Klasyfikacja

Sformułowanie problemu

Metody klasyfikacji

Kryteria oceny metod klasyfikacji

Niniejszy wykład poświęcimy kolejnej metodzie eksploracji danych – klasyfikacji. Na początek kilka słów wprowadzających oraz przedstawimy formalne sformułowanie problemu klasyfikacji. Następnie omówmy szczegółowo poszczególne fazy procesu klasyfikacji. W dalszej części skupimy się na metodach klasyfikacji oraz omówimy kryteria oceny metod klasyfikacji. Na zakończenie omówimy algorytm indukcji drzewa decyzyjnego oraz kryteria oceny podziału węzła w drzewie.

Klasyfikacja (1)

- **Dane wejściowe**

treningowy zbiór krotek (przykładów, obserwacji, próbek), będących listą wartości atrybutów opisowych (tzw. deskryptorów) i wybranego atrybutu decyzyjnego (*ang. class label attribute*)

- **Dane wyjściowe**

model (klasyfikator), przydziela każdej krotce wartość atrybutu decyzyjnego w oparciu o wartości pozostałych atrybutów (deskryptorów)

Jedną z najstarszych jak również najważniejszych metod eksploracji danych, która ma bardzo istotne znaczenie praktyczne jest metoda klasyfikacji (*ang. classification*). Polega ona na znajdowaniu odwzorowania danych w zbiór predefiniowanych klas. Na podstawie zawartości bazy danych budowany jest model (np. drzewo decyzyjne, reguły logiczne), który służy do klasyfikowania nowych obiektów w bazie danych lub głębszego zrozumienia istniejącego podziału obiektów na predefiniowane klasy. Klasyfikacja znalazła szereg zastosowań np.: rozpoznawanie trendów na rynkach finansowych, automatyczne rozpoznawanie obiektów w dużych bazach danych obrazów, wspomaganie decyzji przyznawania kredytów bankowych. Ogromne zastosowanie znalazła w systemach medycznych, przykładowo, w bazie danych medycznych znalezione mogą być reguły klasyfikujące poszczególne schorzenia, a następnie przy pomocy znalezionych reguł automatycznie może być przeprowadzone diagnozowanie kolejnych pacjentów.

Klasyfikacja jest metodą eksploracji danych z nadzorem (z nauczycielem). Proces klasyfikacji składa się z kilku etapów – budowania modelu, po czym następuje faza testowania oraz predykcji nieznanych wartości.

Głównym celem klasyfikacji jest zbudowanie formalnego modelu zwanego klasyfikatorem. Danymi wejściowymi w procesie klasyfikacji jest treningowy zbiór krotek (przykładów, obserwacji, próbek), będących listą wartości atrybutów opisowych (tzw. deskryptorów) i wybranego atrybutu decyzyjnego (ang. *class label attribute*). Wynikiem procesu klasyfikacji jest pewien otrzymany model (klasyfikator), który przydziela każdej krotce (przykładowi) wartość atrybutu decyzyjnego w oparciu o wartości pozostałych atrybutów (deskryptorów).

Klasyfikacja (2)

- Wartości atrybutu decyzyjnego dzielą zbiór krotek na predefiniowane **klasy**, składające się z krotek o tej samej wartości atrybutu decyzyjnego

Klasyfikator
służy do predykcji wartości atrybutu decyzyjnego (klasy) krotek, dla których wartość atrybutu decyzyjnego, tj. przydział do klasy, nie jest znany

Wprowadzimy obecnie kilka pojęć, którymi będziemy się posługiwać w dalszej części wykładu. Spośród zbioru atrybutów zwanego zbiorem treningowym możemy wyznaczyć jeden atrybut, który zwany jest atrybutem decyzyjnym. Wartości atrybutu decyzyjnego dzielą zbiór krotek na predefiniowane klasy, składające się z krotek o tej samej wartości atrybutu decyzyjnego. Mówiąc o klasyfikacji będziemy często używać określenia klasyfikator. Klasyfikator jest modelem, który służy do predykcji wartości atrybutu decyzyjnego (klasy) krotek, dla których wartość atrybutu decyzyjnego, tj. przydział do klasy, nie jest znany.

Czym jest klasyfikacja?

- **Klasyfikacja danych jest dwu-etapowym procesem:**
  - **Etap 1:**

    budowa modelu (klasyfikatora) opisującego predefiniowany zbiór klas danych lub zbiór pojęć

  - **Etap 2:**

    zastosowanie opracowanego modelu do klasyfikacji nowych danych

Klasyfikacja, jak wcześniej zostało wspomniane jest procesem dwuetapowym. W pierwszym etapie konstruujemy model (klasyfikator), opisujący predefiniowany zbiór klas danych lub zbiór pojęć. W drugim etapie klasyfikacji otrzymany model stosujemy do klasyfikacji nowych danych.

