SELEKCJI CECH DANYCH BIOMEDYCZNYCH ZA POMOCĄ ALGORYTMU GENETYCZNEGO

Joanna Lis
Instytut Biocybernetyki
i Inżynierii Biomedycznej PAN
ul. Trojdena 4, 02-109 Warszawa
e-mail:joanna.lis@ibib.waw.pl

Streszczenie.

W pracy opisuje się metodę selekcji cech danych biomedycznych za pomocą algorytmu genetycznego (AG). Funkcja kryterialna (funkcja przystosowania w AG) oceniająca wybrany zestaw cech odzwierciedla dowolne preferencje medyczne (dotyczące jakości klasyfikacji, czasu, uciążliwości i ceny badań, ewentualnie dowolnych innych czynników). Klasyfikacji dokonuje się za pomocą metody K-NN na podstawie wyłonionego zestawu cech. Do oceny jakości klasyfikacji zastosowano zmodyfikowany algorytm "leave-one-out". Ocena ta jest przeprowadzona na zbiorze uczącym, zawierającym wyniki badań i odpowiedzi na postawione pytania. Przeprowadzone eksperymenty wskazują na dużą łatwość kształtowania optymalizowanego kryterium. Uzyskiwane w wyniku optymalizacji wyniki rzeczywiście odzwierciedlają oczekiwany cel poszukiwań, pozwalając na dużo większą swobodę w poszukiwaniach niż tradycyjne wyznaczanie minimalnego zestawu cech dającego zadawalającą jakość klasyfikacji.

1. Wprowadzenie.

Problem selekcji cech pojawia się powszechnie w dziedzinach związanych z rozpoznawaniem obrazów, w szczególności w zagadnieniach biologicznych i diagnostyce medycznej. W praktyce wybór zleconych badań pacjenta uwarunkowany jest wieloma czynnikami. Podstawowym czynnikiem jest możliwość uzyskania na podstawie wyników badań odpowiedzi na pytania o rodzaj choroby pacjenta, stopień jej zaawansowania itp. Oprócz tego istotne jest, czy wykonanie badania jest uciążliwe dla pacjenta, czy może mieć skutki uboczne, czy jest czasochłonne lub kosztowne. Czynniki te mogą mieć różną wagę, np. szybkość badań może być istotniejsza od ceny (w nagłych wypadkach) lub na odwrót.

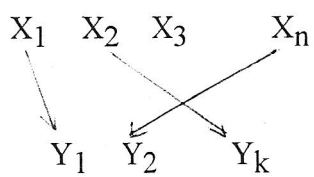
Wybór najlepszego zestawu badań, uwzględniający wymienione uwarunkowania, odpowiada optymalizacji kryterium, zależnego od jakości klasyfikacji i szeroko pojętego "kosztu". Kryterium to może przyjmować skomplikowaną postać, gdyż sam koszt nie zawsze jest sumą kosztów poszczególnych pomiarów (np. uciążliwość dla pacjenta wykonania kilku badań z tej samej próbki krwi jest taka sama jak wykonanie jednego badania z tej próbki, a nie dwa razy większa) i optymalizacja takiego kryterium klasycznymi metodami jest zadaniem bardzo trudnym lub niewykonalnym.

Istniejące algorytmy selekcji cech nie są w stanie podołać temu zadaniu. Najczęściej uwzględniają one tylko kryterium jakości klasyfikacji i ograniczenia liczby cech czy też biorą pod uwagę proste funkcje kosztu, szczególnie w przypadku przyjęcia założeń o postaci rozkładów prawdopodobieństwa w poszczególnych klasach. Natomiast problem selekcji optymalnego zestawu cech w przypadku złożonego kryterium innego niż prosty koszt (a takie są kryteria rzeczywiste) nie jest rozwiązany.

W niniejszej pracy przedstawiono wstępne badania i przemyślenia na temat selekcji cech badań przedmiotowych w chorobach wątroby i uzupełniających informacji klinicznych o badanych pacjentach [1] za pomocą odpowiednio zmodyfikowanego algorytmu genetycznego.

