Zadanie klasyfikacji - drzewa decyzyjne, wykł. 5

Joanna Jędrzejowicz

Instytut Informatyki

Klasyfikacja danych - przykłady

- Bankowość- czy dane podanie o udzielenie kredytu jest obarczone dużym czy małym ryzykiem, czy dana transakcja kartą jest oszustwem,
- Edukacja - umieszczenie studenta w odpowiedniej grupie z uwzględnieniem jego potrzeb,
- Medycyna - diagnozowanie, czy występuje dana choroba,
- Telekomunikacja - czy dany klient zrezygnuje z usług firmy telekomunikacyjnej (zbiór churn)
- Informatyka - czy wiadomosc pocztowa jest spamem

W zadaniu klasyfikacji jest ustalona zmienna celu (inaczej: klasa) np. w powyższych przykładach: ryzyko duże - tak lub nie, grupa studencka o numerach $1, \ldots, 5$, diagnoza - jest choroba- tak lub nie itd.

Na czym polega zadanie klasyfikacji danych

Niech C oznacza zbiór klas. Zakładamy, że algorytm uczący się klasyfikacji otrzymuje, jako wejście, zbiór danych treningowych (uczących)

$$LD = \{< d, c > \mid d \in D, c \in C\} \subset D \times C,$$

gdzie D jest zbiorem rekordów (wierszy atrybutów)
$d = (w_1^d, \ldots, w_n^d)$ gdzie w_i^d dla $i = 1, \ldots, n$ jest wartością i-tego atrybutu, n - jest liczbą atrybutów.

Na czym polega zadanie klasyfikacji danych, cd

Algorytm uczenia klasyfikatora służy do znalezienia najlepszej możliwej aproksymacji $\bar{f}$ nieznanej funkcji f takiej, że $f(d) = c$.

Następnie $\bar{f}$ może być użyta do znalezienia klasy $\bar{c} = \bar{f}(\bar{d})$ dla dowolnego $\bar{d}$ takiego, że $(\bar{d}, \bar{c}) \notin LD$, zatem algorytm będzie mógł być użyty do znajdowania klasy dla przykładów, których nie użyto w uczeniu.
Zbiór danych LD składa się z dwóch podzbiorów $LD = TD \cup TS$, gdzie TD-jest zbiorem treningowym, TS - zbiorem testowym.

Dokładność klasyfikacji

Jak liczyć dokładność klasyfikatora:

- budujemy klasyfikator w oparciu o zbiór treningowy - dla każdego rekordu określona jest jego klasa,
- używając zbudowanego klasyfikatora, dla każdego wiersza ze zbioru testowego wyznaczamy klasę oraz porównujemy z rzeczywistą klasą,
- dokładnośc określamy jako stosunek liczby poprawnie sklasyfikowanych wierszy ze zbioru testowego do mocy tego zbioru.

Ewaluacja- cross-validation - sprawdzian krzyżowy

Ewaluacja metodą '10-fold cross-validation' polega na dokonaniu podziału zbioru danych na 10 cześci $X_1, \ldots, X_{10}$ i wykonaniu 10 doświadczeń.

W i-tym doświadczeniu X_i jest zbiorem testowym, a pozostałe $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_{10}$ podzbiorów stanowi zbiór treningowy; w każdym doświadczeniu oblicza się dokładność klasyfikatora, a następnie liczy się średnią dokładność.