Trening i testowanie (1)

- Zbiór dostępnych krotek (przykładów, obserwacji, próbek) dzielimy na dwa zbiory: zbiór treningowy i zbiór testowy
- Model klasyfikacyjny (klasyfikator) jest budowany dwu-etapowo:

Uczenie (trening) – klasyfikator jest budowany w oparciu o zbiór treningowy danych

Testowanie – dokładność (jakość) klasyfikatora jest weryfikowana w oparciu o zbiór testowy danych

Bazę danych czyli zbiór dostępnych krotek (przykładów, obserwacji czy próbek) dzielimy na dwa zbiory. Pierwszym będzie zbiór treningowy, z którego budujemy model. Drugi zbiór, zwany zbiorem testowym będzie służył do testowania modelu.

Model uczący może być używany do przewidywania klas nowych krotek, dla których atrybut decyzyjny jest utracony lub nieznany. Dwuetapowy proces budowy klasyfikatora składa się z fazy treningowej zwanej uczeniem. Klasyfikator jest budowany w oparciu o zbiór treningowy danych. Druga faza – testowania, polega na weryfikacji dokładności (jakość) klasyfikatora w oparciu o zbiór testowy danych.

Trening i testowanie (2)

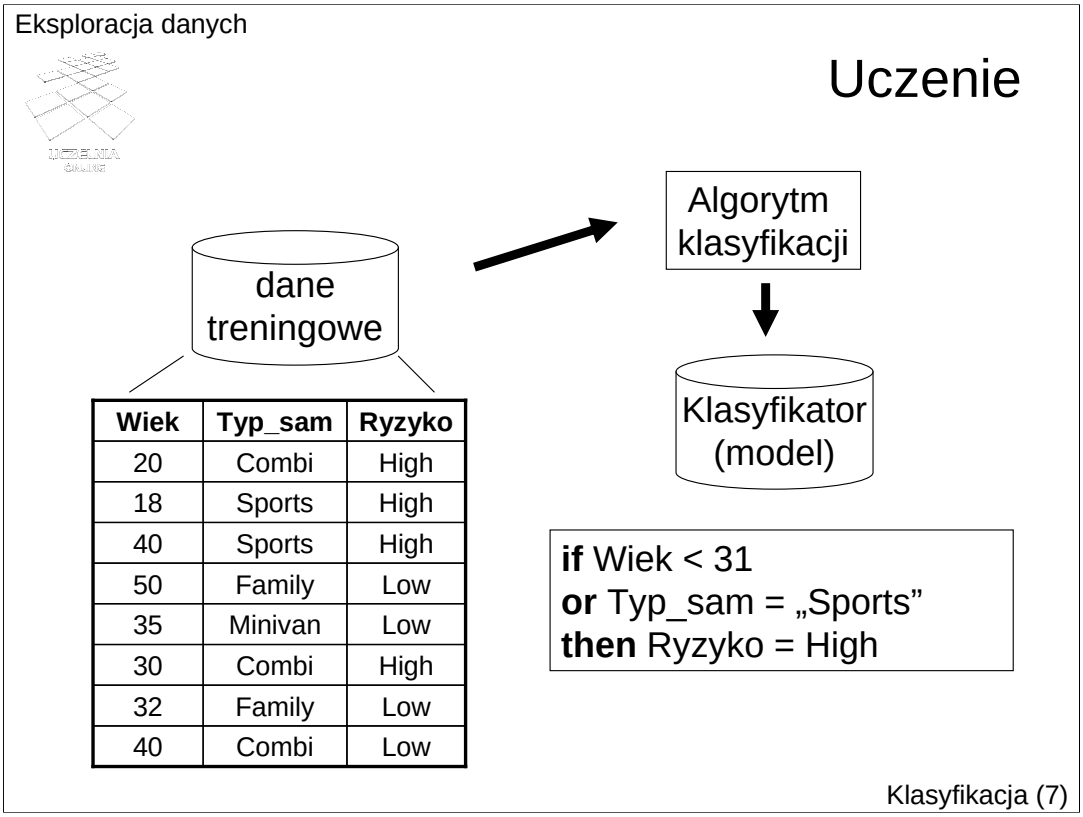
- **Wynik klasyfikacji:**
  - Reguły klasyfikacyjne postaci ***if - then***
  - Formuły logiczne
  - Drzewa decyzyjne
- **Dokładność modelu:**

Dla przykładów testowych, dla których znane są wartości atrybutu decyzyjnego, wartości te są porównywane z wartościami atrybutu decyzyjnego generowanymi dla tych przykładów przez klasyfikator

Współczynnik dokładności (*ang. accuracy rate*) = %
procent przykładów testowych
poprawnie zaklasyfikowanych przez model