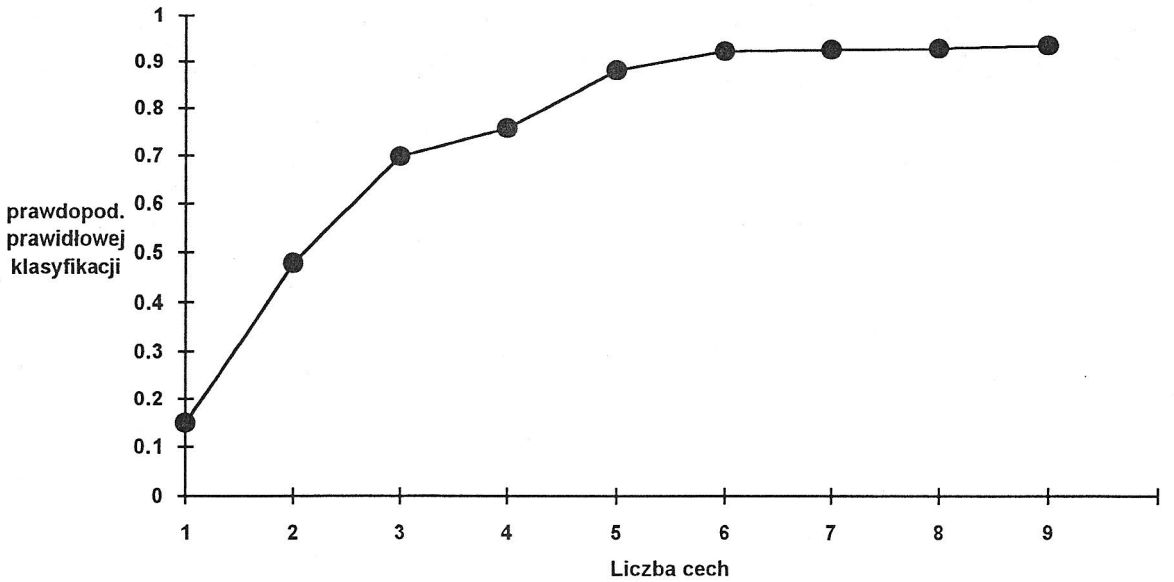
2. Selekcja cech

Selekcja cech jest procesem polegającym na tym, że ze zbioru **X n cech**, opisującego objekt, należy wybrać taki podzbiór **Y k cech** zbioru **X**, by można było względnie łatwo i poprawnie zakwalifikować dany obiekt.



Rysunek 1. Schemat selekcji cech.

Tak wybrany podzbiór powinien zapewniać wysokie prawdopodobieństwo prawidłowej klasyfikacji i powinien być jak najmniej liczny. Pierwszy warunek jest oczywisty, drugi zaś powinien być spełniony dlatego, że pomiar cech bywa kosztowny oraz zbyt wiele cech, szczególnie w przypadku cech zależnych zaciemnia reguły klasyfikacyjne. Rysunek 2 ilustruje, że przy pewnej liczbie cech prawdopodobieństwo prawidłowej klasyfikacji rośnie aż do osiągnięcia takiego stanu, powyżej którego jakość klasyfikacji nie poprawia się znacząco wraz ze wzrostem liczby wybieranych cech.



Rysunek 2. Zależność prawdopodobieństwa prawidłowej klasyfikacji od liczby cech.

Jest również zrozumiałe, że aby realistycznie ocenić jakość klasyfikatora trzeba mieć do dyspozycji znacznie więcej testowych danych niż cech, użytych do konstrukcji klasyfikatora. Biorąc powyższe pod uwagę opracowano wiele algorytmów i technik optymalizacyjnych do znajdowania optymalnego zestawu cech charakteryzujących dany objekt lub zjawisko. Wyczerpujący przegląd technik selekcji cech można znaleźć w książce Devijvera i Kittlera [2], a także w pracy Siedleckiego i Sklansky'ego [7].

Ogólnie można powiedzieć, że w skład algorytmów selekcji cech wchodzą następujące dwie procedury:

procedura poszukiwań

procedura estymacji błędu

Procedura poszukiwań kieruje wyborem kolejnych rozwiązań w celu otrzymania rozwiązania optymalnego. Do oceny tego wyboru posługuje się procedurą estymacji błędu.

