Zaawansowane Metody Eksploracji

Selekcja atrybutów

Prof. dr hab. inż. Grzegorz Dudek
Wydział Matematyki i Informatyki
Uniwersytet Łódzki

Cel

Celem selekcji atrybutów (selekcji cech, *feature selection*) jest wybranie atrybutów, które zapewniają najlepsze działanie modelu. Atrybuty nieistotne, nadmiarowe, nieskorelowane z etykietami przykładów są usuwane.

Zysk z selekcji atrybutów:

- uproszczenie modelu
- poprawa dokładności
- poprawa generalizacji
- redukcja czasu uczenia
- redukcja złożoności pamięciowej
- umożliwienie wizualizacji, gdy liczba atrybutów po selekcji jest nie większa od 3

Metody

Metody selekcji atrybutów:

- filtracyjne – atrybuty do usunięcia określa się na podstawie analizy danych zawartych w zbiorze trenującym (usunięcie atrybutów nieskorelowanych z klasą/wartością docelowa funkcji, usunięcie atrybutów silnie skorelowanych z innymi atrybutami)

- typu *wrapper* – atrybuty do usunięcia określa się na podstawie analizy generowanych przez model wyników (tworzymy hipotezy dla różnych podzbiorów atrybutów i wyznaczamy ich dokładności). Podzbiór atrybutów, który zapewnia największą dokładność hipotezy uznajemy za optymalny. Podzbiory atrybutów możemy generować w pewien systematyczny sposób (podejście deterministyczne) lub wykorzystując stochastyczne metody przeszukiwania

- typu *frapper* – połączenie dwóch powyższych

- wbudowane – stanowiące integralną część modelu, np. drzewa decyzyjne, indukcja reguł

Alternatywna do selekcji jest ekstrakcja atrybutów – wyznaczenie nowych atrybutów, które są funkcjami atrybutów oryginalnych, np. PCA.

Regresja krokowa

Problem: duża liczba atrybutów (zależnych od siebie, nie wpływających na zmienną wyjściową *y*)

Cel: wybrać niewielki podzbiór atrybutów, który pozwoli uprościć model zachowując jego dokładność

Regresja krokowa (*stepwise regression*) pozwala wyłonić w procedurze krokowej atrybuty istotne i zbudować na nich model liniowy, który zapewnia najmniejszy błąd regresji.

Konstrukcja modelu regresji krokowej może przebiegać w trzech trybach:

- Krokowe dodawanie atrybutów
  1. Ustal $\Phi = \{x_1, x_2, \ldots, x_n\}$ – zbiór atrybutów kandydujących i $\Omega = \varnothing$ – zbiór atrybutów istotnych
  2. Powtarzaj dla każdego $x_i \in \Phi$:
     2.1. Zbuduj model z wykorzystaniem wszystkich atrybutów z Ω i *i*-tego atrybutu z Φ; odnotuj błąd tego modelu
  3. Jeśli nie nastąpiła poprawa modelu w p. 2 – zakończ
  4. Wybierz atrybut z Φ, dla którego nastąpiła największa poprawa modelu i przenieś go z Φ do Ω
  5. Powtórz kroki 2-5

Regresja krokowa

- Krokowa eliminacja atrybutów
  1. Ustal $\Omega = \{x_1, x_2, \ldots, x_n\}$ – zbiór atrybutów istotnych
  2. Powtarzaj dla każdego $x_i \in \Omega$:
     2.1. Zbuduj model z wykorzystaniem atrybutów z Ω pomijając *i*-ty atrybut; odnotuj błąd tego modelu
  3. Jeśli nie nastąpiła poprawa modelu w p. 2 – zakończ
  4. Wybierz atrybut z Ω, po pominięciu którego nastąpiła największa poprawa modelu i usuń go z Ω
  5. Powtórz kroki 2-5
- Naprzemienne użycie dwu powyższych trybów

Miarą poprawy modelu *jest* tzw. *p*-wartość (liczbowe wyrażenie istotności statystycznej) testu statystycznego *F*-Snedecora* (stosuje się też inne kryteria).