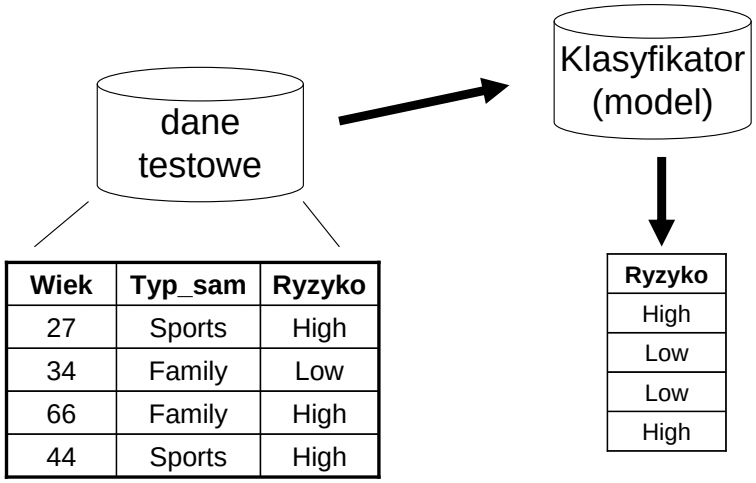
Mamy wiele różnych sposobów reprezentacji modelu uczącego dla klasyfikacji i predykcji, i dla każdego istnieją dedykowane techniki, które mogą być użyte do wnioskowania wyjściowej struktury z danych. Zazwyczaj, w klasyfikacji, model uczący (klasyfikator) jest przedstawiany w postaci drzewa decyzyjnego, tabeli decyzyjnej, lub reguł klasyfikacyjnych postaci IF - THEN.

Istotną sprawą z punktu widzenia poprawności i efektywności modelu jest tzw. dokładność modelu. Dokładność modelu weryfikowana jest w następujący sposób: dla przykładów testowych, dla których znane są wartości atrybutu decyzyjnego, wartości te są porównywane z wartościami atrybutu decyzyjnego generowanymi dla tych przykładów przez klasyfikator. Miarą, która weryfikuje poprawność modelu jest współczynnik dokładności. Współczynnik dokładności modelu jest liczony jako procent przykładów testowych poprawnie zaklasyfikowanych przez model. Jeśli dokładność modelu jest akceptowalna, model może być użyty do klasyfikacji przyszłych danych i przewidywania wartości nowych krotek, dla których wartość atrybutu decyzyjnego jest nieznana.



Przedstawimy obecnie prosty przykład, ilustrujący omówione wcześniej etapy klasyfikacji. Załóżmy, że dana jest baza danych ubezpieczalni zawierająca dane o kierowcach – informacje o spowodowanych przez nich wypadkach. Baza danych jest bardzo prostą relacją zawierającą trzy atrybuty. Atrybut Wiek kierowcy, Typ_sam czyli typ samochodu oraz atrybut Ryzyko związany z informacją, że dany kierowca spowodował wcześniej wypadki czy nie powodował wcześniej wypadku. Jeżeli jest autorem kilku wypadków wartość atrybutu Ryzyko przyjmuje wartość High, w przypadku gdy nie spowodował żadnego wypadku atrybut Ryzyko przyjmuje wartość Low. Atrybut Ryzyko jest atrybutem decyzyjnym. Załóżmy, że z bazy danych ubezpieczalni, wydzielono zbiór danych treningowych. Zbiór ten przedstawiono na slajdzie. Następnie zbiór danych treningowych został poddany algorytmowi klasyfikacji. Algorytm klasyfikacji konstruuje klasyfikator, który może być postaci drzewa decyzyjnego, zbioru reguł decyzyjnych, tabeli decyzyjnych. W naszym przykładzie przedstawionym na slajdzie wynikiem działania algorytmu klasyfikacji jest klasyfikator w postaci pojedynczej reguły decyzyjnej: „Jeżeli wiek kierowcy jest mniejszy niż 31 lub typ samochodu sportowy to Ryzyko jest wysokie”.

Testowanie



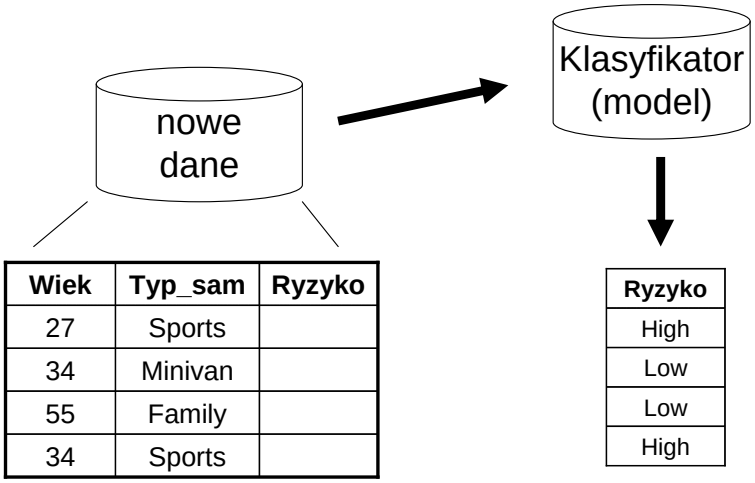
Wiek	Typ_sam	Ryzyko
27	Sports	High
34	Family	Low
66	Family	High
44	Sports	High