Na czym polega zadanie klasyfikacji danych, cd

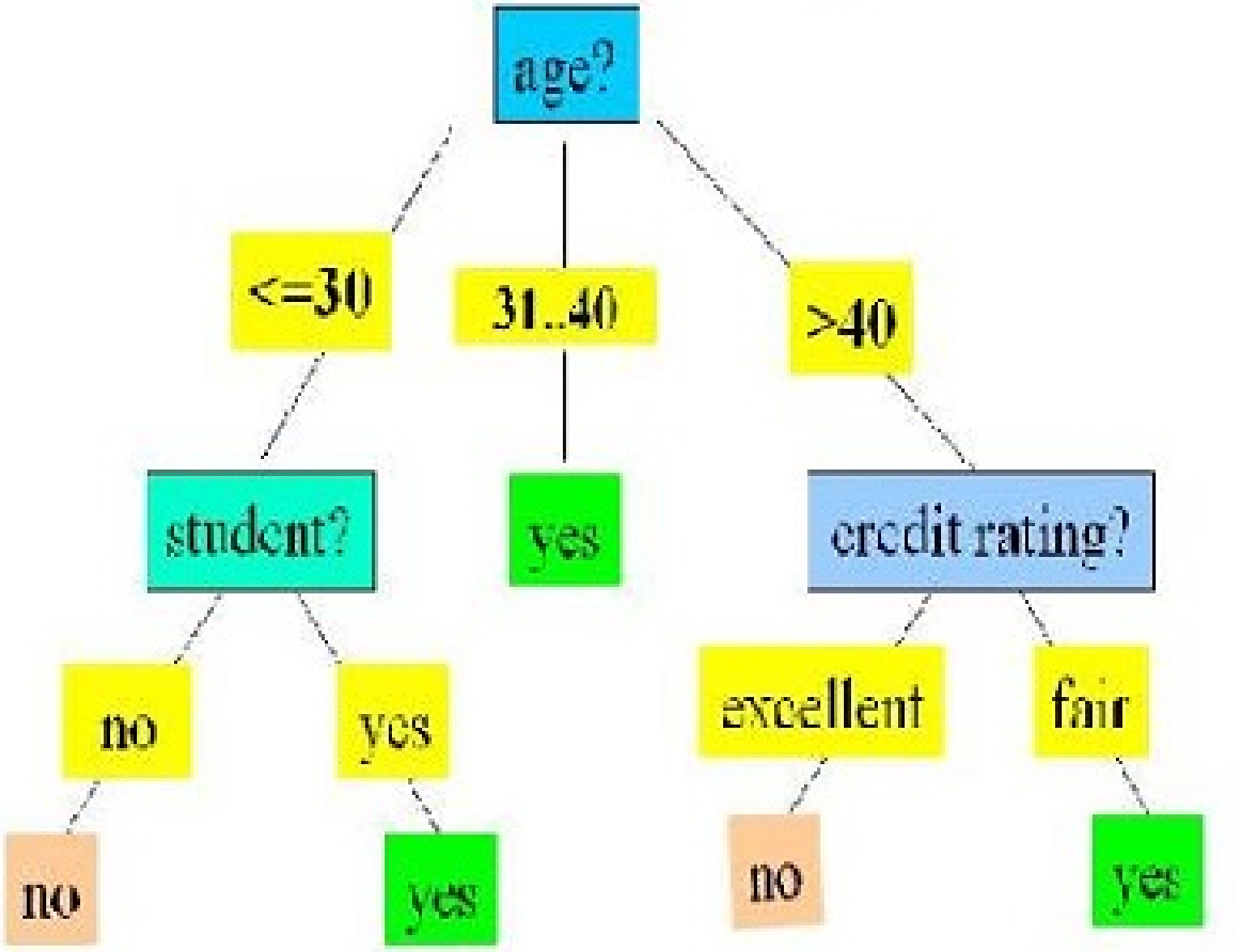
Klasyfikacja jest nazywana uczeniem nadzorowanym, gdyż algorytm uczący ma w zbiorze treningowym pełne dane - informacje, do jakiej klasy należy każdy wektor danych.

Przewidywanie (predykcja) jest podobnym zadaniem do klasyfikacji- w przewidywaniu wynik dotyczy przyszłości, np. przewidywanie cen akcji po upływie pół roku, przewidywanie zwycięzcy rozgrywek sportowych na podstawie wyników z przeszłości itd. Metody i techniki wykorzystywane do klasyfikacji mogą być również użyte, przy poczynieniu dodatkowych założeń, do przewidywania.

Przykładowy zbiór danych 'czy kupi komputer?'

wiek	dochod	stud	zdolKred	czyKupi
≤ 30	duży	nie	umiark	nie
≤ 30	duży	nie	doskonala	nie
$31 \ldots 40$	duży	nie	umiark	tak
> 40	sredni	nie	umiark	tak
> 40	mały	tak	umiark	tak
> 40	mały	tak	doskonala	nie
$31 \ldots 40$	mały	tak	doskonała	tak
≤ 30	sredni	nie	umiark	nie
≤ 30	mały	tak	umiark	tak
> 40	sredni	tak	umiark	tak
≤ 30	sredni	tak	doskonala	tak
$31 \ldots 40$	sredni	nie	doskonala	tak
$31 \ldots 40$	duży	tak	umiark	tak
> 40	sredni	nie	doskonala	nie

Drzewo decyzyjne



Jak tworzy się drzewo decyzyjne?

