Zgłębianie danych - wstępne przygotowanie danych, wykład 3

Joanna Jędrzejowicz

Instytut Informatyki

Co to jest zgłębianie danych?

Data mining (ang.):

- zgłębianie danych
- drążenie danych
- odkrywanie wiedzy z danych

Odkrywanie interesujących (nietrywialnych, nieoczywistych, wcześniej nieznanych i potencjalnie przydatnych) informacji lub prawidłowości z wielkich baz danych.

Przykłady zastosowań

Analiza danych i wspomaganie decyzji

- Analizy rynkowe i zarządzanie - marketing skierowany, zarządzanie relacjami, analiza koszyka rynkowego, segmentacja rynku
- Analiza i zarządzanie ryzykiem - prognozy, utrzymanie klientów, kontrola jakości, analizy konkurencyjności
- Wykrywanie oszustw i wyłudzeń

Inne zastosowania

- Eksploracja tekstu (doniesień agencyjnych, dokumentów, poczty elektronicznej), analiza zasobów informacyjnych Internetu (strony WWW)
- Inteligentne wyszukiwanie informacji

Funkcjonalność eksploracji danych

- Charakteryzowanie i różnicowanie zjawisk i pojęć - uogólnianie, podsumowywanie oraz rozróżnianie atrybutów, np. klient wiarygodny, klient niewiarygodny
- Związki (korelacja i przyczynowość) - związki wielowymiarowe i jednowymiarowe
  $\text{wiek}(X, 0 \ldots 29) \wedge \mathit{dochod}(X, 20 \ldots 29K) \Rightarrow \mathit{kupuje}(X, PC)[\mathit{wsparcie} = 2\%, \mathit{wiarygodnosc} = 60\%]$
  $\text{posiada}(T, \mathit{komputer}) \Rightarrow \mathit{posiada}(T, \mathit{software})[1\%, 75\%]$

Funkcjonalność eksploracji danych cd

- Klasyfikacja i przewidywanie- znajdowanie modeli (funkcji) opisujących i wyróżniających klasy lub pojęcia (w celu przewidywania przyszłości), np. klasyfikuj kraje na podstawie klimatu lub samochody na podstawie zużycia paliwa
- Prezentacja: drzewa decyzji, reguły klasyfikacji, sieci neuronowe
- Przewidywanie: prognozuj nieznane lub brakujące wartości liczbowe
- Grupowanie (klasteryzacja)- znajdź schemat grupowania w klasy – np. pogrupuj domy tak, aby odkryć rozkład podaży
- Grupowanie oparte na zasadzie: maksymalizuj podobieństwo w ramach klasy i minimalizuj podobieństwo pomiędzy klasami

Funkcjonalność eksploracji danych cd

- Analiza wyjątków -wyjątek (outlier): obiekt (dana), który odbiega od ogólnego schematu lub zachowania się, może być szumem lub przypadkiem szczególnym, lecz wyjątek jest użyteczny do wykrywania działań niezgodnych z prawem lub normami
- Analiza trendów i tendencji rozwoju - analiza regresji, odkrywanie sekwencji zdarzeń, analiza cykliczności, wyszukiwanie podobieństw
- Inne rodzaje złożonych analiz statystycznych

Czy wszystkie prawidłowości są interesujące?

- System eksploracji może odkryć setki prawidłowości, lecz nie wszystkie muszą być interesujące. Proponowane podejście: eksploracja zorientowana na użytkownika, skoncentrowana na jego celach
- Miary przydatności: prawidłowość jest interesująca jeśli łatwo ją zrozumieć, można ją walidować na nowych lub testowych danych, jest potencjalnie użyteczna, nowa lub potwierdza hipotezę, którą użytkownik stara się zbadać
- Miary obiektywne i subiektywne:
  Obiektywne: statystyczne, strukturalne - np. wsparcie, wiarygodność
  Subiektywne: oparte na przekonaniu użytkownika – np. nowość

Wstępne przygotowanie danych

- wybór informacji z bazy danych - jakie dane są interesujace, być może agregacja danych z różnych baz, jakie typy poszczególnych atrybutów (numeryczne - rzeczywiste, numeryczne - całkowite, dane kategoryczne, logiczne, tekstowe itp)
- problemy: dane przestarzałe (np. ceny przed denominacją złotego w 1995),
- rekordy z brakującymi polami,
- punkty oddalone (ang. outliers)- wyjątki,
- dane w formacie nieodpowiednim dla modeli DM

Przyjmuje się, że wstępna obróbka danych może przekraczać połowę czasu przeznaczonego na cały proces eksploracji danych.