Podane algorytmy selekcji krokowej są uniwersalne – można je stosować dla dowolnych modeli uczenia maszynowego. Występują pod nazwą *sequential feature selection* (*forward and backward*).

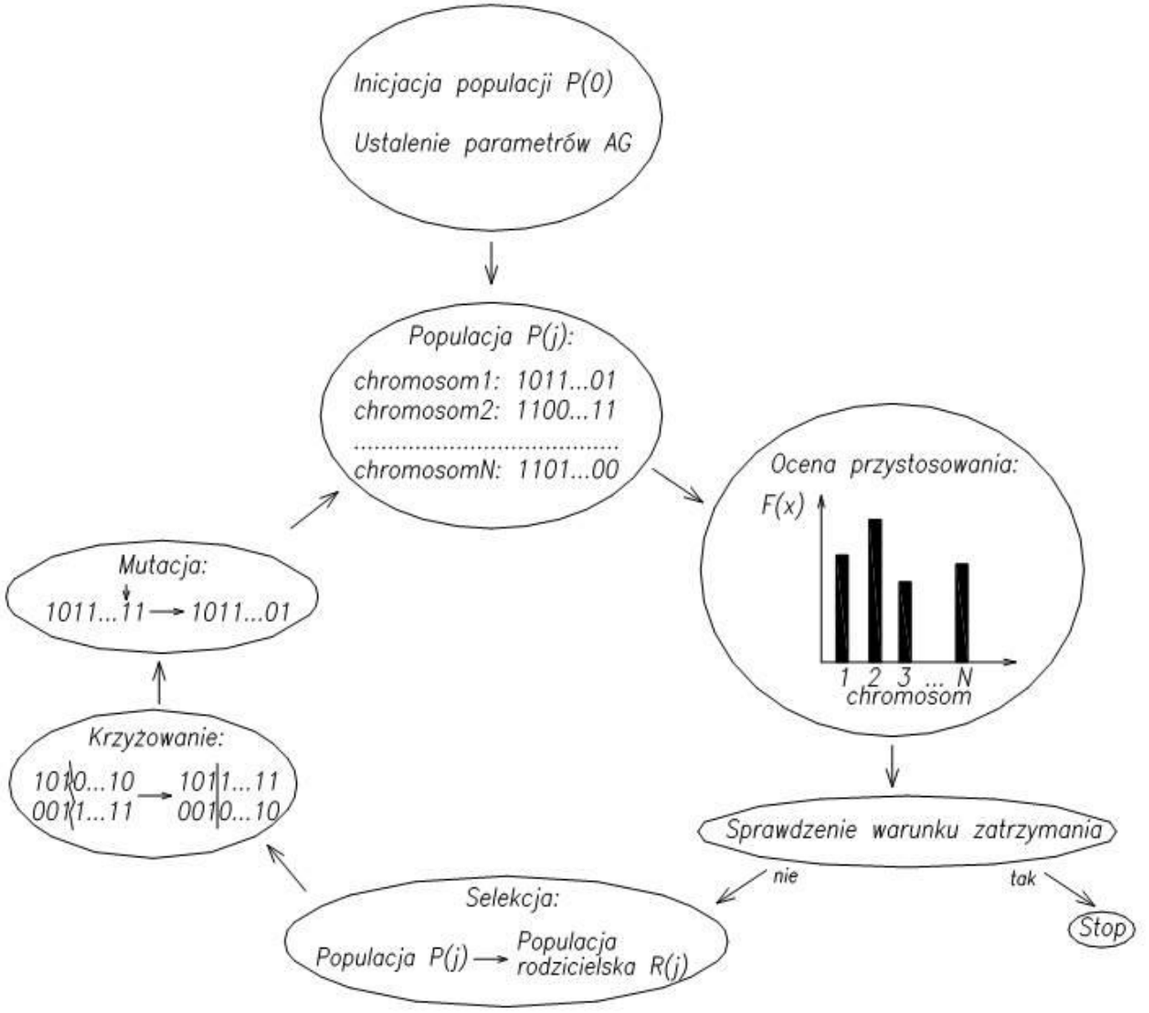
* patrz: Jóźwiak J., Podgórski J.: Statystyka od podstaw. PWE 2001, str. 403

Stochastyczne metody selekcji atrybutów

Sekwencyjna selekcja cech jest deterministyczna, nie zapewnia optymalnych rozwiązań.

