Inteligentne systemy przetwarzania informacji

Nauczyć maszynę

JAN KOMOROWSKI
Centrum Bioinformatyki im. Karola Linneusza
przy Uniwersytecie w Uppsali
i Szwedzkim Uniwersytecie Rolniczym
Jan.Komorowski@lcb.uu.se

Profesor Jan Komorowski, informatyk z wykształcenia, prowadzi interdyscyplinarne badania w biomedycynie molekularnej, w szczególności w transkryptomice i proteomice, stosując metody uczenia maszynowego

JACEK KORONACKI
Instytut Podstaw Informatyki, Warszawa
Polska Akademia Nauk
korona@ipipan.waw.pl

Profesor Jacek Koronacki, dyrektor Instytutu Podstaw Informatyki PAN, jest specjalistą w dziedzinie statystycznej analizy danych i uczenia maszynowego, profesorem wizytującym w ICM UW

Miniony wiek bywa nazywany wiekiem informacji. Prawdą jest, że moce obliczeniowe komputerów oraz pojemności ich pamięci rosły w ostatnich dziesięcioleciach nieomal z dnia na dzień

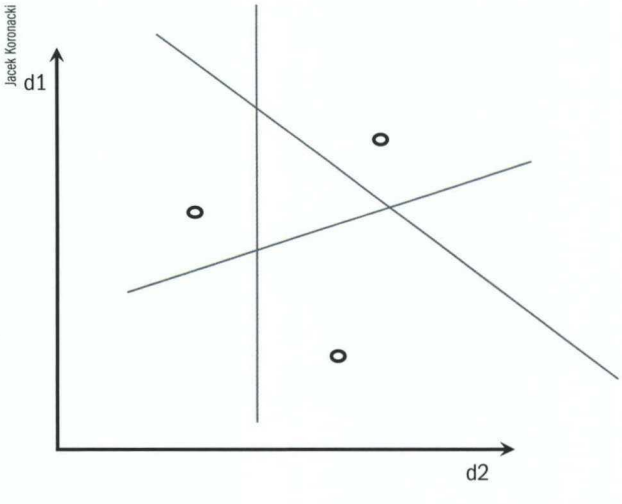
Dziś nierzadko słyszymy, iż weszliśmy w wiek biologii, a nawet szerzej, wiek nauk o życiu. Stało się tak między innymi dlatego, że niezwykle rozwinęły się biotechnologie pozwalające na zbieranie masowych danych o żywych komórkach. Dla przykładu, podczas gdy w końcu lat 90. ubiegłego stulecia mikromacierze pozwalały na równoległą obserwację dziesiątków tysięcy genów, dziś dysponujemy technologiami, które pozwalają zsekwencjonować genom człowieka w ciągu tygodnia, przy czym koszt takiego eksperymentu jest niższy niż 10 tysięcy euro. Zbiór danych z jednego eksperymentu tego typu zawiera około $5 \cdot 10^9$ punktów.

Nic dziwnego, że z końcem ubiegłego wieku nauki o życiu i nauki informacyjne przestały być dziedzinami rozłącznymi. Wcześniej te pierwsze dysponowały małymi ilościami danych i znakomita większość osiąganych rezultatów była opisywana językiem jakościowym o dużym stopniu modalności. Modele budowane przez biologów miały zatem często charakter modeli jakościowych, przypominających zestawy nie w pełni precyzyjnych reguł. Pojawienie się masowych danych jednocześnie umożliwiło i wymusiło zwrócenie się ku całkowicie nowym podejściom. Szybko stało się jasne, jak istotną rolę w obszarze nauk o życiu mają do odegrania metody uczenia maszynowego, zwłaszcza te, które oferują modele czytelne dla człowieka - np. modele w formie drzew lub reguł decyzyjnych. Z jednej strony modele takie pozwalają na modelowanie ilościowe, z drugiej zachowują postać jakościowych reguł.