Ryzyko
High
Low
Low
High

Dokładność = 3/4 = 75%

W drugim etapie klasyfikacji, zwanym etapem testowania dokonujemy weryfikacji dokładności opracowanego modelu. Weryfikacja dokładności modelu jest realizowana w następujący sposób: dla zbioru przykładów testowych, dla których znane są wartości atrybutu decyzyjnego, wartości te są porównywane z wartościami atrybutu decyzyjnego generowanymi dla tych przykładów przez klasyfikator. Na slajdzie przedstawiony jest zbiór danych testowych wyodrębniony z bazy danych ubezpieczalni. Zbiór ten zawiera cztery krotki. Zbiór danych testowych zostaje poddany procesowi klasyfikacji. Klasyfikator generuje dla podanych rekordów wartości atrybutu decyzyjnego. Następnie weryfikujemy dokładność modelu. Miarą weryfikującą dokładność modelu jest tzw. Współczynnik dokładności modelu, który jest liczony jako procent przykładów testowych poprawnie zaklasyfikowanych przez model. Zauważmy, że klasyfikator wygenerował następujące wartości atrybutu decyzyjnego dla zbioru danych testowych: High, Low, Low, High. Jeżeli porównamy wartości atrybutu decyzyjnego, które zostały wygenerowane dla tych przykładów przez klasyfikator i porównamy je z wartościami atrybutu decyzyjnego w zbiorze danych testowych, okazuje się że klasyfikator poprawnie zaklasyfikował 3 z 4 przypadków. A zatem współczynnik dokładności modelu wynosi ¾ czyli 75%.

Klasyfikacja (predykcja)



Wiek	Typ_sam	Ryzyko
27	Sports	
34	Minivan	
55	Family	
34	Sports	

Ryzyko
High
Low
Low
High

Jeżeli dokładność klasyfikatora jest akceptowalna, wówczas możemy wykorzystać klasyfikator do klasyfikacji nowych danych. Celem klasyfikacji, jak pamiętamy jest przyporządkowanie nowych danych dla których wartość atrybutu decyzyjnego nie jest znana do odpowiedniej klasy. Na prezentowanym slajdzie podany jest zbiór danych, dla których wartość atrybutu decyzyjnego Ryzyko nie jest znana. Zbiór ten poddajemy procesowi klasyfikacji, w wyniku czego otrzymujemy następujące wartości atrybutu decyzyjnego Ryzyko: High,Low,Low,High. Co oznacza, że klasyfikator zaklasyfikował nowych kierowców do odpowiednich klas. Kierowca, który ma 27 lat i jeździ samochodem sportowym zaklasyfikował do kierowców wysokiego ryzyka. Kierowcę, który ma 34 lata i jeździ samochodem Minivan zaklasyfikował do kategorii kierowców niskiego ryzyka. Podobnie kierowca mający 55 lat i dysponujący samochodem rodzinnym został zaklasyfikowany do kierowców niskiego ryzyka. Wreszcie kierowca, który ma 34 lata i jeździ samochodem sportowym zaklasyfikowano do kategorii kierowców wysokiego ryzyka.

Klasyfikacja a predykcja

- Dwie metody, które są stosowane do analizy danych i ekstrakcji modeli opisujących klasy danych lub do predykcji trendów:
  - **klasyfikacja:**

    predykcja wartości atrybutu kategorycznego (predykcja klasy)

  - **predykcja:**

    modelowanie funkcji ciągłych

Istnieją dwie metody, które są stosowane do analizy danych, ekstrakcji modeli opisujących klasy danych lub predykcji wartości wybranego atrybutu. Są to KLASYFIKACJA oraz PREDYKCJA. Atrybuty poddane analizie mogą być zarówno numeryczne jak i kategoryczne. Jeśli atrybut decyzyjny jest kategoryczny, wówczas problem predykcji wartości takiego atrybutu jest przedstawiany jako problem klasyfikacji. Jeśli atrybut decyzyjny jest ciągły (numeryczny), problem jest zwany problemem predykcji. Predykcja jest bardzo podobna do klasyfikacji. Jednakże celem predykcji jest zamodelowanie funkcji ciągłej, która by odwzorowywała wartości atrybutu decyzyjnego.