Problem brakujących danych

Jakie są sposoby radzenia sobie z brakującymi danymi

- pomijanie podczas analizy rekordów z brakującymi danymi - niebezpieczne, bo dane mogą tworzyć istotne wzorce; ponadto, marnuje się byc może wartościowe pozostałe dane w rekordach,
- zastępowanie brakujących wartości stałą określoną przez analityka,
- zastępowanie wartością średnią,
- zastępowanie wartością losową z odpowiedniego przedziału- wszystkie powyższe sposoby zastępowania są ryzykowne i mogą zaburzyć analizę

Przekształcanie danych - normalizacja

Atrybuty mają na ogół zakresy, które bardzo różnią się od siebie, np. w powyższych przykładach. Dla pewnych algorytmów atrybuty o dużych zakresach mogą mieć nadmierny wpływ na wyniki.
Stosuje się normalizację: normalizacja min-max polega na wykonaniu przekształcenia

$$new(X) = \frac{X - min(X)}{max(X) - min(X)}$$

Znormalizowane wartości będą należeć do przedziału od 0 do 1.

Przekształcanie danych - standaryzacja

Standaryzacja przekształca dane w taki sposób, aby średnia wartość była 0, zaś odchylenie standardowe równe 1:

$$new(X) = \frac{X - srednie(X)}{\sigma(X)}$$

Po wykonaniu standaryzacji wartość średnia będzie równa 0, odchylenie standardowe 1.

Przypomnienie wzorów

Załóżmy, że mamy n wartości liczbowych atrybutu: $a_1, \ldots, a_n$.
Wartość średnia:

$$\bar{a} = \frac{a_1 + \cdots a_n}{n}$$

odchylenie standardowe

$$\sigma = \sqrt{\frac{(a_1 - \bar{a})^2 + \cdots (a_n - \bar{a})^2}{n-1}}$$

Zbiór danych churn z UCI Repository

- Termin *churn* jest używany, aby wskazać klienta, który rezygnuje z usług jednej firmy na rzecz innej.
- Zbiór danych składa się z 3333 wierszy określających klientów, każdy klient jest opisany przez 21 atrybutów.

Zbiór danych churn z UCI Repository, cd

- *Stan* (atrybut nominalny), *Czas współpracy* (atrybut całkowity), *Kod* (obszaru-wartość nominalna), *Telefon* (numer, zastepuje ID klienta),
- *Plan międzynarodowy* (wartosc tak-nie), *Poczta głosowa* (tak-nie), *Liczba wiadomości* (atrybut całkowity),
- *Dzień minuty*, *Dzień rozmowy*, *Dzień opłata* (wartość rzeczywista),
- *Wieczór minuty*, *Wieczór rozmowy*, *Wieczór opłata* (wartość rzeczywista),
- *Noc minuty*, *Noc rozmowy*, *Noc opłata* (wartość rzeczywista),
- *Międzynarodowe minuty*, *Międzynarodowe rozmowy*, Międzynarodowe opłata (wartość rzeczywista),
- *Liczba rozmów z BOK* (liczba połaczeń z biurem obsługi), *Churn* (atrybut logiczny - czy klient zrezygnował)

Przykładowy wiersz

KS,128,415,382–4657,
nie, tak, 25,
265.10,110,45.07, dzien
197.40, 99, 16.78, wieczor
244.70, 91,11.01,noc
10.00, 3, 2.70, miedzynarodowe
1, Fałsz

Pełen zbiór danych w pliku churnPl.txt

Jakie związki między atrybutami?

Unikać skorelowanych zmiennych w eksploracji danych - użycie skorelowanych danych może wyolbrzymić jakąś część danych i dać niepewne wyniki.
W zbiorze churn zbadać zależności między następującymi atrybutami

- *dzień minuty* i *dzień rozmowy*,
- *dzień rozmowy* i *dzień opłata*,
- *dzień minuty* i *dzień opłata*,

oraz analogiczne zależności dla atrybutów dotyczących rozmów wieczorem, w nocy i międzynarodowych.
Wykrycie skorelowanych zmiennych oznacza, że w modelu mamy do czynienia z nadmiarowością i pewnych atrybutów należy się pozbyć.