- Z korzeniem drzewa związany jest pełny zbiór danych
- Każdy wierzcholek różny od liścia jest węzłem decyzyjnym
- Z każdem węzłem decyzyjnym związany jest atrybut
- Zbiór danych jest dzielony zgodnie z wartościami tego atrybutu i gałęzie prowadzą do następnych węzłow odpowiadających możliwym wartościom atrybutu
- Jeżeli z węzłem jest związany zbiór danych z jednej klasy, to węzeł jest liściem etykietowanym tą klasą
- Jeżeli w węźle występują dane z takimi samymi wartościami wszystkich atrybutów (nie można dalej dzielic!), to przyjmuje się głosowanie większościowe

Kiedy można zastosować drzewo decyzyjne

- Zbiór treningowy powinien być bogaty i różnorodny, zapewniający reprezentatywną grupę typów rekordów, którch klasyfikacja może być potrzebna w przyszłości,
- klasy zmiennej celu muszą być dyskretne, to znaczy nie można zastosować analizy drzew decyzyjnych do ciągłej zmiennej celu,
- drzewa decyzyjne starają się stworzyć zbiór liści, które są 'najczystsze', to znaczy gdy każdy z rekordów w danym liściu należy do tej samej klasy.

Uwaga: jeżeli wartości atrybutu są rzeczywiste, to trzeba podzielić zbiór wartości na pewną liczbę przedziałów

Jakie mogą się pojawić problemy

Co zrobić, jeżeli dany węzeł zawiera różne rekordy, z niejednakową wartością zmiennej celu i nie można dokonać dalszego podziału?

id	wiek	dochod	czyKupi
20	>40	duzy	tak
21	>40	duzy	tak
22	>40	duzy	nie
23	>40	duzy	tak

Jak dokonywać wyboru atrybutów w poszczególnych węzłach- algorytm C4.5

Dla każdego atrybutu liczy się zysk informacji związany z tym atrybutem i wybiera się atrybut, dla którego wartość jest największa.
Załóżmy, że (w ustalonym węźle) zbiór danych S składa się z s wierszy. Załóżmy, że zbiór klas (wartości zmiennej celu) składa się z m różnych elementów $C_1, \ldots, C_m$. Niech s_i oznacza liczbę wierszy z klasy C_i.
Ilosć informacji niezbędna do zaklasyfikowania rekordu z danego zbioru:

$$I(s_1, \ldots, s_m) = -\sum_{i=1}^{m} p_i \times \log p_i$$

gdzie

$$p_i = \frac{s_i}{s}$$

Jak dokonywać wyboru atrybutów w poszczególnych węzłach,cd

Jeżeli A jest atrybutem, który przyjmuje v możliwych wartości $a_1, \ldots, a_v$, to atrybut A dzieli zbiór S na v podzbiorów $S_1, \ldots, S_v$. Niech s_{ij} oznacza liczbę wierszy z klasy C_i należących do podzbioru S_j (czyli wartość atrybutu A jest w tym podzbiorze równa a_j).
Entropia

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \cdots + s_{mj}}{s} \times I(s_{1j}, \ldots, s_{mj})$$

$$I(s_{1j}, \ldots, s_{mj}) = -\sum_{i=1}^{m} p_{ij} \times \log p_{ij}$$

gdzie

$$p_{ij} = \frac{s_{ij}}{|S_j|}$$

Jak dokonywać wyboru atrybutów w poszczególnych węzłach,cd

Zysk informacji dla atrybutu A:

$$Gain(A) = I(s_1, \ldots, s_m) - E(A)$$

Obliczenia dla przykładowych danych

Mamy 4 atrybuty:

- wiek (z trzema możliwymi wartościami: '<=30', '31..40','>40'),
- dochód (z trzema możliwymi wartościami)
- student(tak lub nie)
- zdolKred (dwie wartości)

Przyjmijmy klasa C_1 ='tak', klasa C_2='nie'

Wyliczenie zysku informacji dla atrybutu 'wiek'

$$I(s_1, s_2) = -\frac{9}{14}\log\frac{9}{14} - \frac{5}{14}\log\frac{5}{14} = 0,940$$

Zysk informacji dla atrybutu 'wiek'

Dla wartości '<=30':

$$s_{11} = 2,\ s_{21} = 3,\ I(s_{11}, s_{21}) = -\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5} = 0,971$$

Dla wartości '31..40':

$$s_{12} = 4,\ s_{22} = 0,\ I(s_{12}, s_{22}) = 0$$

Dla wartości '>=40':

$$s_{13} = 3,\ s_{23} = 2,\ I(s_{13}, s_{23}) = -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} = 0,971$$