Kryteria porównawcze metod klasyfikacji (1)

- **Dokładność predykcji** *(ang. predictive accuracy)*:

  zdolność modelu do poprawnej predykcji wartości atrybutu decyzyjnego (klasy) nowego przykładu

- **Efektywność** (ang. *speed)*:

  koszt obliczeniowy związany z wygenerowaniem i zastosowaniem klasyfikatora

- **Odporność modelu** *(ang. robustness)*:

  zdolność modelu do poprawnej predykcji klas w przypadku braku części danych lub występowania danych zaszumionych

W dalszej części wykładu skoncentrujemy się na metodach klasyfikacji. W literaturze zaproponowano wiele modeli klasyfikacji, są to np.: drzewa decyzyjne, tabele decyzyjne, metoda Bayesa, sieci neuronowe, algorytmy genetyczne, metoda k-najbliższych sąsiadów (ang. *k-nearest neighbor*), zbiory przybliżone i wiele innych statystycznych metod, które można zastosować przy klasyfikacji danych. Różnorodność modeli spowodowała wyspecyfikowanie kryteriów porównawczych metod, dzięki którym można dokonać odpowiedniego wyboru metody dla danego zastosowania. Istnieje szereg kryteriów porównawczych, które obecnie krótko scharakteryzujemy. Pierwszym kryterium porównawczym jest kryterium dokładności predykcji (ang. *predictive accuracy*). Pod pojęciem predykcji rozumiemy zdolność modelu do poprawnej predykcji wartości atrybutu decyzyjnego (klasy) nowego przykładu. Innym kryterium jest kryterium efektywności (ang. *speed*) związaną z kosztem obliczeniowym wynikającym z wygenerowania i zastosowania klasyfikatora. W analizowanych danych pojawiają się czasem „luki” czyli brakujące wartości danych, lub też dane przypadkowe, zaciemniające obraz (tzw. dane zaszumione). Zdolność modelu do poprawnej predykcji klas w przypadku pojawienia się wcześniej wspomnianych rodzajów danych niepożądanych ocenia się w kategorii kryterium odporności modelu (ang. robustness).

Kryteria porównawcze metod klasyfikacji (2)

- **Skalowalność** *(ang. scalability)*:

  zdolność do konstrukcji klasyfikatora dla dowolnie dużych wolumenów danych

- **Interpretowalność** *(ang. interpretability)*:

  odnosi się do stopnia w jakim konstrukcja klasyfikatora pozwala na zrozumienie mechanizmu klasyfikacji danych

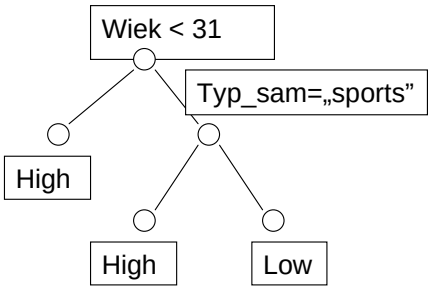
- **Kryteria dziedzinowo-zależne**

Przy analizie danych pochodzących z dużej bazy danych, istotnym czynnikiem oceny metody jest skalowalność czyli zdolność metody do konstrukcji klasyfikatora dla dowolnie dużych wolumenów danych. Otrzymany wynik metody klasyfikacji powinien być łatwo interpretowalny, z czym wiąże się następne kryterium interpretowalności (ang. *interpretability*). Kryterium to odnosi się do stopnia w jakim konstrukcja klasyfikatora pozwala na zrozumienie mechanizmu klasyfikacji danych. Istnieją metody klasyfikacji, które charakteryzują się dużą dokładnością, np. sieci neuronowe, które jednakże nie pozwalają na zrozumienie mechanizmu samej klasyfikacji danych. W praktyce metody te są w praktyce mało przydatne. Każda dziedzina ma swoje specyficzne potrzeby, w związku z tym przy wyborze metody klasyfikacji kieruje się własnymi kryteriami, które spełniają specjalistyczne oczekiwania. Kryterium to ogólnie przyjęło nazwę kryterium dziedzinowo-zależnego.

Sformułowanie problemu

- Dana jest baza danych przykładów, z których każdy należy do określonej klasy, zgodnie z wartością atrybutu decyzyjnego. Celem klasyfikacji jest znalezienie modelu dla każdej klasy

Wiek	Typ_sam	Ryzyko
20	Combi	High
18	Sports	High
40	Sports	High
50	Family	Low
35	Minivan	Low
30	Combi	High
32	Family	Low
40	Combi	Low



Przejdziemy teraz do sformułowania problemu klasyfikacji danych. Mamy daną bazę danych rekordów (przykładów), z których każdy posiada etykietę klasy, do której należy, zgodnie z wartością atrybutu decyzyjnego. Celem klasyfikacji będzie znalezienie modelu dla każdej klasy czyli opisu rekordów każdej z klas. Przykładowym problemem klasyfikacji może być automatyczny podział kierowców na powodujących i niepowodujących wypadki drogowe. Na powyższym slajdzie umieszczona została przykładowa baza danych, zawierającą informacje o wieku kierowcy, typie posiadanego samochodu oraz ryzyku związane z możliwością spowodowania wypadku. Dla powyższych danych, został zbudowany przykładowy model - klasyfikator w postaci drzewa decyzyjnego przedstawionego na slajdzie. Z podanego drzewa decyzyjnego, możemy odczytać następujące reguły decyzyjne: „Jeżeli kierowca ma poniżej 31 lat to ryzyko spowodowania wypadku jest duże”, inną reguła decyzyjną jest reguła: „Jeżeli kierowca ma powyżej 31 lat i dysponuje sportowym samochodem to ryzyko spowodowania wypadku jest duże”, wreszcie „Jeżeli kierowca ma powyżej 31 lat i typ samochodu jest różny od sportowego to ryzyko spowodowania wypadku jest niskie”.