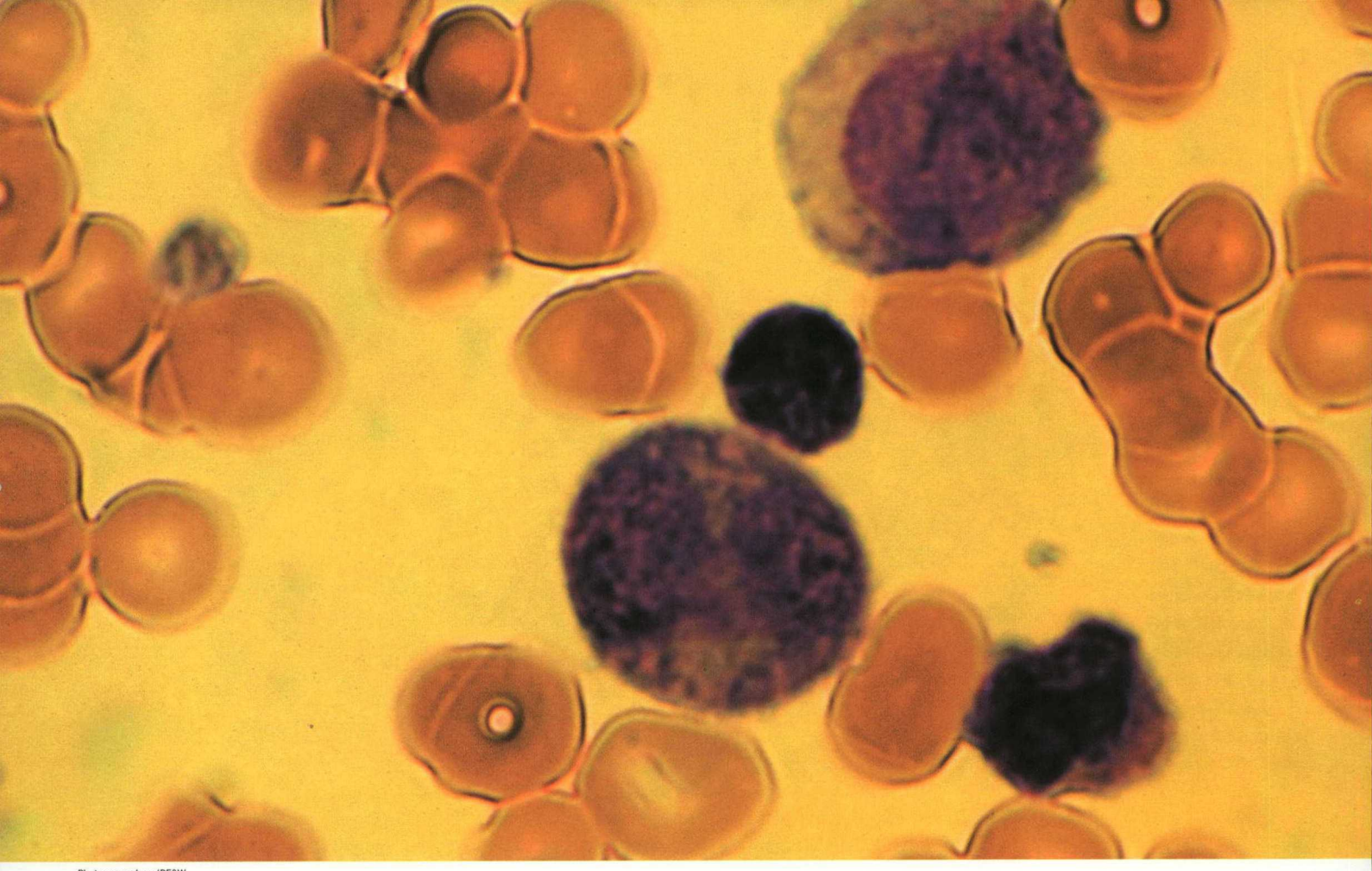
Nauczanie na przykładach

Chodzi o to, że skoro dysponujemy ogromnymi zbiorami informacji, to musimy także dysponować środkami „inteligentnego", a zarazem automatycznego przetwarzania tychże informacji. Człowiek musi być w swoich badaniach wspomagany przez komputer wyposażony w algorytmy stosownie pojętego porządkowania zgromadzonych informacji i wydobywania z nich użytecznej wiedzy. Owemu wydobywaniu wiedzy służą algorytmy uczenia maszynowego, zdolne sprawić, że komputer „uczy się" na podstawie przedstawionych mu przykładów.