Istnieje wiele algorytmów selekcji cech. Najprostszymi ale i najbardziej kosztownymi obliczeniowo są algorytmy sprawdzające wszystkie podzbiory cech. Chociaż łatwo je zaimplementować na komputerze, to ich złożoność obliczeniowa rośnie wykładniczo przez co są bardzo czasochłonne. Istnieje także liczna grupa algorytmów selekcji cech zbliżona do algorytmu *"hill climbing"* . Algorytmy należące do tej grupy (*K-factor selection, Bhattacharyya distance selection, sequential forward selection, sequential backward selection, sequential p forward q backforward selection)* polegają na sekwencyjnym dodawaniu lub odejmowaniu cech od wektora cech stanowiącego aktualne rozwiązanie problemu. Jednakże metody te są dość podatne na wpadanie w lokalne minima co w rezultacie nie prowadzi do otrzymania optymalnego zestawu cech. Inną grupę algorytmów

selekcji cech stanowią algorytmy bazujące na algorytmie *branch-and-bound* Narendry i Fukunagi [5]. Algorytmy te opierają się na założeniu monotoniczności funkcji błędu ze względu na zawieranie się podzbiorow cech. Takie kryteria jak: dyskriminanta Fishera, odległość Mahalanobisa, odległość Bhattacharyya czy entropia Vajda spełniają warunek monotoniczności. Jednakże rozwiązania osiągane za pomocą tych algorytmów mogą nie być optymalne w sensie bayesowskim.

Natomiast w przypadku problemów o dużej liczbie cech, ze względu na czas obliczeń i jakość uzyskiwanych wyników, przewagę uzyskują algorytmy genetyczne [3], [8].

3. Proponowana metoda selekcji cech.

Proponowana metoda selekcji cech stosuje następujące rozwiązania.

1. Funkcja kryterialna oceniająca wybrany zestaw cech odzwierciedla dowolne preferencje medyczne (dotyczące jakości klasyfikacji, czasu, uciążliwości i ceny badań, ewentualnie dowolnych innych czynników).

2. Zakłada się, że klasyfikacja na podstawie wyłonionego zestawu cech przeprowadzana będzie za pomocą metody K-NN (K - najbliższych sąsiadów).

3. Do oceny jakości klasyfikacji zastosowany jest zmodyfikowany algorytm leave-one-out. Ocena ta jest przeprowadzona na zbiorze uczącym, zawierającym wyniki badań i odpowiedzi na postawione pytania.

Metoda leave-one-out jest dobrym estymatorem prawdopodobieństwa prawidłowej klasyfikacji, ale jest czasochłonna (przy dużych zbiorach danych w szczególności). Ma to duże znaczenie i w przypadku algorytmu genetycznego, gdzie funkcja przystosowania jest wielokrotnie obliczana, ta duża czasochłonność jest niepożądana. Aby to przezwyciężyć zmodyfikowano metodę "leave-one-out" następująco:

- *tylko losowy podzbiór zbioru uczącego jest klasyfikowany na podstawie reszty zbioru, a podzbiór ten jest losowany dla każdego chromosomu oddzielnie.*

Pozwala to znacząco przyspieszyć obliczanie funkcji przystosowania w stosunku do standardowego algorytmu leave-one-out.

4. Optymalizacja funkcji kryterialnej na zbiorze wektorów binarnych (zestawów cech) dokonywana jest za pomocą algorytmu genetycznego. Zastosowanie tego algorytmu polega na odpowiedniej konstrukcji operatorów genetycznych i doborze parametrów kontrolnych algorytmu.

4. Algorytm Genetyczny w selekcji cech biomedycznych

Reprezentacja rozwiązań

Zastosowano binarną reprezentację rozwiązań. Każdy członek populacji oznaczany jest jako binarny string (chromosom), którego długość równa jest liczbie możliwych do wyboru cech. Przy czym, na kolejnych pozycjach w chromosomie: 1 oznacza, że dana cecha została wybrana do procesu klasyfikacji, zaś 0 oznacza, że dana cecha została odrzucona i nie bierze udziału w klasyfikacji. Na przykład: chromosom (11111) oznacza, że wszystkie cechy (w tym przypadku 5 cech) biorą udział w klasyfikacji, natomiast chromosom (10100) oznacza, że tylko cechy o numerze 1 i 3 zostały wzięte pod uwagę podczas gdy resztę cech pominięto.

