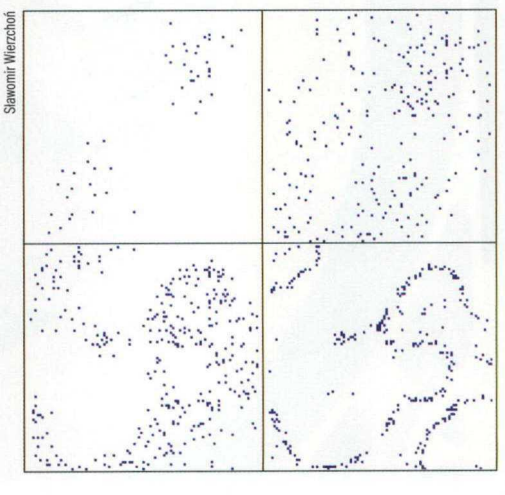
Ciesielski K., Wierzchoń S.T., Kłopotek M.A. (2006). An immune network for contextual text data clustering. *Proc. of the International Conference on Artificial Immune Systems*, 432-445.

Ciesielski K., Kłopotek M.A., Wierzchoń S.T. (2008). Term distribution-based initialization of fuzzy text clustering. *Proc. of the International Symposium on Methodologies for Intelligent Systems.*

De Castro, L.N., Timmis, J. (2002). *Artificial Immune Systems: A New Computational Intelligence Approach*, Londyn: Springer-Verlag.

Wierzchoń S.T. (2007). *Zastosowanie algorytmów immunologicznych w eksploracyjnej analizie danych.* [W:] Kulczycki P., Hryniewicz O., Kacprzyk J. *Techniki Informacyjne w Badaniach Systemowych.* Warszawa: WNT.

http://www.ipipan.eu/~klopotek/BEATCA



Kolejne etapy formowania się samopodtrzymującej się struktury (od górnego lewego do dolnego prawego): po 1300, 1500, 2000 i 4000 iteracji; ostatnia struktura jest prawie niezmienna