SYSTEMY UCZĄCE SIĘ

WYKŁAD 10. PRZEKSZTAŁCANIE ATRYBUTÓW

Dr hab. inż. Grzegorz Dudek
Wydział Elektryczny
Politechnika Częstochowska

Częstochowa 2014

INFORMACJE WSTĘPNE

Hipotezy do uczenia się lub tworzenia pojęć są funkcjami zależnymi od wartości atrybutów. Przekształcenia atrybutów mają na celu poprawienie jakości hipotez uzyskiwanych przez algorytmy indukcyjnego uczenia się.

Na skutek przekształcenia atrybutów uczeń nie zyskuje żadnej nowej informacji o przykładach (część informacji może nawet stracić), lecz może lepiej reprezentować wiedzę. Ten sam algorytm uczenia się, działając na przekształconych atrybutach może tworzyć dokładniejsze i mniej złożone hipotezy.

Metody przekształcania atrybutów:

- przeskalowanie, standaryzacja, normalizacja
- dyskretyzacja atrybutów
- selekcja atrybutów (usuwanie)
- dodawanie atrybutów (jako funkcji atrybutów oryginalnych*)
- ekstrakcja atrybutów

* patrz rozszerzona reprezentacja – wykład 6.

SKALOWANIE, STANDARYZACJA

- Przeskalowanie atrybutu do zadanego przedziału:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}(b - a) + a$$

gdzie: [a, b] – nowy przedział atrybutu (np. [–1, 1], [0, 1]), [$x_{\min}$, $x_{\max}$] – stary przedział atrybutu.

Np. jeśli x = 736, przedział zmienności x: [12, 980], nowy przedział: [0, 1], to x' = 0,716

- Standaryzacja atrybutu:

$$x' = \frac{x - \bar{x}}{\sigma_x}$$

gdzie: $\bar{x}$ – wartość średnia atrybutu x, σ_x – odchylenie standardowe atrybutu x.

Np. jeśli x = 736, $\bar{x}$ = 486, σ_x = 58, to x' = 4,31

Po standaryzacji wartość średnia atrybutu wynosi 0, a jego odchylenie standardowe wynosi 1.

NORMALIZACJA

- Normalizacja przykładu **x** (wektora):

$$\mathbf{x}' = \frac{\mathbf{x}}{\sqrt{\sum_{i=1}^{n} x_i^2}}$$

Po normalizacji długość wektora jest równa 1.

Np. **x** = [–4, 2, 1, –12] po normalizacji: **x**'= [–0.3114, 0.1557, 0.0778, –0.9342]

Inny sposób normalizacji:

$$\mathbf{x}' = \frac{\mathbf{x} - \bar{x}_w}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x}_w)^2}}$$

gdzie: $\bar{x}_w$ – wartość średnia składowych wektora.

Po normalizacji długość wektora jest równa 1, jego średnia wynosi zero, a odchylenia standardowe wszystkich unormowanych wektorów są jednakowe.

Np. **x** = [–4, 2, 1, –12] po normalizacji: **x**'= [–0.0677, 0.4739, 0.3836 –0.7898]

DYSKRETYZACJA ATRYBUTÓW

Dyskretyzacja atrybutów ciągłych polega na zastąpieniu każdego z nich atrybutem o wartościach dyskretnych, odpowiadających pewnym przedziałom ciągłych wartości oryginalnego atrybutu.

Dyskretyzacja atrybutów ciągłych pozwala zastosować algorytm przystosowany do atrybutów dyskretnych (np. algorytm uczenia się reguł) lub uprościć algorytm działający na atrybutach ciągłych (np. drzewo decyzyjne).

Odpowiednikiem dyskretyzacji dla atrybutów porządkowych jest ich agregacja. Agregacja atrybutu porządkowego polega na jego zastąpieniu innym atrybutem porządkowym o mniejszej liczbie wartości.