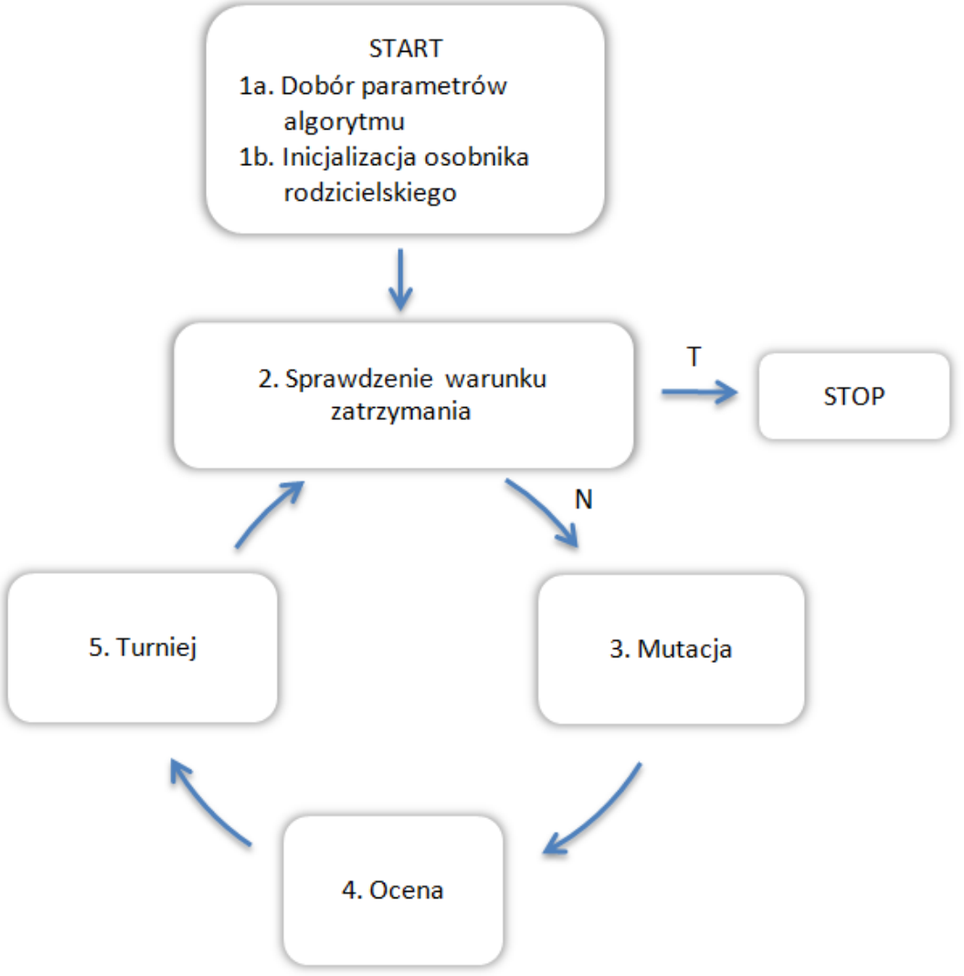
Metody stochastyczne są odporne na minima lokalne. W tych metodach rozwiązanie reprezentowane jest jako ciąg bitów: 01001010…100. Zero na *i*-tej pozycji oznacza, że *i*-ty atrybut jest nieistotny (wypada z modelu), a jedynka – że jest istotny (pozostaje w modelu).