Postanowiliśmy krótko opisać zaledwie dwa, ale mamy nadzieję pouczające sposoby wykorzystania w genomice i proteomice metod maszynowego uczenia klasyfikacji obiektów do zadanych klas. Każdy obiekt opisany jest za pomocą wektora, czyli skończonej liczby ustawionych w ustalonym porządku atrybutów. W przypadku różnych obiektów atrybuty



Zbudowanie klasyfikatora rozróżniającego przypadki ostrej białaczki limfoblastycznej i szpiku pomijałoby istotę problemu, gdyż przestrzeń p-wymiarową (u nas p=7129) można rozdzielić na dwa dowolne podzbiory aż (p+1) punktów, których w naszym wypadku jest zaledwie 38



Photoresearchers/BE&W

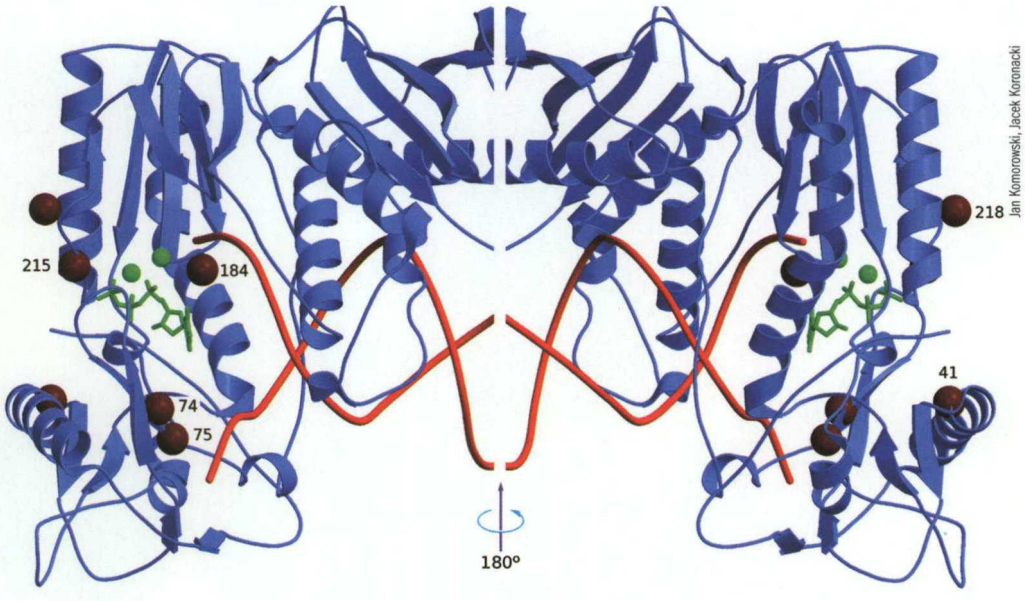
Jak odróżnić poszczególne rodzaje białaczki? Tu obraz mikroskopowy rozmazu krwi obwodowej pacjenta chorującego na ostrą białaczkę szpikową

przyjmują na ogół inne wartości, pochodzące z ustalonych zbiorów wartości możliwych. Algorytm klasyfikacji do zadanych klas uczy się (możliwie trafnego) klasyfikowania obiektów na podstawie przedstawionego mu zbioru uczącego, czyli zbioru obiektów, których przynależności do klas są znane. W dalszym ciągu tego opisu spotkamy się najpierw z obiektami, z których każdy opisany jest wektorem tzw. ekspresji ustalonych genów. Atrybutem jest tu zatem ekspresja konkretnego genu. Obiektem jest w tym wypadku komórka, w której zmierzono ekspresje wybranych genów. W naszym przykładzie będziemy mieli do czynienia ze zbiorem uczącym 38 obiektów, z których każdy opisany jest wektorem ekspresji 7129 genów. Wiadomo przy tym, że 27 obiektów (wiadomo, które to obiekty) należy do klasy obiektów będących przykładami ostrej białaczki limfoblastycznej (w skrócie ALL) i że pozostałych 11 obiektów należy do klasy przykładów ostrej białaczki szpiku (w skrócie AML; jest to słynny zbiór uczący, zebrany i po raz pierwszy przeanalizowany przez Goluba i in.).