Korzyści z dyskretyzacji:

- poprawa efektywności obliczeniowej procesu uczenia się: zastąpienie wielu wartości atrybutu ciągłego niewielką liczbą wartości zdyskretyzowanych może znacznie obniżyć nakłady obliczeń,
- zwiększenie prostoty i czytelności hipotez: hipotezy bezpośrednio wykorzystujące atrybuty ciągłe mogą być złożone i nieczytelne, dyskretyzacja zaś może prowadzić do hipotez prostszych i łatwiejszych do interpretacji,
- poprawa dokładności hipotez: dyskretyzacja atrybutów ciągłych może korzystnie wpływać na dokładność generowanych hipotez, zwłaszcza w przypadku nie w pełni poprawnych danych trenujących, ponieważ jest pewnym sposobem zapobiegania nadmiernemu dopasowaniu.

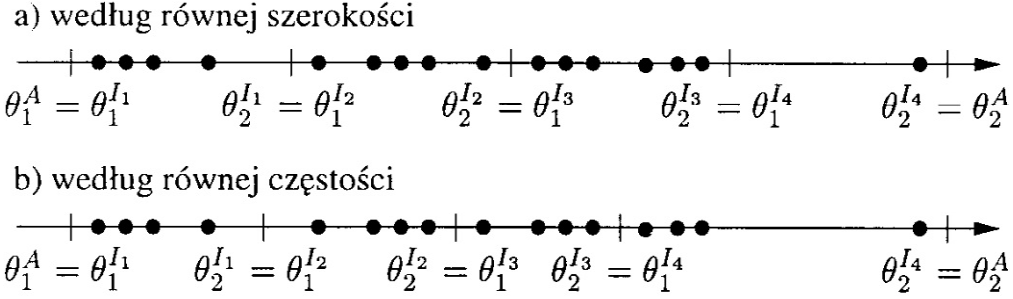
DYSKRETYZACJA ATRYBUTÓW

Podziały metod dyskretyzacji:

- metody prymitywne i zaawansowane (nieuwzględniające/uwzględniające rozkład wartości atrybutów i klas)
- metody globalne i lokalne (dyskretyzacja jednakowa w całej dziedzinie/dyskretyzacja różna w różnych obszarach dziedziny)
- metody bez nadzoru i z nadzorem (etykiety klas przykładów są nieznane lub nieuwzględniane podczas dyskretyzacji/etykiety klas są znane i uwzględniane podczas dyskretyzacji)

Prymitywne metody dyskretyzacji:

- dyskretyzacja według równej szerokości (równe długości przedziałów, każdemu przedziałowi przypisywana jest jedna wartość dyskretna)
- dyskretyzacja według równej częstości (jednakowa lub zbliżona liczba przykładów w każdym przedziale)



Rys. 7.1. Ilustracja prymitywnych metod dyskretyzacji

SELEKCJA ATRYBUTÓW

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

Zysk z selekcji atrybutów:

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

Metody selekcji atrybutów:

- filtracyjne – atrybuty do usunięcia określa się na podstawie analizy danych zawartych w zbiorze trenującym (usuniecie 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 ucznia hipotez (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 element modelu, np. drzewa decyzyjne, indukcja reguł

[†] patrz krokowe dodawanie i eliminacja atrybutów w regresji krokowej – wykład 6.

EKSTRAKCJA ATRYBUTÓW

Ekstrakcja atrybutów – wyznaczenie nowych atrybutów, które są funkcjami atrybutów oryginalnych.

Ekstrakcja atrybutów podobnie jak selekcja atrybutów pozwala zmniejszyć wymiarowość problemu (liczba nowych atrybutów zwykle jest mniejsza niż liczba atrybutów oryginalnych). Często okazuje się, że mniejsza liczba nowych atrybutów zawiera nieomal tyle samo informacji o strukturze danych, co atrybuty oryginalne.

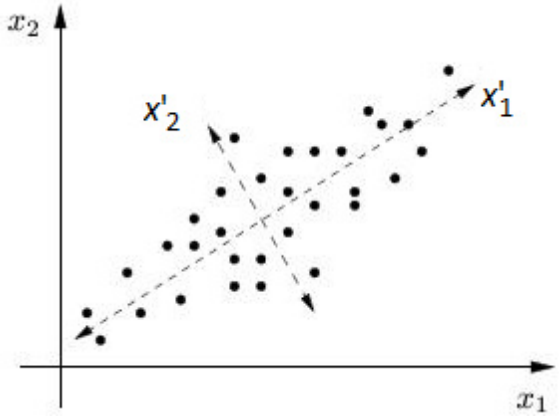
Najpopularniejszą metodą ekstrakcji atrybutów jest analiza głównych składowych (*principal component analysis* PCA).