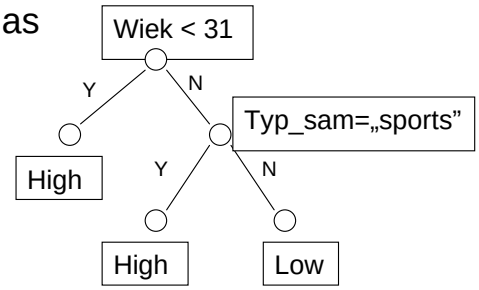
Metody klasyfikacji

- Klasyfikacja poprzez indukcję drzew decyzyjnych
- Klasyfikatory Bayes'owskie
- Sieci Neuronowe
- Analiza statystyczna
- Metaheurystyki (np. algorytmy genetyczne)
- Zbiory przybliżone
- k-NN – k-najbliższe sąsiedztwo

W literaturze zaproponowano wiele metod klasyfikacji, jest to np: klasyfikacja poprzez indukcję drzew decyzyjnych, klasyfikatory Bayes'owskie, sieci neuronowe, analiza statystyczna, metaheurystyki (np. algorytmy genetyczne), zbiory przybliżone, metoda k-NN czyli k-najbliższych sąsiadów (ang. *k-nearest neighbor*) i wiele innych statystycznych metod, które cały czas są rozwijane. Wśród wymienionych metod klasyfikacji najczęstszym jest metoda klasyfikacji poprzez indukcję drzew decyzyjnych, która jest szczególnie atrakcyjne dla eksploracji danych. Po pierwsze, dzięki intuicyjnej reprezentacji końcowy/otrzymany model klasyfikacji jest zrozumiały dla człowieka. Po drugie, drzewa decyzyjne mogą być konstruowane stosunkowo szybko w porównaniu z innymi metodami klasyfikacji. Kolejnym atutem drzew decyzyjnych jest, skalowalność dla dużych zbiorów danych i możliwość użycia wielowymiarowych danych. W dodatku dokładność drzew decyzyjnych jest porównywalna z innymi metodami klasyfikacji. Większość komercyjnych dostępnych narzędzi do eksploracji danych opiera się na modelu drzew decyzyjnych. Główną wadą drzew decyzyjnych jest niemożność wychwycenia korelacji pomiędzy atrybutami bez dodatkowych obliczeń. W następnej części wykładu przyjrzymy się bliżej wyżej wymienionej metodzie klasyfikacji.

Klasyfikacja poprzez indukcję drzew decyzyjnych (1)

- Drzewo decyzyjne jest grafem o strukturze drzewiastej, gdzie
  - każdy wierzchołek wewnętrzny reprezentuje test na atrybucie (atrybutach),
  - każdy łuk reprezentuje wynik testu,
  - każdy liść reprezentuje pojedynczą klasę lub rozkład wartości klas



Wynikiem klasyfikacji w metodzie indukcji drzew decyzyjnych jest drzewo decyzyjne. Drzewo decyzyjne jest skierowanym acyklicznym grafem o strukturze drzewiastej, w którym każdy wierzchołek reprezentuje test na atrybucie (atrybutach), każdy łuk reprezentuje wynik testu, każdy liść reprezentuje pojedynczą klasę lub rozkład wartości klas. Najwyższy węzeł / wierzchołek drzewa nazywany rootem (korzeniem drzewa). Przyjrzyjmy się jeszcze raz przedstawionemu drzewu klasyfikacyjnemu z poprzedniego przykładu. W liściach mamy pojedynczą klasę lub rozkład wartości klas atrybuty decyzyjnego „Ryzyko”: (High,Low). Każdy wierzchołek wewnętrzny jest testem na atrybucie, natomiast łuk jest wynikiem testu (Y lub N).

Klasyfikacja poprzez indukcję drzew decyzyjnych (2)

- Drzewo decyzyjne rekurencyjnie dzieli zbiór treningowy na partycje do momentu, w którym każda partycja zawiera dane należące do jednej klasy, lub, gdy w ramach partycji dominują dane należące do jednej klasy
- Każdy wierzchołek wewnętrzny drzewa zawiera tzw. **punkt podziału** (*ang. split point***),** którym jest test na atrybucie (atrybutach), który dzieli zbiór danych na partycje

Drzewo decyzyjne rekurencyjnie dzieli zbiór treningowy na partycje do momentu, w którym każda partycja zawiera dane należące do jednej klasy, lub, gdy w ramach partycji dominują dane należące do jednej klasy, natomiast rozmiar partycji jest ograniczony. Każdy wierzchołek wewnętrzny drzewa zawiera tzw. punkt podziału (*ang. split point*), którym jest test na atrybucie (atrybutach), który dzieli zbiór danych na partycje.