Do stochastycznego przeszukiwania przestrzeni rozwiązań (ciągów binarnych) można zastosować algorytmy genetyczne:



Stochastyczne metody selekcji atrybutów

lub przeszukiwanie turniejowe:

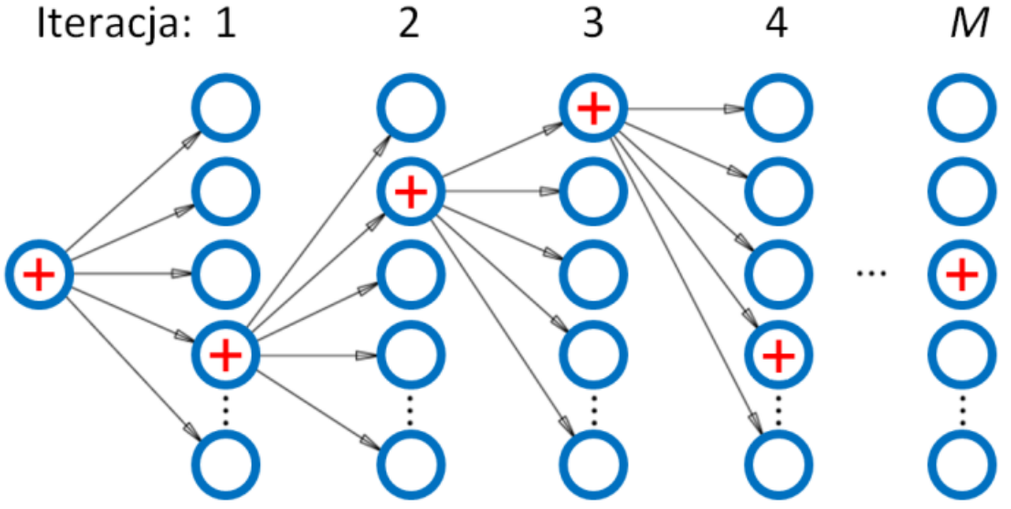


Mutacja

- tworzymy *L* klonów osobnika rodzicielskiego
- w każdym klonie zmieniamy jeden wybrany losowo bit (w każdym klonie inny)

Turniej

- wybieramy osobnika potomnego z najwyższą oceną, który staje się rodzicem w następnej iteracji



⇒ Dudek G.: **Tournament searching method to feature selection problem**. ICAISC 2010, Springer LNCS 6114, pp. 437-444, 2010

Regresja grzbietowa

Problem: duże wagi w_i w modelu liniowym (co do modułu) sprawiają, że wyjście y jest wrażliwe na małe zmiany wejść x_i

Cel: zmniejszyć wagi (co do modułu)

W regresji grzbietowej (*ridge regression*) kryterium zawiera sumę kwadratów wag jako składnik kary[†]:

$$E(h_{\mathbf{w}}) = \sum_{\mathbf{x}\in P} (y - h_{\mathbf{w}}(\mathbf{x}))^2 + \lambda \sum_{i=1}^{n} w_i^2 = (\mathbf{Y} - \mathbf{X}\mathbf{w})^T (\mathbf{Y} - \mathbf{X}\mathbf{w}) + \lambda \mathbf{w}^T \mathbf{w}$$

gdzie $\lambda \geq 0$ jest parametrem określającym stopień uwzględnienia kary w kryterium.

Dla $\lambda = 0$ otrzymujemy zwykły model regresji liniowej; dla $\lambda = \infty$ otrzymujemy zerowe wagi.

Aby wyrównać wpływ poszczególnych wag na wartość kary przed wykonaniem obliczeń należy sprowadzić wartości wszystkich atrybutów do tej samej skali (wariancja próbkowa wszystkich atrybutów powinna wynosić 1).