Zysk informacji dla atrybutu 'wiek', cd

$$E(wiek) = \frac{5}{14} \times I(s_{11}, s_{21}) + \frac{4}{14} \times I(s_{12}, s_{22}) + \frac{5}{14} \times I(s_{13}, s_{23}) = 0,694$$

$$Gain(wiek) = I(s_1, s_2) - E(wiek) = 0,940 - 0,694 = 0,246$$

Zadanie Policzyć zysk informacji dla pozostałych atrybutów. Sprawdzić, że największa wartość jest osiągana dla atrybutu 'wiek'. Narysować drzewo decyzyjne (wyklad5.xls)

Błędy popełniane przez modele klasyfikacyjne

- błąd treningowy klasyfikatora - stosunek niepoprawnie zaklasyfikowanych rekordów zbioru treningowego do łacznej liczby rekordów w zbiorze treningowym,
- błąd testowy klasyfikatora - jak wyżej dla zbior testowego,
- błąd generalizacji - oczekiwany błąd na zbiorze nowych rekordów, zakłada się że minimalizacja błędu testowego prowadzi do minimalizacji błędu generalizacji

Przykładowy zbiór testowy

wiek	dochod	stud	zdolKred	czyKupi
≤ 30	mały	tak	umiark	tak
> 40	duży	nie	doskonała	tak
$31 \cdots 40$	duży	tak	umiark	tak

Jaka jest błąd testowy, jaki treningowy?

Przeuczenie klasyfikatora

Problem silnego dopasowania drzewa decyzyjnego (por. przykład):

- błąd treningowy mały,
- dane mogą być mało reprezentatywne,
- może powodować duży błąd generalizacji

Przycinanie drzewa decyzyjnego (ang. pruning)

Usunięcie poddrzewa i zastąpienie liściem, któremu przypisuje się etykietą dominującą w zbiorze rekordów związanych z wierzcholkiem.

Przykład. Odcinamy wierzchołek zdolKred i zastepujemy liściem 'tak'. Jak zmieni się błąd treningowy i testowy?

Przycinanie drzewa decyzyjnego

Metody wykorzystujace strategię przycinania wstępnego: na etapie konstrukcji drzewa zatrzymuje sie tworzenie nowych węzłów korzystając z nast. kryteriów:

- miara jakosci podziału zbioru treningowego (np. zysk informacji) jest poniżej zadanego progu,
- zbiór treningowy związany z danym wierzchołkiem nie jest dostatecznie liczny,
- rozkład klas (np. większosc rekordów należy do jednej klasy)

Przycinanie drzewa decyzyjnego, cd

Metody wykorzystujace strategię przycinania końcowego: konstruuje się pełne drzewo, a następnie metodą wstępującą, poczynając od lisci i przesuwając się w kierunku korzenia dokonuje się przycięcia. Możliwe kryteria:

- wylicza się błąd klasyfikacji dla drzewa zredukowanego i porównuje z miarą dla drzewa niezredukowanego, jeżeli przycięcie pogarsza to przywraca się zredukowane poddrzewo
- ewentualnie kryteria jak dla przycinania wstępnego

Reguły decyzyjne

Korzystając z drzewa decyzyjnego możemy wygenerować reguły decyzyjne - wystarczy zapisać informację towrzyszącą przejściu od korzenia do liścia, np.
if wiek $< 40 \wedge$ wiek $>= 31$ **then** czyKupi=**tak**
Dla reguł określa się dwie miary: wsparcie (ang. support) oraz ufność (ang. confidence).
Dla reguły **if** A **then** B

- wsparcie: stosunek liczby rekordów zawierających A i B do liczby wszystkich rekordów,
- ufność: stosunek liczby rekordów zawierających A i B do liczby rekordów zawierających A.
- Jakie jest wsparcie i ufność dla powyższej reguły?

wsparcie $\frac{4}{14}$, ufność $\frac{4}{4}$

Metoda klasyfikacji Random Forest (Breiman 2001)

Klasyfikator zespołowy wykorzystujący drzewa decyzyjne oraz następujące dwie zasady:

- 'bagging' - do budowy kolejnych drzew losuje się się dane ze zbioru treningowego,
- przy generowaniu konkretnego drzewa, dla każdego wierzchołka losuje się M atrybutów i sposród nich wybiera jeden z największym zyskiem informacji (M jest parametrem).
- po skonstruowaniu drzew klasyfikuje się nowe przykłady stosując głosowanie większościowe.