Zadaniem algorytmu klasyfikacji jest skonstruowanie na podstawie zbioru uczącego reguł klasyfikacyjnych, zwanych także regułami decyzyjnymi, które pozwolą mu klasyfikować do zadanych klas nowe obiekty, których przynależności do klas nie znamy. Popularnym rodzajem algorytmów klasyfikacyjnych są tzw. drzewa klasyfikacyjne. Ich reguły decyzyjne mają postać implikacji, w każdej z nich występuje koniunkcja warunków nakładanych na wartości wybranych atrybutów oraz konkluzja o przynależności do klasy; na przykład, gdy zadanymi klasami są, powiedzmy, klasy Φ, Ψ i Ω, jedna z reguł może mieć postać: {*jeżeli i-ty atrybut obiektu O przyjmuje wartość większą niż a, j-ty atrybut tego obiektu przyjmuje wartość nie mniejszą niż b oraz k-ty atrybut przyjmuje wartość większą niż c, to obiekt O należy do klasy* Ω}.

Rozpoznać białaczkę

Wracając do przykładu danych dotyczących białaczki, wypada najpierw stwierdzić, iż samo zbudowanie klasyfikatora rozróżniającego obiekty z klas ALL i AML pomijałoby istotę problemu. Po pierwsze, zadanie sprowadzałoby się do podzielenia na dwa rozłączne podzbiory zbioru zaledwie 38 punktów w przestrzeni euklidesowej o wymiarze wynoszącym aż 7129, a to – z geometrycznego punktu widzenia – jest zadaniem trywialnym. Można bowiem udowodnić, że w przestrzeni p-wymiarowej można hiperpłaszczyzną rozdzielić na dwa dowolne i rozłączne podzbiory $(p+1)$ punktów, jeśli tylko punkty te nie leżą w jakiejś podprzestrzeni przestrzeni p-wymiarowej, czyli że w przestrzeni o wymiarze 7129 moglibyśmy łatwo podzielić na dwie klasy nie 38, ale 7130 punktów (rysunek



Jan Komorowski, Jacek Koronacki

Drugim przykładem zastosowania naszej metody jest modelowanie oporności retrowirusa HIV-1 na leki w celu wskazania, które mutacje aminokwasów składających się na widoczną na schemacie odwrotną transkryptazę (RT) są odpowiedzialne za oporność na leki. Czerwone kule to atomy C-alfa aminokwasów wskazanych przez algorytm jako ważne dla oporności na didanozynę

na str. 28, gdzie *p* wynosi zaledwie 2). Samo rozwiązanie takiego zadania klasyfikacji mogłoby przeto być mało interesujące. Ale co znacznie ważniejsze, można oczekiwać, iż o podziale na klasy decyduje w rzeczywistości niewielki ułamek wszystkich 7129 genów. W istocie zatem budowę klasyfikatora chcielibyśmy poprzedzić - albo przynajmniej połączyć z - dowiedzeniem się, które geny mają związek z danym typem choroby. Innymi słowy, interesuje nas nie tylko klasyfikacja, lecz także - jeśli nie przede wszystkim - odpowiedź na pytanie, które geny odpowiadają za zachorowanie na dany typ choroby lub ekspresja których genów wskazuje na rozwiniętą już w organizmie chorobę. Powstało bardzo wiele metod opartych na najszerzej pojętej analizie danych, które próbują dać odpowiedź na takie pytania, jak to właśnie zadane.