Populacja początkowa

Metoda generacji populacji rozwiązań początkowych polega na losowym wyborze chromosomów z przestrzeni rozwiązań, którą w tym przypadku stanowią stringi binarne z wszelkimi możliwymi kombinacjami cech.

Funkcja Przystosowania

Funkcja przystosowania jest zdefiniowana w ten sposób, by zawierała główne kryteria optymalnego wyboru zestawu cech w diagnostyce medycznej, jakimi są **prawidłowość klasyfikacji, szybkość, koszt oraz uciążliwość pomiaru cech**. Funkcja ta przybiera następującą postać:

$$F(\mathbf{X}) = P(\mathbf{X}) - \alpha K(\mathbf{X}) - \beta C(\mathbf{X}) - \gamma U(\mathbf{X}) - \lambda \qquad (*)$$

gdzie

$\mathbf{X} = (X_1, X_2, \ldots, X_n)$, $X_i \in \{0, 1\}$

n - liczba dostępnych cech

P - prawdopodobieństwo prawidłowej klasyfikacji

K - koszt pomiaru cech

C - czas pomiaru cech

U - uciążliwość pomiaru cech

α, β, γ, λ - współczynniki

Operatory genetyczne

Przed oceną kolejnych populacji chromosomy poddawane są działaniu operatorów genetycznych (z częstością zależną od stałych charakteryzujących algorytm) [4] :

Operator krzyżowania wytwarza na podstawie dwóch chromosomów rodzicielskich dwa chromosomy potomne i zgodnie z przyjętą strategią powstawania następnego pokolenia, chromosomy rodzicielskie zostają zapomniane, na ich miejsce wchodzą chromosomy potomne. Stosowany jest tzw. jednorodny operator krzyżowania (uniform crossover). Operator ten dla każdej kolejnej cechy (i-tej cechy) wykonuje następujące działania:

- z równym prawdopodobieństwem losowany jest jeden z rodzicielskich zestawów cech
- z wylosowanego rodzicielskiego zestawu cech przepisywana jest do pierwszego zestawu potomnego informacja o występowaniu i-tej cechy
- z pozostałego, niewylosowanego rodzicielskiego zestawu cech przepisywana jest do drugiego zestawu potomnego informacja o występowaniu i-tej cechy

Operator mutacji

Z określonym niewielkim prawdopodobieństwem informacja o występowaniu każdej z cech w zestawie jest zmieniana (cecha jest losowo usuwana lub dodawana do zestawu).

Selekcja

Ze starej, istniejącej populacji tworzona jest nowa populacja za pomocą tzw. ruletki.

Parametry kontrolne

Parametrami kontrolnymi algorytmu genetycznego są rozmiar populacji, prawdopodobieństwo krzyżowania i prawdopodobieństwo mutacji. W przypadku problemów o reprezentacji binarnej znane są wartości tych parametrów zapewniające dobre rezultaty dla szerokiego zakresu optymalizowanych problemów [6] :

rozmiar populacji	20 - 30
prawdopodobieństwo krzyżowania	0.75 - 0.95
prawdopodobieństwo mutacji	0.005 - 0.01

5. Eksperymenty

Weryfikację opisanej metody przeprowadzono na danych hepatologicznych pochodzących z systemu komputerowego HEPAR, opracowanego w Instytucie Biocybernetyki i Inżynierii Biomedycznej PAN przy współudziale Centrum Medycznego Kształcenia Podyplomowego oraz Kliniki Gastroenterologii Instytutu Żywności i Żywienia w Warszawie. Zbiór danych składa się z 40 wyników badań (cech) - rezultatów wywiadu lekarskiego, badań przedmiotowych oraz badań laboratoryjnych (Tabela 1) wykonanych dla każdego z 62 pacjentów, cierpiących na 4 różne rodzaje chorób watroby: *hepatitis chronica activa, hiperbilirubinaemia functionalis, cirrhosis hepatis compensata, cirrhosis hepatis decompensata.*

Przeprowadzono eksperymenty badające działanie opisanego powyżej algorytmu dla różnych celów poszukiwań (przy różnych optymalizowanych kryteriach).