Wagi minimalizujące powyższe kryterium wyznacza się ze wzoru:

$$\mathbf{w} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{Y}$$

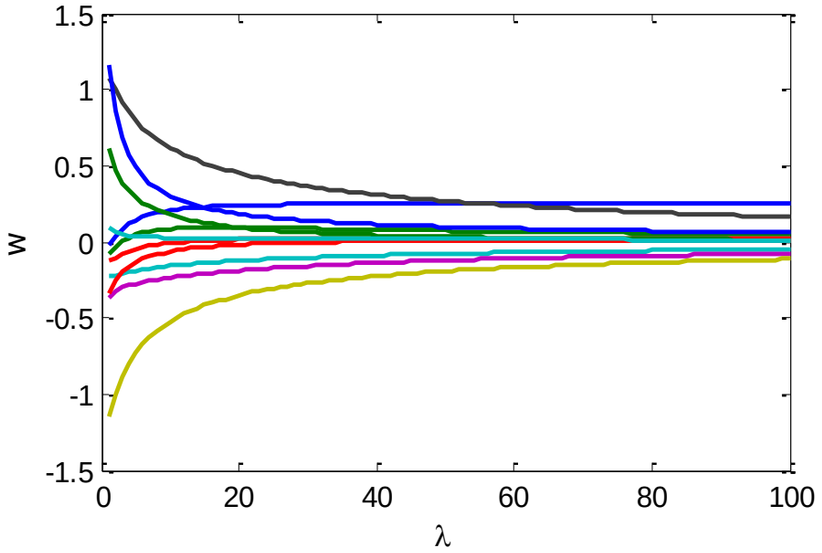
gdzie **I** jest macierzą jednostkową (z jedynkami na przekątnej) o wymiarach $n \times n$.

[†] Zapis w postaci macierzowej wymaga wcześniejszego wyeliminowania wyrazu wolnego w_0 i scentrowania atrybutów. Wtedy macierz **X** ma rozmiary $N \times n$, a **w** – $N \times 1$.

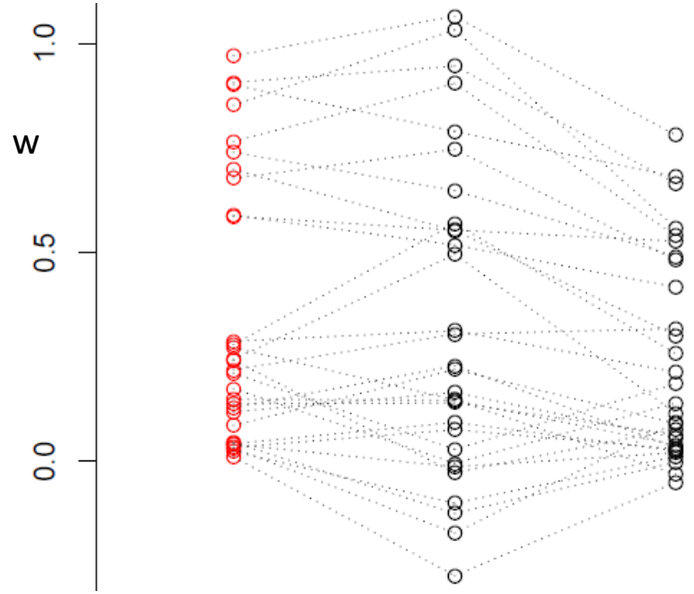
Regresja grzbietowa

Optymalną wartość parametru λ dobiera się w procedurze kroswalidacji.

Rys. Przykładowe wartości wag dla różnych wartości λ.



Rys. Wartości wag (od lewej): rzeczywiste, estymowane w regresji liniowej, estymowane w regresji grzbietowej



Wzbogacenie kryterium o karę w postaci sumy kwadratów wag nazywa się regularyzacją Tichonowa. Regularyzacja zapobiega przeuczeniu modelu (nadmiernemu dopasowaniu).