Przypomnienie wzorów,cd

Załózmy, że mamy n wartości liczbowych dwóch atrybutów: $x_1, \ldots, x_n$ oraz $y_1, \ldots, y_n$.
Wartość współczynnika korelacji liniowej Pearsona na podstawie n elementowej próbki:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \times \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Przypomnienie wzorów, cd

Interpretacja współczynnika korelacji liniowej Pearsona

- $$r \in [-1, 1]$$
- współczynnik korelacji jest miarą związku liniowego, $r = 0$ oznacza brak zależności liniowej, na tej podstawie nie można wnioskować o niezależności zmiennych,
- gdy $r > 0$ - korelacja dodatnia - wzrostowi wartości zmiennej x towarzyszy wzrost wartości zmiennej y,
- gdy $r < 0$ - korelacja ujemna - wzrostowi wartości zmiennej x towarzyszy spadek wartości zmiennej y,
- im $|r|$ bliższy 1 tym zależność liniowa jest silniejsza,

Kwadrat Anscomba

John Anscombe (1918-2001) podał przykład uzasadniający rolę wykresów, jako waznego elementu analizy danych.

Podał 4 zbiory po dwie serie danych (X_i, Y_i) oraz badał zależność wykorzystując współczynnik korelacji Pearsona, równy dla wszystkich 4 zbiorów, choć zależności między zmiennymi są różne.

Odpowiednie dane i wykresy w pliku kwadratAnscombe.xls

Poszukiwanie związków między wartościami atrybutów w zbiorze churn

Warto sprawdzić

- czy wartość zmiennej churn zależy od wartości atrybutu Plan międzynarodowy,
- jaka jest zależność churn od atrybutu liczba rozmów z BOK
- prześledzić histogram atrybutu - Wieczór minuty - klienci z dużą liczbą wykorzystanych minut wieczorem mają małą tendencję do rezygnacji,
- klienci z dużą liczbą wykorzystanych minut w ciągu dnia częściej rezygnują.

Odległość - metryka

Metryka d jest funkcją o wartościach rzeczywistych spełniającą następujące trzy warunki:

1. $d(x, y) \geq 0$ oraz $d(x, y) = 0 \iff x = y$
2. $d(x, y) = d(y, x)$
3. $d(x, z) \leq d(x, y) + d(y, z)$

na przykład dla atrybutów liczbowych:
odległość euklidesowa

$$d(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$$

metryka miasto (metryka Manhattan)

$$d(x, y) = \sum_i |x_i - y_i|$$

Metryki dla różnych typów atrybutów

Metryka VDM (Value Diffrence Metric: Stanfill, Walz 1986) dla atrybutów nominalnaych (inaczej kategorycznych).
Niech: C oznacza zbiór klas, a - ustalony atrybut, x, y - wartości atrybutu

$$vdm_a(x,y) = \sum_{c=1}^{C} |\frac{N_{a,x,c}}{N_{a,x}} - \frac{N_{a,y,c}}{N_{a,y}}|$$

gdzie $N_{a,x}$ oznacza liczbę przykładów, które dla atrybutu a przyjmują wartość x,
gdzie $N_{a,x,c}$ oznacza liczbę przykładów, które dla atrybutu a przyjmują wartość x i pochodzą z klasy c,

Metryka HEOM

Metryka HEOM (Heterogenenous Value Diffrence Metric) może być użyta dla przykładów, których część atrybutów jest liczbowa i część kategoryczna:

$$HEOM(x,y) = \sqrt{\sum_{a=1}^{m} d_a^2(x_a, y_a)}$$

$$d_a(x,y) = \begin{cases} 1 & \text{jezeli } x \text{ lub } y \text{ nieokreślone, inaczej} \\ vdm_a(x,y) & \text{atrybut } a \text{ kategoryczny} \\ diff_a(x,y) & \text{atrybut } a \text{ jest liczbowy} \end{cases}$$

$$diff_a(x,y) = \frac{|x-y|}{4 \cdot \sigma_a}$$

σ_a - odchylenie standardowe a

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

Metryka HEOM przykład

Załóżmy, że korzystajac z danych z przykładu należy obliczyć wartosć metryki HEOM dla rekordów:
X=(wiek<=30, dochod=sredni, stud=tak, zdolKred=umiar)
Y=(wiek<=30, dochod=sredni, studt=nie, zdolKred=umiar)

$$HEOM(X,Y) = \sqrt{0^2 + 0^2 + (|\frac{6}{7} - \frac{3}{7}| + |\frac{1}{7} - \frac{4}{7}|)^2 + 0^2} = \frac{6}{7}$$

wszystkich wierszy: 14, klasa 'tak' - 9, klasa 'nie '- 5