PCA na podstawie wektorów oryginalnych **x** tworzy nowe wektory **x**' o składowych wzajemnie nieskorelowanych (tzw. składowe główne). Składowe główne wyjaśniają w maksymalnym stopniu całkowitą wariancję składowych oryginalnych. Składowe wektora **x**' są liniowymi kombinacjami składowych wektora **x** (zakłada się, że wektory **x** mają zerową średnią, jeśli to nie jest spełnione, wektory **x** należy przekształcić, odejmując od każdej składowej wartość średnią wektora):

$$x'_{i,j} = \sum_{l=1}^{n} a_{j,l} x_{i,l} = \mathbf{a}_j^{\mathsf{T}} \mathbf{x}_i, \quad j = 1,2,...,n$$

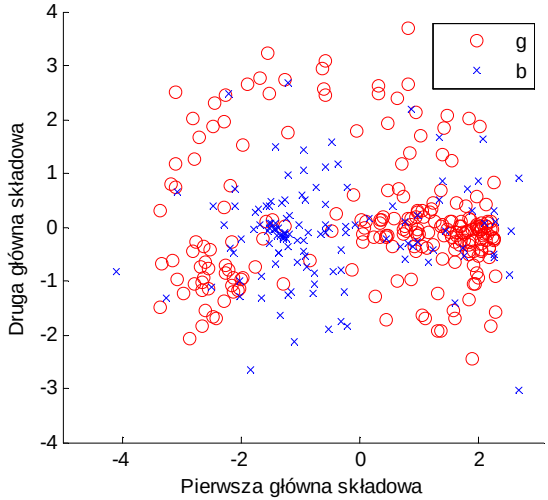
gdzie wektor współczynników $\mathbf{a}_j$ jest wektorem charakterystycznym odpowiadającym kolejnym największym wartościom własnym macierzy kowariancji z próby $\mathbf{x}_1$, $\mathbf{x}_2$, .., $\mathbf{x}_N$. Składowe główne są unormowane i wzajemnie ortogonalne, tj. $\mathbf{a}_j^{\mathsf{T}} \mathbf{a}_j = 1$ i $\mathbf{a}_j^{\mathsf{T}} \mathbf{a}_k = 0$ dla wszystkich $j \neq k$.

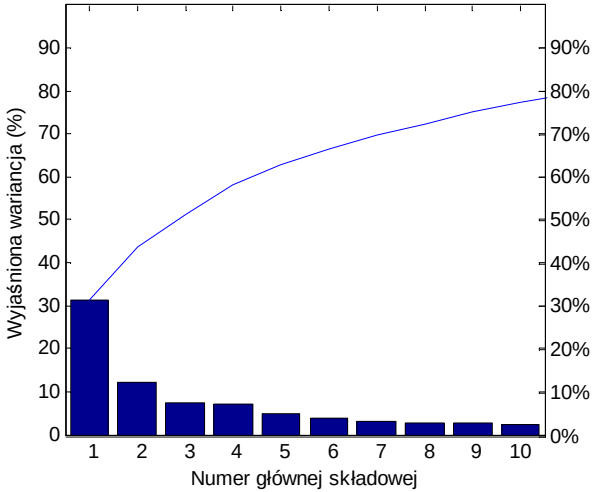
Pierwsza główna składowa $x'_{i,1}$ wykazuje największą wariancję, kolejne główne składowe mają wariancje coraz mniejsze. Ponieważ najwięcej informacji przenoszą początkowe składowe, końcowe składowe o najmniejszych wariancjach (często nieprzekraczających szumu pomiarowego) można odrzucić.



ANALIZA GŁÓWNYCH SKŁADOWYCH – PRZYKŁAD

Wyznacz główne składowe zbioru danych Ionosphere[‡] (351 przykładów, 34 atrybuty, dwie klasy: g i b). Przedstaw dwie pierwsze składowe na wykresie. Przedstaw wariancje wyjaśniane przez kolejne składowe.





[‡] patrz ćwiczenie lab. pt.: Klasyfikacja danych za pomocą drzew decyzyjnych