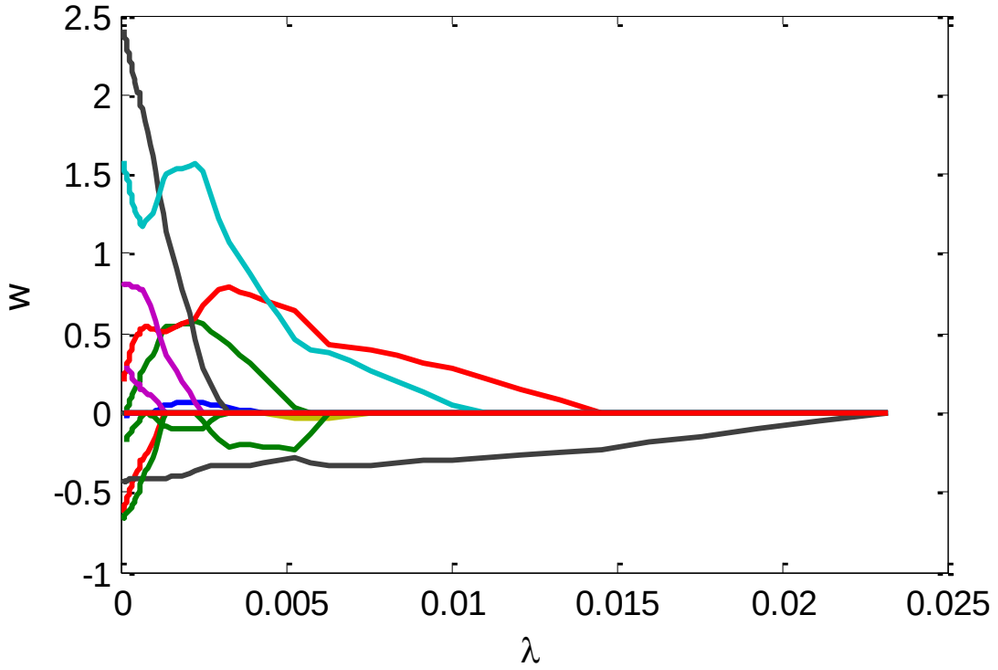
Regularyzacja pozwala zredukować błąd średniokwadratowy (MSE). Odbywa się to poprzez redukcję wariancji, chociaż obciążenie modelu wzrasta.

LASSO

LASSO (*Least Absolute Shrinkage and Selection Operator*) jest metodą regularyzacji modelu regresji liniowej, w której kryterium zawiera sumę modułów wag jako składnik kary:

$$E(h_{\mathbf{w}}) = \sum_{\mathbf{x} \in P} (y - h_{\mathbf{w}}(\mathbf{x}))^2 + \lambda \sum_{i=1}^{n} |w_i|$$

Wprowadzenie kary w postaci sumy modułów zamiast sumy kwadratów wag ma ciekawe konsekwencje. W regresji grzbietowej wagi zmniejszają się waz ze wzrostem λ, ale nigdy nie osiągają zera. W LASSO wagi mogą się zerować przy odpowiednio dużych wartościach λ. LASSO jest jednocześnie algorytmem regularyzacji i selekcji atrybutów (atrybuty z zerowymi wagami nie są uwzględniane w modelu).

