





PLATFORMA PRZYSZŁOŚCI 4.0

Wprowadzenie do metod eksploracji danych w odkrywaniu wiedzy




Eksploracja danych

- Wiele systemów informacyjnych generuje i magazynuje olbrzymie ilości danych, nierzadko wielkości peta bajtów (n.p. NASA's Earth Observation System (EOSDIS) gromadzi dane rzędu 10^{15} bajtów rocznie (zdjęcia satelitarne)).
- Zasoby danych tego rodzaju mogą stanowić olbrzymi potencjał wartościowej wiedzy.
- Jednakże... ich ogrom przekracza ludzkie możliwości analizowania i odkrywania wzorców.
- Eksploracja danych (data mining) jest dziedziną zajmującą się efektywnym odkrywaniem nowych, interesujących i pożytecznych wzorców z wielkich zasobów danych.




Etapy odkrywania wiedzy z danych

- Gromadzenie danych (np. w bazach danych, hurtowniach danych, logach...)
- Wstępne przetwarzanie danych (czyszczenie, integracja, zastępowanie wartości brakujących, agregacja rekordów, tworzenie atrybutów wyprowadzonych, ...)
- Eksploracja danych (której rezultatem jest wiedza odkryta z eksplorowanych danych)**
- Ocena uzyskanych wyników
- Prezentacja i/lub (wielokrotne) użycie uzyskanych (wybranych) rezultatów



Zadania eksploracji danych

- Klasyfikacja
- Predykcja
- Grupowanie pojęciowe/segmentacja i wykrywanie wyjątków
- Znajdowanie reguł asocjacyjnych, epizodów i reguł epizodycznych, wzorców sekwencyjnych
- Wyszukiwanie reduktów
- ...



Zapalczywa klasyfikacja „nielojalnych” klientów...


Id klienta	Dzwoni	Wiek	Żonaty	Zmiana
2	Często	Średni	Tak	Nie
4	Rzadko	Senior	Nie	Nie
5	Często	Senior	Nie	Nie
6	Często	Senior	Tak	Nie
7	Rzadko	Średni	Tak	Nie
1	Rzadko	Średni	Nie	Tak
3	Często	Młody	Nie	Tak

Do której klasy należy zaklasyfikować klienta: (Dzwoni = Rzadko, Wiek = Senior, Żonaty = Nie)?

Drzewo decyzyjne:

```

    graph TD
        A[Żonaty?] -- Nie --> B[Wiek?]
        A -- Tak --> C[Zmiana = Nie]
        B -- Młody --> D[Zmiana = Tak]
        B -- Średni --> E[Zmiana = Tak]
        B -- Senior --> F[Zmiana = Nie]
    
```



Zapalczywa klasyfikacja „nielojalnych” klientów...

Id klienta	Dzwoni	Wiek	Żonaty	Zmiana
2	Często	Średni	Tak	Nie
4	Rzadko	Senior	Nie	Nie
5	Często	Senior	Nie	Nie
6	Często	Senior	Tak	Nie
7	Rzadko	Średni	Tak	Nie
1	Rzadko	Średni	Nie	Tak
3	Często	Młody	Nie	Tak

Do której klasy należy zaklasyfikować klienta: (Dzwoni = Rzadko, Wiek = Senior, Żonaty = Nie)?

Drzewo decyzyjne:

```

    graph TD
        A[Wiek?] -- Młody --> B[Zmiana = Tak]
        A -- Średni --> C[Żonaty?]
        A -- Senior --> D[Zmiana = Nie]
        C -- Nie --> E[Zmiana = Tak]
        C -- Tak --> F[Zmiana = Nie]
    
```

Zapalczywa klasyfikacja „niełojalnych” klientów

Id klienta	Dzwoni	Wiek	Żonaty	Zmiana
1	Rzadko	Średni	Nie	Tak
2	Często	Średni	Tak	Nie
3	Często	Młody	Nie	Tak
4	Rzadko	Senior	Nie	Nie
5	Często	Senior	Nie	Nie
6	Często	Senior	Tak	Nie
7	Rzadko	Średni	Tak	Nie

Reguły decyzyjne:

- Jeśli (Wiek = Młody), to (Zmiana = Tak) [1, 100%]
- Jeśli (Wiek = Senior), to (Zmiana = Nie) [3, 100%]
- Jeśli (Wiek = Średni) i (Żonaty = Nie), to (Zmiana = Tak) [1, 100%]
- Jeśli (Wiek = Średni) i (Żonaty = Tak), to (Zmiana = Nie) [2, 100%]
- Jeśli (Wiek = Średni), to (Zmiana = Nie) [2, 67%]

Leniwa klasyfikacja „niełojalnych” klientów z wykorzystaniem k najbliższych sąsiadów

Do której klasy należy zaklasyfikować klienta:
(Dzwoni = Rzadko, Wiek = Senior, Żonaty = Nie)?