Mamy prawo sądzić, że metoda zaproponowana ostatnio przez nasz zespół warta jest szczególnej uwagi. Nasza metoda jest koncepcyjnie bardzo prosta i naturalna, chociaż wymaga powtarzania przez komputer wielu prostych obliczeń. Dany atrybut uznajemy za ważny w procesie specyfikowania danej klasy oraz jej odróżnienia od innych klas, czyli za niosący informację o klasie, jeśli atrybut ten „ma tendencję" do brania udziału w procesie klasyfikowania obiektów ze zbioru uczącego. Owa tendencja lub „gotowość" atrybutu do brania udziału w procesie klasyfikacji, nazywana odtąd ważnością atrybutu, mierzona jest za pośrednictwem budowy wielu (tysięcy) drzew klasyfikacyjnych na różnych - losowo wybranych - podwektorach wektora atrybutów. Metoda wykorzystuje tysiące klasyfikatorów, ale nie po to, by zbudować klasyfikator, lecz po to, by wskazać atrybuty ważne dla danego problemu klasyfikacji. Dopiero mając zestaw takich atrybutów, można, ale nie trzeba, zbudować dowolny klasyfikator. Co przy tym istotne, nasza metoda nie tylko wybiera ważne atrybuty, lecz także dokonuje rankingu tych atrybutów ze względu na ich wagę. W przypadku danych Goluba i in. otrzymaliśmy wyniki, które można uznać za wyjątkowo interesujące.

Wcześniejsze podejścia do problemu znalezienia obiektywnie ważnych atrybutów - bez odwoływania się do wiedzy biologów - sprowadzały się zwykle do znalezienia tych, których wartości różnią się istotnie (w statystycznym sensie), gdy obiekty należą do różnych klas. Analizowano pojedyncze atrybuty, nie uwzględniając ich możliwych interakcji z innymi. W rezultacie znajdowano tylko te geny, których ekspresje odpowiadają późnym stadiom rozważanych typów nowotworu i wyrażają szczególnie dramatyczne zmiany będące skutkami nowotworu. W naszym podejściu poszukujemy atrybutów, których wartości - samodzielnie lub razem z innymi atrybutami - pozwalają rozróżnić obiekty z różnych klas. I w przypadku danych Goluba i in. wykryliśmy również słabo, choć dostatecznie wyraźnie zróżnicowane geny, które są związane z powstawaniem lub skłonnością do danego typu nowotworu. Taki wynik sugeruje, że może-

my diagnozować wczesne stadia nowotworu, a być może nawet przewidywać skłonność do jego powstawania.

Wykrywanie sprawców oporności na lek

Zaproponowana metoda rankingu atrybutów pozwala zarazem *explicite* opisać współzależności między atrybutami oraz ocenić siłę tych współzależności w procesie klasyfikowania obiektów do klas. Drugim przykładem zastosowania naszej metody jest modelowanie oporności retrowirusa HIV-1 na leki. Retrowirus HIV-1 mutuje niezwykle szybko. W ciągu trzech dni po zarażeniu w komórkach gospodarza występuje około 10 milionów wirusów i praktycznie wszystkie jego mutacje. Niektóre z tych mutacji są oporne na dany lek.

Rysunki na str. 30 i 31 przedstawiają najsilniejsze współzależności, jakie wykryliśmy w danych dotyczących oporności na lek o nazwie didanozyna, którego zadaniem jest zahamowanie replikacji retrowirusa (dane pochodzą z *HIV Resistance Database* Uniwersytetu Stanfordzkiego). Dane zawierały 706 obiektów, każdy opisany 560 aminokwasami i każdy należący do jednej z trzech klas: oporny na lek, podatny na lek i o oporności umiarkowanej.

