Algorytmy eksploracji danych: Wykład 7

Wojciech Kempa

Politechnika Śląska
Wydział Matematyki Stosowanej

Wprowadzenie do klasyfikacji

- Załóżmy, że dysponujemy pewnym zbiorem obiektów opisanych za pomocą wielu atrybutów, którego elementy zostały przyporządkowane do pewnej liczby klas.

Wprowadzenie do klasyfikacji

- Załóżmy, że dysponujemy pewnym zbiorem obiektów opisanych za pomocą wielu atrybutów, którego elementy zostały przyporządkowane do pewnej liczby klas.
- Do której z tych klas należy przyporządkować nowy obiekt opisany za pomocą tych samych atrybutów?

Wprowadzenie do klasyfikacji

- Załóżmy, że dysponujemy pewnym zbiorem obiektów opisanych za pomocą wielu atrybutów, którego elementy zostały przyporządkowane do pewnej liczby klas.
- Do której z tych klas należy przyporządkować nowy obiekt opisany za pomocą tych samych atrybutów?
- Jak sprawdzić, czy stworzona w ten sposób reguła klasyfikacyjna jest dokładna i sprawdzi się w przyszłości w przypadku większej liczby obiektów?

Wprowadzenie do klasyfikacji

- Załóżmy, że dysponujemy pewnym zbiorem obiektów opisanych za pomocą wielu atrybutów, którego elementy zostały przyporządkowane do pewnej liczby klas.
- Do której z tych klas należy przyporządkować nowy obiekt opisany za pomocą tych samych atrybutów?
- Jak sprawdzić, czy stworzona w ten sposób reguła klasyfikacyjna jest dokładna i sprawdzi się w przyszłości w przypadku większej liczby obiektów?
- Odpowiedzi na te i podobne pytania są domeną jednej z dziedzin eksploracji danych zwanej **klasyfikacją**.

Wprowadzenie do klasyfikacji

- Załóżmy, że dysponujemy pewnym zbiorem obiektów opisanych za pomocą wielu atrybutów, którego elementy zostały przyporządkowane do pewnej liczby klas.
- Do której z tych klas należy przyporządkować nowy obiekt opisany za pomocą tych samych atrybutów?
- Jak sprawdzić, czy stworzona w ten sposób reguła klasyfikacyjna jest dokładna i sprawdzi się w przyszłości w przypadku większej liczby obiektów?
- Odpowiedzi na te i podobne pytania są domeną jednej z dziedzin eksploracji danych zwanej **klasyfikacją**.
- Zasadniczym celem jest tu zbudowanie ogólnego modelu klasyfikacyjnego zbioru obiektów na podstawie danych historycznych, a następnie zastosowanie go do predykcji klasy nowego obiektu.

Atrybuty warunkowe i atrybut decyzyjny

- Dane wejściowe stanowi pewien zbiór obiektów (obserwacji itp.) D zwany **zbiorem treningowym**, którego elementy (obiekty) opisane są za pomocą pewnej skończonej liczby tzw. **atrybutów warunkowych** $A_1, A_2, \ldots, A_s$ i jednego tzw. **atrybutu decyzyjnego** C określającego przyporządkowanie obiektu do danej klasy.

Atrybuty warunkowe i atrybut decyzyjny

- Dane wejściowe stanowi pewien zbiór obiektów (obserwacji itp.) D zwany **zbiorem treningowym**, którego elementy (obiekty) opisane są za pomocą pewnej skończonej liczby tzw. **atrybutów warunkowych** $A_1, A_2, \ldots, A_s$ i jednego tzw. **atrybutu decyzyjnego** C określającego przyporządkowanie obiektu do danej klasy.
- Atrybut decyzyjny jest atrybutem kategorycznym, jego wartości $C_1, \ldots, C_m$, $m \geq 2$, to tzw. **klasy decyzyjne**.

Atrybuty warunkowe i atrybut decyzyjny

- Dane wejściowe stanowi pewien zbiór obiektów (obserwacji itp.) D zwany **zbiorem treningowym**, którego elementy (obiekty) opisane są za pomocą pewnej skończonej liczby tzw. **atrybutów warunkowych** $A_1, A_2, \ldots, A_s$ i jednego tzw. **atrybutu decyzyjnego** C określającego przyporządkowanie obiektu do danej klasy.

- Atrybut decyzyjny jest atrybutem kategorycznym, jego wartości $C_1, \ldots, C_m$, $m \geq 2$, to tzw. **klasy decyzyjne**.
- Atrybuty warunkowe mogą być atrybutami ilościowymi, binarnymi lub też jakościowymi, określonymi za pomocą skali porządkowej lub nominalnej.