Id klienta	Dzwoni	Wiek	Żonaty	Zmiana
2	Często	Średni	Tak	Nie
4	Rzadko	Senior	Nie	Nie
5	Często	Senior	Nie	Nie
6	Często	Senior	Tak	Nie
7	Rzadko	Średni	Tak	Nie
1	Rzadko	Średni	Nie	Tak
3	Często	Młody	Nie	Tak

Jego k=3 najbliższych sąsiadów:

- w klasie dec. Zmiana = Nie: 2 obiekty (#4 i #5)
- w klasie dec. Zmiana = Tak: 1 obiekt (#1)

Inne zastosowania klasyfikacji...

- Firma telekomunikacyjna może użyć klasyfikatora np. do klasyfikowania nowych lub potencjalnych klientów do następujących klas:
 - bardzo aktywny, aktywny, bierny,
 - reagujący na kampanie reklamowe i ignorujący je.
- To z kolei może wskazywać jakie ogłoszenia i kampanie promocyjne należy do nich kierować. Wiedza tego typu może znacząco zmniejszyć koszty marketingowe i wpłynąć na pozyskanie nowych klientów.

Inne zastosowania klasyfikacji

- Klasyfikatory mogą wspierać wykrywanie potencjalnych oszustów, którzy prawdopodobnie nie wniosą opłat za żądane usługi.
- Klasyfikatory mogą pomóc w wykrywaniu nielegalnych połączeń (wykonywanych ze skradzionego telefonu lub z użyciem skradzionej karty...).

Predykcja ruchu w komórce...

Id komórki	Populacja	Dochód	TypTerenu	Ruch
2	1847	1800	9870	2,53
1	5146	2727	1250	6,79
3	6465	1500	1500	7,22
6	7653	1850	1780	8,22
7	7257	1900	2500	9,95
4	15702	1700	3900	11,18
5	12799	1750	5000	19,93

Wielokrotna regresja liniowa:
Obciążenie = $(\alpha \times \text{Populacja}) + \beta$

Regresja wielomianowa:
Obciążenie = $(\alpha_2 \times \text{Populacja}^2) + (\alpha \times \text{Populacja}) + \beta$

Predykcja ruchu w komórce

Id komórki	Populacja	Dochód	TypTerenu	Ruch
2	1847	1800	9870	2,53
1	5146	2727	1250	6,79
3	6465.00	1500	1500	7,22
6	7653	1850	1780	8,22
7	7257	1900	2500	9,95
4	15702	1700	3900	11,18
5	12799	1750	5000	19,93

Sieć neuronowa:

- ♦ $w1 = f_{akt}((\alpha_{P-w1} \times P) + (\alpha_{D-w1} \times D) + (\alpha_{T-w1} \times T) + \theta_{w1})$
- ♦ ...
- ♦ $w4 = f_{akt}((\alpha_{w1-w4} \times w1) + (\alpha_{w2-w4} \times w2) + (\alpha_{w3-w4} \times w3) + \theta_{w4})$
- ♦ $R = w4$.

Grupowanie

- Obiekty położone blisko w przestrzeni wielowymiarowej powinny być przypisane do tego samego klastra (grupy/segmentu).
- Obiekty położone daleko w przestrzeni wielowymiarowej powinny być przypisane do różnych klastrów.
- Wyjątki – obiekty odległe od (prawie) wszystkich pozostałych.
- Wyjątki - szum lub ... potencjalnie bardzo interesujące anomalie (np. użyteczne przy wykrywaniu oszustw finansowych).

13

Reguły asocjacyjne, wzorce sekwencyjne

- Przykład reguły asocjacyjnej:
Jeśli klient prowadzi firmę, to wybiera plan taryfy biznesowej z prawdopodobieństwem 95% i wsparciem 25%.
- Przykład wzorca sekwencyjnego:
45% klientów korzystających z usług firmy tel. X, po wybraniu usługi "do komórek za 19 zł", zmienia plan taryfowy w przeciągu 2 miesięcy na „wszystko za 29 zł”, a w przeciągu następnych 2 miesięcy zażąda usługi „dołącz drugi numer”.

Klient	Czas	Zdarzenie
1	01.05.17	do komórek za 19
1	01.06.17	wszystko za 29
1	01.07.17	wszystko za 39
1	15.08.17	dołącz drugi numer
2	01.06.17	do komórek za 19
2	01.08.17	wszystko za 29
2	01.10.17	dołącz drugi numer
3	01.04.17	wszystko za 29
...		

14

Epizody i reguły epizodyczne na potrzeby zarządzania alarmami w sieci

- 4 fińskie firmy telekomunikacyjne:
 - Nokia Telecommunications,
 - Helsinki Telephone Corp. HPY,
 - Radiolinja and Tampere Telephone Corp.)
 - the Technology Development Centre of Finland (Tekes))
 prowadziły w latach 1994-1997 współpracę z Uniwersytetem w Helsinkach w zakresie odkrywania epizodów i reguł epizodycznych.
- W rezultacie powstał system TASA (Telecommunication Alarm Sequence Analyzer) wykrywający wiedzę typu:


```
IF (alarm połączeniowy) AND NEXT(nieudane połączenie),
            THEN (alarm o wysokim współczynniku uszkodzenia) [5] [60]
            zaufanie [90%] częstość [151/168].
```
- Znaczenie reguły: w 90% przypadków, jeśli w przeciągu 5s. wystąpił najpierw alarm połączeniowy, a następnie zarejestrowano nieudane połączenie, to w przeciągu 60s. wystąpił alarm o wysokim współczynniku uszkodzenia. Wszystkie 3 zdarzenia wystąpiły razem 151 razy, a 2 zdarzenia z części IF reguły wystąpiły 168 razy.