Klasyfikacja poprzez indukcję drzew decyzyjnych (3)

- **Algorytm podstawowy**:
  algorytm zachłanny, który konstruuje rekurencyjnie drzewo decyzyjne metodą top-down
- Wiele wariantów algorytmu podstawowego (źródła):
  - uczenie maszynowe (ID3, C4.5)
  - statystyka (CART)
  - rozpoznawanie obrazów (CHAID)
- Podstawowa różnica: kryterium podziału

Podstawowym algorytmem konstrukcji drzew decyzyjnych używanym w etapie konstrukcji jest algorytm zachłanny, który tworzy drzewo decyzyjne w rekurencyjny sposób techniką top-down w sposób „dziel i rządź” (ang. *divide-and-conquer*). Istnieje wiele wariantów algorytmu podstawowego. Najczęściej stosowanymi algorytmami, pochodzącymi z uczenia maszynowego są algorytmy ID3 oraz C4.5 Inną techniką jest pochodzącą ze statystyki metoda CART, czy też metoda związana z rozpoznawaniem obrazów CHAID. Podstawową różnicą powyższych algorytmów jest przyjęte kryterium podziału, czyli sposobu w jaki tworzone są nowe węzły wewnętrzne w drzewie decyzyjnym, używanego podczas fazy budowania drzewa decyzyjnego. Metoda podziału powinna maksymalizować dokładność konstruowanego drzewa decyzyjnego, lub innymi słowy minimalizować błędną klasyfikację rekordów danych.

Klasyfikacja poprzez indukcję drzew decyzyjnych (4)

- **Algorytm jest wykonywany w dwóch fazach:**

  - **Faza 1:**

    Konstrukcja drzewa decyzyjnego w oparciu o zbiór treningowy

  - **Faza 2:**

    Obcinanie drzewa w celu poprawy dokładności, interpretowalności i uniezależnienia się od efektu przetrenowania

Drzewo decyzyjne jest zwykle konstruowane w dwóch fazach. W fazie pierwszej, zwanej fazą budowania, fazą wzrostu lub fazą indukcji drzew decyzyjnych, drzewo decyzyjne jest tworzone z treningowej bazy danych. W fazie drugiej, zwanej fazą obcinania lub redukcji drzewa (ang. *pruning*), następuje obcinanie drzewa w celu poprawy dokładności, interpretowalności i uniezależnienia się od efektu przetrenowania. W fazie obcinania następuje identyfikacja i usunięcie gałęzi reprezentujące punkty osobliwe i szum.

Z przycinaniem drzewa wiążą się dwie główne strategie postpruning, w którym konstruujemy pełne drzewo decyzji i usuwamy z niego zawodne części. Strategia druga – prepruning - przestaje rozwijać gałąź, gdy informacje zaczynają być zawodne. W praktyce preferowany jest postpruning, gdyż prepruning często powoduje efekt „wczesnego stopu”.

Konstrukcja drzewa

- W fazie konstrukcji drzewa, zbiór treningowy jest dzielony na partycje, rekurencyjnie, w punktach podziału do momentu, gdy każda z partycji jest „czysta” (zawiera dane należące wyłącznie do jednej klasy) lub liczba elementów partycji dostatecznie mała (spada poniżej pewnego zadanego progu)
- Postać testu stanowiącego punkt podziału zależy od kryterium podziału i typu danych atrybutu występującego w teście:

dla atrybutu ciągłego A, test ma postać wartość(A) < x, gdzie x należy do dziedziny atrybutu A, $x \in dom(A)$

dla atrybutu kategorycznego A, test ma postać wartość(A) $\in$ X, gdzie $X \subset dom(A)$

Przyjrzymy się obecnie nieco dokładniej fazie konstrukcji drzewa decyzyjnego. W fazie konstrukcji drzewa, zbiór treningowy jest dzielony na partycje, rekurencyjnie, w punktach podziału do momentu, gdy każda z partycji jest „czysta” (zawiera dane należące wyłącznie do jednej klasy) lub liczba elementów partycji dostatecznie mała (spada poniżej pewnego zadanego progu). Postać testu stanowiącego punkt podziału zależy od kryterium podziału i typu danych atrybutu występującego w teście. Jak pamiętamy atrybuty mogą być typu ciągłego oraz typu kategorycznego. Dla atrybutu ciągłego A, test ma postać wartość(A) < x, gdzie x należy do dziedziny atrybutu A, x należy do dom(A), gdzie X zawiera się w dom(A). Dla atrybutu kategorycznego A, test ma postać wartość(A) należy do X, gdzie X jest podzbiorem dom(A).