Klasyfikator

- Celem procesu klasyfikacji jest znalezienie **funkcji klasyfikującej (klasyfikatora)**, która pozwoli na jednoznaczne przyporządkowanie nowego obiektu $X = (A_1 = x_1, \ldots, A_s = x_s)$ do jednej z m klas decyzyjnych $C_1, \ldots, C_m$.

Klasyfikator

- Celem procesu klasyfikacji jest znalezienie **funkcji klasyfikującej (klasyfikatora)**, która pozwoli na jednoznaczne przyporządkowanie nowego obiektu $X = (A_1 = x_1, \ldots, A_s = x_s)$ do jednej z m klas decyzyjnych $C_1, \ldots, C_m$.

- Konstrukcja klasyfikatora odbywa się z istotnym wykorzystaniem zbioru treningowego, czyli zbioru obiektów, które **już są przyporządkowane do odpowiednich klas decyzyjnych**.

Przykład 1 (1)

Jako przykład rozważmy zbiór treningowy D opisujący klientów pewnego banku (mężczyzn), w którym atrybutami warunkowymi są

- A_1: *Wiek* – atrybut ilościowy ciągły;
- A_2: *Status* – atrybut jakościowy dany w skali nominalnej (Kawaler, Żonaty, Rozwiedziony);
- A_3: *Dochód* – atrybut jakościowy dany w skali porządkowej (Niski, Średni, Wysoki);
- A_4: *Liczba dzieci* – atrybut ilościowy dyskretny.

Atrybut decyzyjny C jest związany z pozytywnym bądź negatywnym zaopiniowaniem wniosku kredytowego klienta. Wyróżniamy w nim dwie klasy związane z poziomem ryzyka:

- C_1: Wysokie;
- C_2: Niskie.

Przykład 1 (2)

Baza danych D może mieć następującą postać:

L.p.	*Wiek* (A_1)	*Status* (A_2)	*Dochód* (A_3)	*Dzieci* (A_4)	*Ryzyko* (C)
1	26	Kawaler	Niski	0	Wysokie
2	35	Żonaty	Średni	1	Niskie
3	36	Żonaty	Wysoki	2	Niskie
4	49	Rozwiedziony	Średni	2	Niskie
5	27	Żonaty	Niski	1	Wysokie
6	39	Rozwiedziony	Wysoki	0	Wysokie
7	31	Żonaty	Niski	2	Wysokie
8	50	Żonaty	Wysoki	2	Niskie
9	51	Kawaler	Średni	1	Niskie
10	38	Rozwiedziony	Niski	2	Wysokie
11	29	Kawaler	Wysoki	0	Niskie
12	42	Rozwiedziony	Średni	4	Wysokie
13	41	Żonaty	Średni	1	Niskie
14	56	Rozwiedziony	Wysoki	2	Niskie
15	32	Kawaler	Średni	0	Niskie
16	45	Żonaty	Średni	4	Wysokie

Tablica 1: Przykładowy zbiór treningowy D (źródło: T. Morzy *Eksploracja danych*)

Konstrukcja modelu klasyfikacyjnego (1)

Tak naprawdę, w praktyce, na początku procesu klasyfikacyjnego z reguły całą dostępną bazę danych dzielimy na zbiór treningowy, na podstawie którego konstruowany będzie klasyfikator, oraz **zbiór testowy**, który posłuży do weryfikacji jakości zbudowanego klasyfikatora. Konstrukcja modelu klasyfikacyjnego przebiega zatem dwuetapowo, obejmując następujące dwa kroki:

- **krok uczenia (treningu)**, w którym z wykorzystaniem zbioru treningowego konstruowany jest odpowiedni klasyfikator (za pomocą tzw. **algorytmu uczącego**);
- **krok testowania**, w którym weryfikowana jest jakość zbudowanego modelu na podstawie bazy danych stanowiących zbiór testowy.

Konstrukcja modelu klasyfikacyjnego (2)

W praktyce istnieje wiele różnych metod (algorytmów uczących) konstruowania klasyfikatorów. Najważniejsze z nich to m.in.

- drzewa klasyfikacyjne (decyzyjne);
- klasyfikatory bayesowskie;
- reguły klasyfikacyjne typu "if–then–else";
- sztuczne sieci neuronowe.