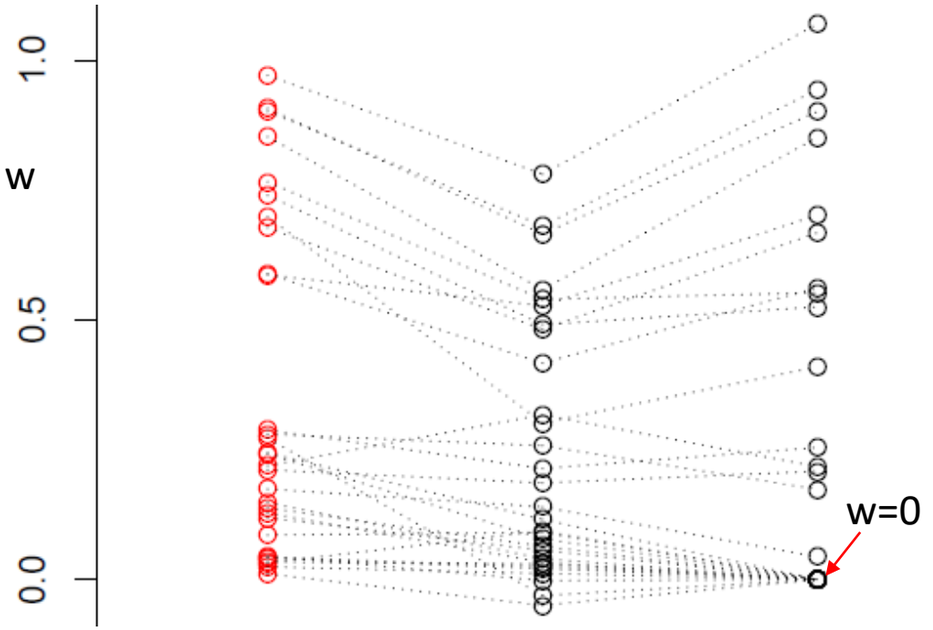


LASSO

Podobnie jak w regresji grzbietowej redukcja MSE odbywa się poprzez redukcję wariancji, chociaż obciążenie modelu wzrasta.

Wyznaczenie wartości wag minimalizujących kryterium używane w LASSO nie jest możliwe na drodze analitycznej jak w regresji grzbietowej, lecz wymaga algorytmu iteracyjnego.

Rys. Wartości wag (od lewej): rzeczywiste, estymowane w regresji grzbietowej, estymowane w LASSO



Elastyczna sieć (*elastic net*) łączy regresję grzbietową z LASSO. Składnik kary ma tutaj postać:

$$\lambda \sum_{i=1}^{n} (\alpha |w_i| + (1-\alpha) w_i^2)$$

gdzie $\alpha \in [0, 1]$. Dla $\alpha = 0$ otrzymujemy regresję grzbietową, dla $\alpha = 1$ otrzymujemy LASSO.

Wbudowana selekcja cech

Drzewa decyzyjne

W procesie konstrukcji drzewa, w każdym węźle zachodzi selekcja atrybutu do testu. Kryterium selekcji ocenia przyrost informacji wynikający z zastosowania testu opierającego się na danym atrybucie do podziału zbioru przykładów na dwa podzbiory.

Atrybuty, które ostatecznie zostają wybrane do testów, to atrybuty istotne. Atrybuty niewybrane można pominąć.