A) Przeprowadzono selekcję cech w ten sposób, aby uzyskać maksymalne prawdopodobieństwo prawidłowej klasyfikacji, przy jak najmniejszej liczbie użytych cech. Użycie jak najmniejszej liczny cech mialo drugorzędne znaczenie - współczynniki

optymalizowanej funkcji zostały tak dobrane, że nawet jeden dodatkowy dobrze zaklasyfikowany punkt miał większy wpływ na optymalizowaną funkcję, niż dowolnie duże ograniczenie liczby wykorzystywanych cech.

Fukcja klasyfikacji miała postać:

$F(X) = P(X) - 0.01*K(X)$

gdzie: **P(X)** - prawdopodobieństwo prawidłowej klasyfikacji

K(X) - koszt pomiaru cech zdefiniowany został jako liczba użytych cech podzielona przez liczbę dostępnych cech (przyjeto równy koszt pomiaru cech).

Otrzymano zestaw **10** cech, w tym **5** odpowiadających badaniom laboratoryjnym (cechy o numerach 30, 34, 35, 37, 39), prawdopodobieństwo prawidłowej klasyfikacji zostało ocenione na **88.53%** (Tabela 1)

1. osłabienie, apatia	21. pajączki naczyniowe
2. bóle głowy	22. obrzęki tkanki podskórnej
3. świąd skóry	23. wątroba powiększona
4. bóle w nadbrzuszu	24. wątroba bolesna
5. ucisk w prawym podżebrzu	25. wątroba nierówna
6. bóle stawowo mięśniowe	26. wątroba nadmiernie spoista
7. skaza krwotoczna	**27. brzeg wątroby ostry**
8. żółtaczka	**28. brzeg wątroby nierówny**
9. upośledzone łaknienie	29. wysięk w jamie brzusznej
10. wzdęcie jelitowe	**30. OB**
11. biegunki	31. krwinki płytkowe (w tys.)
12. zła tolerancja tłuszczów	32. wskaźnik protrombinowy
13. zła tolerancja alkoholu	33. BUN
14. picie alkoholu	**34. bilirubina całkowita**
15. WZW	**35. AlAT**
16. przebyte wstrzyknięcia	36. AspAt
17. leki hepatotoksyczne	**37. fosfotaza alkaliczna**
18. WZW w rodzinie	38.GGTP
19. zażółcenie skóry	**39.trójglicerydy**
20. zmiany krwotoczne	40.HBsAg w surowicy

Tabela 1. Zestaw 10 cech wybranych podczas selekcji.

B) Przeprowadzono selekcję cech w ten sposób, aby uzyskać maksymalne prawdopodobieństwo prawidłowej klasyfikacji, przy jak najmniejszej liczbie użytych cech odpowiadających badaniom laboratoryjnym. Eksperyment ten symulował poszukiwanie zestawu cech przydatnego do szybkiego diagnozowania - głównie na podstawie wywiadu lekarskiego. Inaczej niż w poprzednim eksperymencie, użycie jak najmniejszej liczny badań laboratoryjnych miało porównywalne znaczenie z osiągnięciem dobrej jakości klasyfikacji. Ilość użytych cech odpowiadających wywiadowi lekarskiemu nie miała wpływu na optymalizowane kryterium.

Fukcja klasyfikacji miała postać:

$$F(X) = P(X) - 0.05*C(X)$$

gdzie: **P(X)** - prawdopodobieństwo prawidłowej klasyfikacji

C(X) - czas pomiaru cech zdefiniowany został jako liczba użytych cech odpowiadających badaniom laboratoryjnym.

Otrzymano zestaw **8** cech, w tym **1** cecha odpowiada badaniom laboratoryjnym (cecha numer 30), prawdopodobieństwo prawidłowej klasyfikacji zostało ocenione na **85.25%** (Tabela 2).