Dotychczasowe badania nad modelowaniem oporności skupiały się na zbudowaniu jak najlepszego klasyfikatora i pomijały problem wyjaśnienia mechanizmu oporności. W naszym podejściu zastąpiliśmy nazwy aminokwasów ich fizyko-chemicznymi własnościami. Używając wiedzy biochemicznej, wyodrębniliśmy 7 najważniejszych własności, które razem dały 3920 atrybutów. Najsilniejsze współzależności okazały się dotyczyć w sumie 6 pozycji (aminokwasów). Z tych 5 było poprzednio uznanych przez ekspertów za istotne w formowaniu oporności, a 1 aminokwas okazał się nowy. Mając informacje, które pozycje są ważne i jakie są między nimi współzależności, mogliśmy przejść do 3-wymiarowej struktury odwrotnej transkryptazy (RT) i przeanalizować mechanizmy oporności - jej nabywania lub utraty (dla ułatwienia pomijamy tutaj oporność umiarkowaną). Podział strukturalny RT wyróżnia domeny obrazowo nazywane kciukiem, dłońmi i palcami. W domenie kciuka mutacje nie występują. Pierwszy wykryty mechanizm dotyczy miejsc aktywnych w dłoni, drugi związany jest z ruchem domeny palców. Mutacja w miejscach aktywnych ma oczywiste konsekwencje - wirus nabrał oporności na lek. O palcach można obrazowo powiedzieć, że przytrzymują one DNA w taki sposób, by kolejne nukleotydy mogły być dopisywane do tworzącego się obiektu. Mutacja w takim miejscu powoduje, że lek, który dotychczas uniemożliwiał palcom utrzymywanie DNA w odpowiednim miejscu, przestaje działać i funkcja palców jest przywrócona.

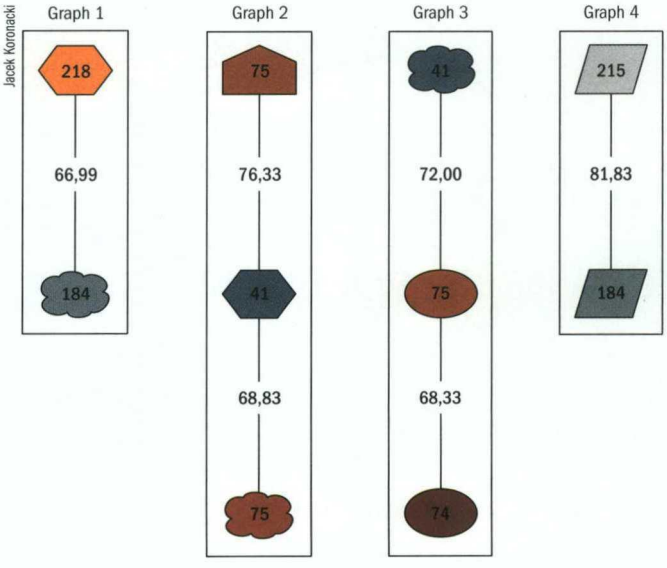
Obecnie zaczęliśmy stosować naszą metodę do problemów, w których rozwiązania nie są jeszcze znane. W ten sposób mamy nadzieję realizować cele biologii obliczeniowej, która buduje *in silicium* modele zjawisk zachodzących *in vivo* lub *in vitro* i przedstawia hipotezy do ostatecznego sprawdzenia w laboratorium.

Zespół realizujący projekt skorzystał z konsultacji dra Krzysztofa Ginalskiego i dra Witolda Rudnickiego z Centrum Modelowania Matematycznego i Komputerowego Uniwersytetu Warszawskiego. ■

Chcesz wiedzieć więcej?

Dramiński M., Rada-Iglesias A., Enroth S., Wadelius C., Koronacki J., Komorowski J. (2008). Monte Carlo Feature Selection for Supervised Classification. *Bioinformatics, 24 (1)*, 110-117.

Dramiński M., Kierczak M., Koronacki J., Komorowski J. (2009). *Monte Carlo Feature Selection and Interdependency Discovery in Supervised Classification.* [W:] Koronacki J., Raś Z., Wierzchoń S., Kacprzyk J. (Red.). *Recent Advances in Machine Learning*, Springer (w druku).



Schemat przedstawia najsilniejsze współzależności, jakie wykryliśmy w danych dotyczących oporności na didanozynę, której zadaniem jest zahamowanie replikacji retrowirusa HIV. Kształt odpowiada własności biochemicznej, kolor - numerowi pozycji, odcinki łączące pozycje wskazują współzależność, liczby na odcinkach - siłę współzależności