Eksploracja danych

Algorytm konstrukcji drzewa (1)

```
Make Tree (Training Data D)
{
      Partition(D)
}
Partition(Data S)
{
      if (all points in S are in the same class)
      then
            return
      for each attribute A do
            evaluate splits on attribute A;
      use best split found to partition S into S1 and S2
      Partition(S1)
      Partition(S2)
}
```

Klasyfikacja (20)

Przejdźmy obecnie do przedstawienia algorytmu konstrukcji drzewa decyzyjnego. Prezentowany na slajdzie algorytm konstruuje binarne drzewo decyzyjne. Nie jest to algorytm ogólny, gdyż niektóre algorytmy klasyfikacji metodą indukcji drzew decyzyjnych konstruują drzewa decyzyjne, które nie są binarne, niemniej algorytm ten dobrze ilustruje mechanizm konstrukcji drzewa decyzyjnego. Podstawową procedurą prezentowanego algorytmu jest procedura Make Tree(Training Data D), której argumentem wejściowym jest cały zbiór danych treningowych D. Procedura Make Tree wywołuje procedurę Partition, której na początku parametrem wejściowym jest zbiór danych treningowych D. Budowa drzewa rozpoczyna się od pojedynczego węzła/wierzchołka zwanego korzeniem (root N node) reprezentującego treningową bazę danych D. Jeśli wszystkie krotki w D należą do tej samej klasy C, wówczas, węzeł N staje się liściem z etykietą C, i algorytm kończy swoje działanie. W przeciwnym razie, zbiór atrybutów A jest sprawdzany zgodnie z metodą selekcji podziału (split selection) SS i wybierany jest atrybut podziału zwany „best-split”. Atrybut podziału partycjonuje/dzieli zbiór treningowy D na zbiór oddzielnej klasy próbek S1, S2, ... Sv, gdzie Si=1,..,v zawiera wszystkie próbki ze zbioru D razem z punktem podziału. Gałąź, z etykietą Vi, jest tworzona dla każdej wartości ai atrybutu podziału, i dla każdej gałęzi Vi przydzielony jest zbiór próbek. Procedura partycjonowania jest powtarzana rekurencyjnie dla każdego węzła/wierzchołka potomka, przez co jest formowane drzewo decyzyjne dla każdej partycji przykładów. Procedura się kończy gdy każda z partycji jest „czysta” (zawiera dane należące wyłącznie do jednej klasy) lub liczba elementów partycji dostatecznie mała (spada poniżej pewnego zadanego progu).

Algorytm konstrukcji drzewa (2)

- W trakcie budowy drzewa decyzyjnego, wybieramy taki atrybut i taki punkt podziału, określający wierzchołek wewnętrzny drzewa decyzyjnego, który „najlepiej” dzieli zbiór danych treningowych należących do tego wierzchołka
- Do oceny jakości punktu podziału zaproponowano szereg kryteriów (wskaźników)

W trakcie budowy drzewa decyzyjnego, musimy zwrócić szczególną uwagę na wybór takiego atrybutu i takiego punkt podziału, który określi wierzchołek wewnętrzny drzewa decyzyjnego, innymi słowy „najlepiej” podzieli zbiór danych treningowych należących do tego wierzchołka. Najczęstszą metodą wybieraną w systemach komercyjnych jest metoda, która polega na wyborze takiego atrybutu i takiego punktu podziału, który będzie minimalizował przyjęta miarę „zanieczyszczenia” zbioru danych. Metoda ta znajduje atrybut podziału wierzchołków drzewa decyzyjnego poprzez minimalizację miary „zanieczyszczenia”. Do oceny jakości punktu podziału zaproponowano szereg wskaźników (kryteriów), które przedstawimy na kolejnym slajdzie.

Kryteria oceny podziału

- Indeks Gini (algorytmy CART, SPRINT)

  Wybieramy atrybut, który minimalizuje indeks Gini

- Zysk informacyjny (algorytmy ID3, C4.5)

  Wybieramy atrybut, który maksymalizuje redukcję entropii

- Indeks korelacji $\chi 2$ (algorytm CHAID)

  Mierzymy korelację pomiędzy każdym atrybutem i każdą klasą (wartością atrybutu decyzyjnego)

  Wybieramy atrybut o maksymalnej korelacji

W literaturze zaproponowano szereg kryteriów oceny jakości punktu podziału, w praktyce w systemach komercyjnych wykorzystuje się trzy podstawowe kryteria. Mianowicie indeks gini, zysk informacyjny oraz indeks korelacji x2. W pierwszym przypadku wybieramy atrybut, który minimalizuje wartość indeksu gini, stosowany w algorytmach CART i SPRINT. W przypadku zysku informacyjnego stosowanego w algorytmach ID3 oraz C4.5 wybieramy atrybut, który maksymalizuje redukcję entropii. W przypadku indeksu korelacji x2 stosowanego w algorytmie CHAID mierzymy korelację pomiędzy każdym atrybutem i każdą klasą (wartością atrybutu decyzyjnego), ostatecznie wybieramy atrybut o maksymalnej korelacji.