Wśród metod wykorzystywanych w klasyfikacji wyróżnia się też takie, w których klasyfikacja przebiega „na bieżąco", tzn. bez konstruowania modelu klasyfikacyjnego. Najważniejszą z takich metod jest **metoda k najbliższych „sąsiadów"** oznaczana w skrócie jako **metoda k-NN** (ang. *k-Nearest Neighbors*).

Drzewa decyzyjne (1)

- Indukcja tzw. **drzewa decyzyjnego** to jedna z najważniejszych technik stosowanych w klasyfikacji.

Drzewa decyzyjne (1)

- Indukcja tzw. **drzewa decyzyjnego** to jedna z najważniejszych technik stosowanych w klasyfikacji.
- Znajduje ona bardzo wiele zastosowań w naukach technicznych (np. w modelowaniu ryzyka awarii i systemach zabezpieczeń), ekonomicznych (np. w modelowaniu ryzyka inwestycyjnego, ubezpieczeniowego), ale także w naukach przyrodniczych i medycznych (np. w konstrukcji tzw. szybkiej ścieżki diagnostycznej dla konkretnego schorzenia).

Drzewa decyzyjne (1)

- Indukcja tzw. **drzewa decyzyjnego** to jedna z najważniejszych technik stosowanych w klasyfikacji.
- Znajduje ona bardzo wiele zastosowań w naukach technicznych (np. w modelowaniu ryzyka awarii i systemach zabezpieczeń), ekonomicznych (np. w modelowaniu ryzyka inwestycyjnego, ubezpieczeniowego), ale także w naukach przyrodniczych i medycznych (np. w konstrukcji tzw. szybkiej ścieżki diagnostycznej dla konkretnego schorzenia).
- **Drzewo decyzyjne** to graf bez cykli (pętli), w którym istnieje tylko jedna ścieżka między dwoma różnymi węzłami.

Drzewa decyzyjne (1)

- Indukcja tzw. **drzewa decyzyjnego** to jedna z najważniejszych technik stosowanych w klasyfikacji.
- Znajduje ona bardzo wiele zastosowań w naukach technicznych (np. w modelowaniu ryzyka awarii i systemach zabezpieczeń), ekonomicznych (np. w modelowaniu ryzyka inwestycyjnego, ubezpieczeniowego), ale także w naukach przyrodniczych i medycznych (np. w konstrukcji tzw. szybkiej ścieżki diagnostycznej dla konkretnego schorzenia).
- **Drzewo decyzyjne** to graf bez cykli (pętli), w którym istnieje tylko jedna ścieżka między dwoma różnymi węzłami.
- Reprezentuje ono pewien proces danego podziału zbioru obiektów na rozłączne klasy względem atrybutu decyzyjnego.

Drzewa decyzyjne (1)

- Indukcja tzw. **drzewa decyzyjnego** to jedna z najważniejszych technik stosowanych w klasyfikacji.
- Znajduje ona bardzo wiele zastosowań w naukach technicznych (np. w modelowaniu ryzyka awarii i systemach zabezpieczeń), ekonomicznych (np. w modelowaniu ryzyka inwestycyjnego, ubezpieczeniowego), ale także w naukach przyrodniczych i medycznych (np. w konstrukcji tzw. szybkiej ścieżki diagnostycznej dla konkretnego schorzenia).
- **Drzewo decyzyjne** to graf bez cykli (pętli), w którym istnieje tylko jedna ścieżka między dwoma różnymi węzłami.
- Reprezentuje ono pewien proces danego podziału zbioru obiektów na rozłączne klasy względem atrybutu decyzyjnego.
- Tworzenie drzewa rozpoczynamy od jego **korzenia (wierzchołka)**, z którego wychodzą **krawędzie (gałęzie)** reprezentujące wartości cech, na podstawie których dokonujemy podziału zbioru obiektów.

Drzewa decyzyjne (2)

- Wewnętrzne węzły drzewa opisują sposób dokonania tego podziału. Zakończenia krawędzi czyli **liście** drzewa odpowiadają klasom, do których należą obiekty.

Drzewa decyzyjne (2)

- Wewnętrzne węzły drzewa opisują sposób dokonania tego podziału. Zakończenia krawędzi czyli **liście** drzewa odpowiadają klasom, do których należą obiekty.
- Pierwsze algorytmy indukowania drzew klasyfikacyjnych pojawiły się w latach 70. dwudziestego wieku.