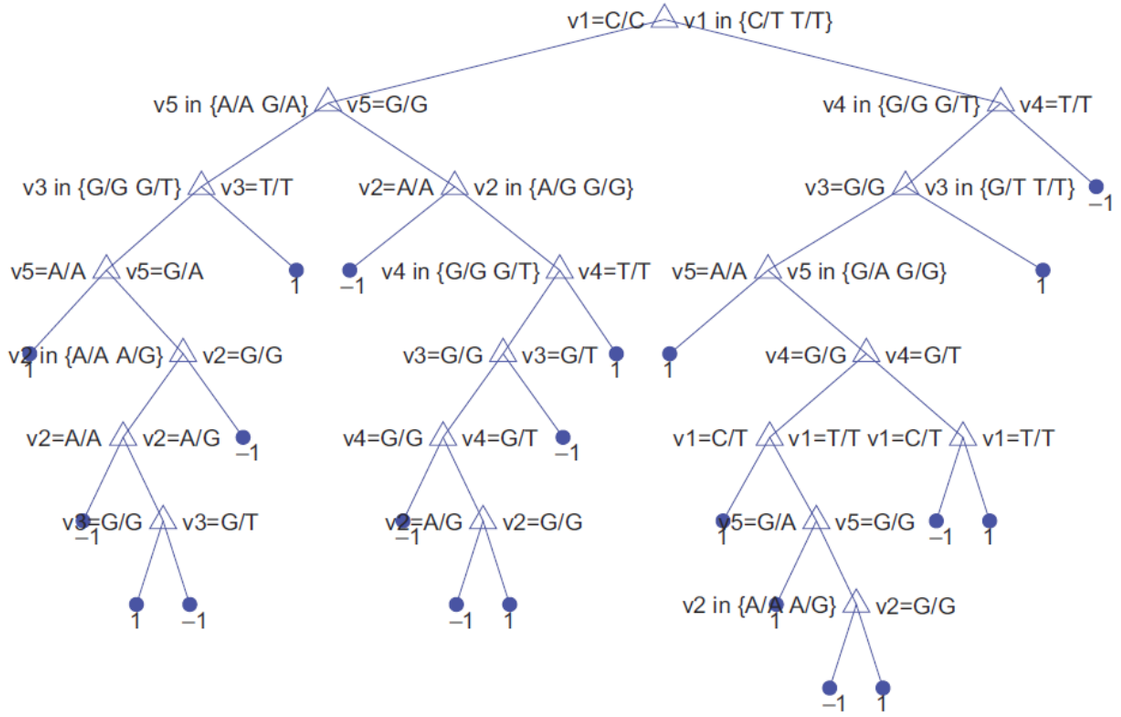


Fig 2. DT model

Wbudowana selekcja cech

Algorytm immunologiczny z lokalną selekcją cech – umożliwia selekcję różnych podzbiorów atrybutów w różnych rejonach przestrzeni wejściowej

Trening (kreacja pamięci immunologicznej)

1. Załadowanie zbioru antygenów uczących.
2. Generacja populacji początkowej przeciwciał.
3. Wykonuj dla każdego przeciwciała.
   3.1. Wykonuj do spełnienia warunku stopu (selekcja klonalna, przesz. turniejowe).
      3.1.1. Klonowanie.
      3.1.2. Hipermutacja klonalna.
      3.1.3. Obliczenie promieni krosreaktywności i powinowactw klonów do antygenów.
      3.1.4. Ocena klonów.
      3.1.5. Selekcja najlepszego klonu i zastąpienie nim przeciwciała macierzystego.
4. Apoptoza nadmiarowych przeciwciał (opcjonalnie).

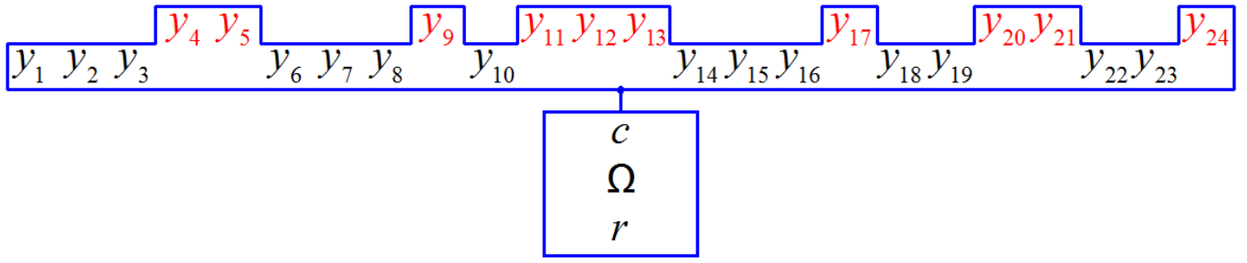
Test

5. Prezentacja antygenu testowego i obliczenie awidności przeciwciał.
6. Przypisanie antygenowi klasy o największej awidności.

Elementy rozpoznające – przeciwciała (PC) z paratopami odpowiednio ukształtowanymi w procesie treningu (dojrzewania powinowactwa)

PC jest czwórką $\langle \mathbf{y}, c, \Omega, r \rangle$

- $\mathbf{y} = [y_1, y_2, \ldots, y_n]$ – wektor atrybutów
- $c \in \{1, 2, \ldots, C\}$ – numer klasy
- Ω – paratop, zbiór atrybutów istotnych (wybieranych adaptacyjnie)
- $r \geq 0$ – próg krosreaktywności (dobierany adaptacyjnie) definiujący obszar recepcyjny PC w $\mathbb{R}^{|\Omega|}$ (hiperkula o środku w $\mathbf{y}' = [y_i]$, $i \in \Omega$ i promieniu r)



⇒ Dudek G.: **Artificial immune system for classification with local feature selection**. IEEE Transactions on Evolutionary Computation, vol. 16, issue 6, pp. 847-860, 2012.