Odkrywanie zależności funkcyjnych i przybliżonych w dużych bazach danych...

- Ustalenie zależności funkcyjnych jest jednym z niezbędnych, podstawowych zadań w procesie projektowania bazy danych.
- Przykład zależności funkcyjnej:
 - pesel → imię, nazwisko;
 - kod pocztowy → miasto;
 - miasto, ulica → kod pocztowy.
- Próby automatycznego odkrywania zależności funkcyjnych napotykały na następujące problemy:
 - mało danych → prawdziwe, ale i nieprawdziwe zależności funkcyjne;
 - dużo danych → brak efektywnych, skalowalnych metod;
 - przekłamanie w danych → problemy z wykrywaniem prawdziwych zależności.

Odkrywanie zależności funkcyjnych i przybliżonych w dużych bazach danych

- Potrzeba odkrywania zależności przybliżonych
- imię → płeć
 - Ale Jan Maria Rokita jest mężczyzną...
 - Jednak imię zazwyczaj właściwie określa płeć!
- TANE (Huhtala et al.) i Dep-miner (Lopes et al.) – zaproponowali efektywne, skalowalne algorytmy odkrywania zależności funkcyjnych i przybliżonych z dużych baz danych.
- W algorytmach tych umiejętnie wykorzystano techniki odkrywania wzorców częstych.

nazwisko	imię	płeć
Adamska	Maria	k
Kowalska	Maria	k
Pośnik	Maria	k
Rokita	Maria	m

Redukcja atrybutów wg teorii zbiorów przybliżonych

Redukt


Id klienta	Dzwoni	Wiek	Żonaty	Zmiana
1	Rzadko	Średni	Nie	Tak
2	Często	Młody	Nie	Tak
3	Rzadko	Senior	Tak	Tak
4	Często	Senior	Nie	Tak
5	Często	Senior	Nie	Nie
6	Rzadko	Średni	Tak	Nie
7	Często	Średni	Tak	Nie

Przybliżenie dolne

Przybliżenie górne

Obszar brzegowy

18




Języki eksploracji danych

- Rozwój i standaryzacja języków eksploracji danych, które mają umożliwić użytkownikowi specyfikację poszukiwanej wiedzy.
- Przykład. Pracownik firmy telekomunikacyjnej chce poznać przyczyny częstych skarg, które występują z prawdopodobieństwem co najmniej 65% w co najmniej 10% zarejestrowanych przypadków.


```
SELECT FROM MINE(T) R
WHERE R.ANTECEDENT >= {kategoria = *}
AND R.CONSEQUENT = {częstość_skarg = "wysoka"}
AND R.SUPPORT > 10% AND R.CONFIDENCE >= 65%;
```

- Przykładowe wyniki:
 - IF kategoria IN {"niski poziom sygnału"} AND od = {"aktywacja"},
 - THEN częstość_skarg = "wysoka" [wsparcie = 12%, zaufanie = 95%].
 - IF kategoria IN {"echo"} AND od = {"miesiąc"},
 - THEN częstość_skarg = "wysoka" [wsparcie = 15%, zaufanie = 70%].




Podsumowanie

- Odkrywanie wiedzy jest dziedziną zorientowaną na zastosowania, w której problemy badawcze są często motywowane dostępnością i specyfiką zasobów informacyjnych o świecie rzeczywistym.
- Proces odkrywania wiedzy często wymaga współpracy analityków, ekspertów w dziedzinie rozważanego problemu, informatyków i statystyków.
- Innowacyjne metody eksploracji danych są coraz powszechniej używane m. in. w zakresie:
 - prowadzenia skutecznej działalności marketingowej,
 - wykrywania przypadków nielegalnego korzystania ze świadczonych usług,
 - zapobiegania utracie klientów,
 - rozpoznawania potrzeb klientów,
 - budowy, pielęgnacji i eksploatacji sieci.



Literatura

- Han J., Kamber M., Pei, J., Data Mining: Concepts and Techniques, The Morgan Kaufmann Series in Data Management Systems, 3rd edition, Morgan Kaufmann, 2011
- Morzy T., Eksploracja danych, Metody i algorytmy, Wydawnictwo Naukowe PWN, 2013
- Kryszkiewicz M., Zastosowania i trendy rozwoju metod odkrywania wiedzy, ICS Research Report 9/2003, Warsaw, June 2003



Ćwiczenia

- Jakie etapy składają się na proces odkrywania wiedzy z danych?
- Jakie są zadania eksploracji danych?