Drzewa decyzyjne (2)

- Wewnętrzne węzły drzewa opisują sposób dokonania tego podziału. Zakończenia krawędzi czyli **liście** drzewa odpowiadają klasom, do których należą obiekty.
- Pierwsze algorytmy indukowania drzew klasyfikacyjnych pojawiły się w latach 70. dwudziestego wieku.
- Jedne z pierwszych to **algorytm CHAID** (ang. *Chi-squared Automatic Interaction Detection*), oryginalnie zaproponowany tylko dla atrybutów danych w skali nominalnej, oraz **algorytm Id3** (Quinlan, 1986).

Drzewa decyzyjne (2)

- Wewnętrzne węzły drzewa opisują sposób dokonania tego podziału. Zakończenia krawędzi czyli **liście** drzewa odpowiadają klasom, do których należą obiekty.
- Pierwsze algorytmy indukowania drzew klasyfikacyjnych pojawiły się w latach 70. dwudziestego wieku.
- Jedne z pierwszych to **algorytm CHAID** (ang. *Chi-squared Automatic Interaction Detection*), oryginalnie zaproponowany tylko dla atrybutów danych w skali nominalnej, oraz **algorytm Id3** (Quinlan, 1986).
- Modyfikacja algorytmu Id3, czyli **algorytm C4.5** (Quinlan, 1993), oraz zaproponowany przez Breimana w 1984 roku **algorytm CART** stanowią swoistą bazę, na podstawie której w późniejszych latach powstało bardzo wiele innych algorytmów, m.in. **algorytm SPRINT**, wykorzystywany w komercyjnym systemie IBM Intelligent Miner.

Algorytmy wyboru atrybutu „podziałowego”

Absolutnie kluczową kwestią w przypadku konstrukcji drzewa decyzyjnego jest określenie **sposobu wyboru atrybutu „podziałowego”** na każdym etapie konstrukcji drzewa, a zatem atrybutu, który będzie generował podział drzewa na gałęzie: początkowo w korzeniu drzewa, a później na etapie węzłów pośrednich. W praktyce stosuje się w tym celu trzy różne podejścia, a co za tym idzie różne miary ilościowe:

- entropię (np. algorytmy Id3, C4.5);
- tzw. indeks Gini (Giniego) (np. algorytmy CART, SPRINT);
- statystykę χ^2 (np. algorytm CHAID).

Ogólny schemat konstrukcji drzewa decyzyjnego (1)

Zanim omówimy w szczegółach i na przykładach wymienione miary, przedstawmy ogólny schemat konstrukcji drzewa decyzyjnego. Obejmuje on następujące etapy.

1. Tworzymy korzeń drzewa reprezentujący dany zbiór treningowy D. Jeżeli wszystkie obiekty zbioru D reprezentują jedną i tę samą klasę C_i atrybutu decyzyjnego C, to przypisujemy mu etykietę C_i (korzeń staje się liściem drzewa) i procedura się kończy.
2. Jeżeli obiekty zbioru D reprezentują różne klasy atrybutu decyzyjnego, analizowany jest zbiór atrybutów warunkowych A_i, spośród których wybierany jest atrybut „podziałowy”. Jeśli zbiór atrybutów warunkowych jest pusty, korzeniowi przyporządkowujemy etykietę klasy C_i dominującej w zbiorze D i procedura się kończy (korzeń staje się liściem drzewa).

Ogólny schemat konstrukcji drzewa decyzyjnego (2)

3. Po przeprowadzeniu podziału w korzeniu drzewa analizujemy zbiór atrybutów warunkowych w każdym z nowo powstałych węzłów (oczywiście nie bierzemy już pod uwagę atrybutu, który był atrybutem „podziałowym" w pierwszym etapie). Jeśli zbiór atrybutów jest pusty, to bieżący węzeł staje się liściem drzewa z etykietą klasy dominującej. Jeśli zbiór atrybutów jest niepusty, dokonujemy kolejnego wyboru atrybutu „podziałowego" itd.
4. Podczas podziału w węźle drzewa dla każdej wartości (kategorii) atrybutu „podziałowego" tworzona jest krawędź (gałąź) drzewa o etykiecie odpowiadającej tej wartości. Gałąź kończy się nowym węzłem drzewa.

Algorytm oparty na entropii (1)

Jeżeli X jest zmienną losową typu dyskretnego, przyjmującą wartości $x_1, \ldots, x_m$ z prawdopodobieństwami, odpowiednio, $p_1, \ldots, p_m$, wówczas **entropię** zmiennej X definiujemy w następujący sposób:

$$H(X) \stackrel{def}{=} -\sum_{i=1}^{m} p_i \log_2 p_i.$$

Entropia jest miarą nieuporządkowania, rozproszenia zbioru danych. Rzeczywiście, przy równomiernym rozkładzie prawdopodobieństwa (maksymalnym rozproszeniu), tzn. gdy $p_i = \frac{1}{m}$, jej wartość jest maksymalna i wynosi $\log_2 m$.