1. osłabienie, apatia	21. pajączki naczyniowe
2. bóle głowy	22. obrzęki tkanki podskórnej
3. świąd skóry	23. wątroba powiększona
4. bóle w nadbrzuszu	24. wątroba bolesna
5. ucisk w prawym podżebrzu	25. wątroba nierówna
6. bóle stawowo mięśniowe	26. wątroba nadmiernie spoista
7. skaza krwotoczna	27. brzeg wątroby ostry
8. żółtaczka	**28. brzeg wątroby nierówny**
9. upośledzone laknienie	**29. wysięk w jamie brzusznej**
10. wzdęcie jelitowe	**30. OB**
11. biegunki	31. krwinki płytkowe (w tys.)
12. zła tolerancja tłuszczów	32. wskaźnik protrombinowy
13. zła tolerancja alkoholu	33. BUN
14. picie alkoholu	34. bilirubina całkowita
15. WZW	35. AlAT
16. przebyte wstrzyknięcia	36. AspAt
17. leki hepatotoksyczne	37. fosfotaza alkaliczna
18. WZW w rodzinie	38.GGTP
19. zażółcenie skóry	39.trójglicerydy
20. zmiany krwotoczne	40.HBsAg w surowicy

Tabela 2. Zestaw 8 cech wybranych podczas selekcji.

7. Wnioski

Zarówno wymienione eksperymenty, jak i szereg innych, analogicznych prób, wskazują na dużą łatwość kształtowania optymalizowanego kryterium za pomocą funkcji postaci (*). Uzyskiwane w wyniku optymalizacji wyniki rzeczywiście odzwierciedlają oczekiwany cel poszukiwań, pozwalając dużo większą swobodę w poszukiwaniach niż tradycyjne wyznaczanie minimalnego zestawu cech dającego zadawalającą jakość klasyfikacji.

Pełne wykorzystanie opisywanej metody wymagać będzie dokładnego określenia czasochłonności i kosztu wykonywanych pomiarów, a także wprowadzenia skali ich uciążliwości dla pacjenta. Po spełnieniu tych warunków, zastosowanie opisanej metody prowadzić będzie do:

1. wczesnego wyboru optymalnego postępowania z chorymi
2. ograniczenia inwazyjności sposobów badania do minimum i minimalizacji ilości koniecznego do badania materiału.
3. zmniejszenie kosztów finansowych zestawu badań koniecznych dla ustalenia sposobu postępowania lekarskiego

Niniejsza praca została zrealizowana w ramach projektu badawczego Nr 8T11E 01411 finansowanego przez KBN.

7. Literatura

[1] Bobrowski L. i inni, "Hepar - komputerowy system wspomagania diagnostyki i analizy danych biomedycznych", Prace Instytutu Biocybernetyki i Inżynierii Biomedycznej PAN, Nr 31, 1992.

[2] Devijver P. A. and J. Kittler, "Pattern Recognition: a statistical approach", Prentice-Hall, London, 1982.

[3] Kunchewa L., "Genetic algorithm for feature selection for parallel classifiers", Information Processing Letters 46, pp. 163-168, 1993.

[4] Lis J. "Algorytmy syntezy klasyfikacyjnych sieci neuronowych", Praca doktorska, Instytut Biocybernetyki i Inżynierii Biomedycznej PAN, 1994.

[5] Narendra P. M. and K. Fukunaga, "A branch and bound algorithm for feature subset selection" IEEE Trans. Comp. 26, pp.917-922, 1977.

[6] Schaffer J. D., R. A. Caruana, L. J. Eshelman and R. Das "A study of control parameters affecting online performance of genetic algorithms for function optimization." In Proc 3rd. Int. Conf. Genetic Algorithms, Morgan Kaufman, Los Altos, 1989.

[7] Siedlecki W. and J. Sklansky, "On automatic feature selection", International Journal of Pattern Recognition and Artificial Intelligence, vol.2, no. 2, pp.197-220, 1988.

[8] Vriesenga M.R., "Genetic selection and neural modelling for designing pattern classifiers", Ph. Dissertation, University of California, Irvine, 1995.