Algorytm oparty na entropii (2)

- W teorii informacji entropia wyraża średnią ważoną ilości informacji dostarczanej przez pojedynczą wiadomość generowaną przez źródło informacji, przy czym wagami są tu prawdopodobieństwa nadania (wygenerowania) poszczególnych wiadomości.

Algorytm oparty na entropii (2)

- W teorii informacji entropia wyraża średnią ważoną ilości informacji dostarczanej przez pojedynczą wiadomość generowaną przez źródło informacji, przy czym wagami są tu prawdopodobieństwa nadania (wygenerowania) poszczególnych wiadomości.

- W praktyce można spotkać się z definicją entropii, w której logarytmy obliczane są przy podstawie innej niż 2. Jeżeli podstawa logarytmu wynosi 2, wówczas jednostką entropii jest **bit**.

Algorytm oparty na entropii (3)

Zastosowanie pojęcia entropii w procedurze indukowania drzewa decyzyjnego przebiega w następujący sposób. Podzielmy cały n-elementowy zbiór treningowy D na m rozłącznych **partycji** $D_1, \ldots, D_m$ związanych z wartościami (kategoriami) atrybutu decyzyjnego C. Oznaczmy przez n_i liczbę elementów partycji D_i, $i = 1, \ldots, m$. Entropia zbioru treningowego jest wówczas równa

$$H(D) = -\sum_{i=1}^{m} p_i \log_2 p_i = -\sum_{i=1}^{m} \frac{n_i}{n} \log_2 \frac{n_i}{n},$$

ponieważ prawdopodobieństwa p_i estymujemy za pomocą częstości empirycznych $\frac{n_i}{n}$. Jeżeli któraś z kategorii atrybutu decyzyjnego nie występuje w zbiorze treningowym (odpowiednie $n_i = 0$), wówczas przyjmujemy $\log_2 p_i = 0$.

Algorytm oparty na entropii (4)

Załóżmy, że atrybut warunkowy A przyjmuje r różnych wartości (występuje w r różnych kategoriach) $a_1, \ldots, a_r$ i niech $n_{i,j}$ oznacza liczbę elementów partycji D_i, w których $A = a_j$, $i = 1, \ldots, m$, $j = 1, \ldots, r$. **Entropię kategorii a_j atrybutu A w zbiorze treningowym** możemy zdefiniować w następujący sposób:

$$H(a_j) = -\sum_{i=1}^{m} p_{i,j} \log_2 p_{i,j} = -\sum_{i=1}^{m} \frac{n_{i,j}}{n_j} \log_2 \frac{n_{i,j}}{n_j},$$

gdzie $n_{i,j}$ oznacza liczbę obiektów zbioru treningowego należących do partycji D_i, dla których równocześnie $A = a_j$.

Algorytm oparty na entropii (5)

Entropię atrybutu A definiujemy jako następującą ważoną sumę:

$$H(A) = \sum_{j=1}^{r} \widehat{n}_j H(a_j) = \frac{1}{n} \sum_{j=1}^{r} (n_{1,j} + \ldots + n_{m,j}) H(a_j).$$

Iloraz $\widehat{n}_j = \frac{n_{1,j} + \ldots + n_{m,j}}{n}$ jest częstością względną wystąpienia kategorii a_j w zbiorze treningowym D.

Algorytm oparty na entropii (6)

Mając zdefiniowaną entropię atrybutu A, definiujemy teraz tzw. **zysk informacyjny dla atrybutu** A w następujący sposób:

$$Gain(A) = H(D) - H(A).$$

Zysk informacyjny wyraża różnicę pomiędzy ilością informacji niezbędnej do sklasyfikowania dowolnego obiektu ze zbioru D przed podziałem tego zbioru oraz po jego podziale z wykorzystaniem atrybutu A jako atrybutu „podziałowego".
W praktyce na każdym etapie konstrukcji drzewa klasyfikacyjnego obliczamy zysk informacyjny dla każdego z dostępnych jeszcze atrybutów warunkowych. **Wybieramy jako atrybut „podziałowy" ten, dla którego zysk informacyjny jest największy (czyli, równoważnie, entropia jest minimalna)**.