



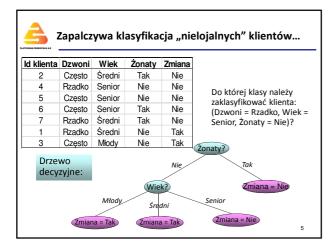
Etapy odkrywania wiedzy z danych

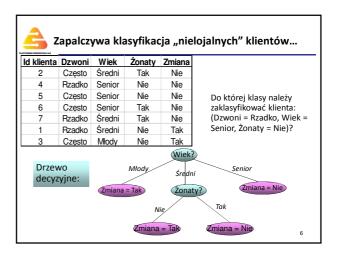
- Gromadzenie danych (np. w bazach danych, hurtowaniach danych, logach...)
- Wstępne przetwarzanie danych (czyszczenie, integracja, zastępowanie wartości brakujących, agregacja rekordów, tworzenie atrybutów wyprowadzonych, ...)
- Eksploracja danych (której rezultatem jest wiedza odkryta z eksplorowanych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

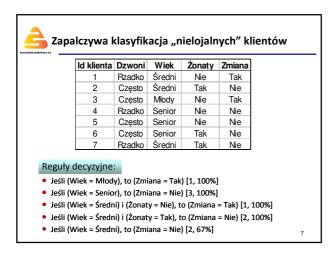


Zadania eksploracji danych

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











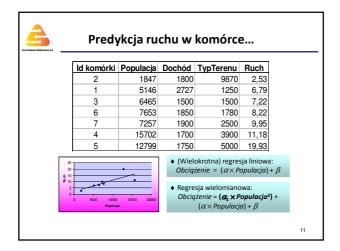
Inne zastosowania klasyfikacji...

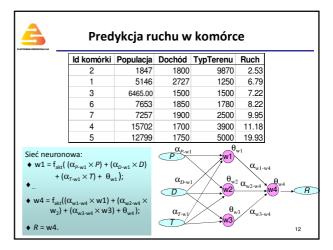
- Firma telekomunikacyjna może użyć klasyfikatora np. do klasyfikowania nowych lub potencjalnych klientów do następujących klas:
 - bardzo aktywny, aktywny, bierny,
- reagujący na kampanie reklamowe i ignorujący je.
- To z kolei może wskazywać jakie ogłoszenia i kampanie promocyjne należy do nich kierować.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

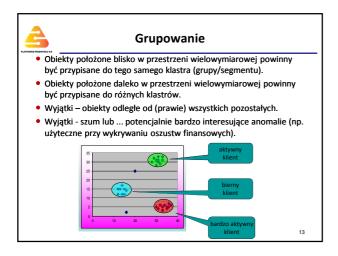


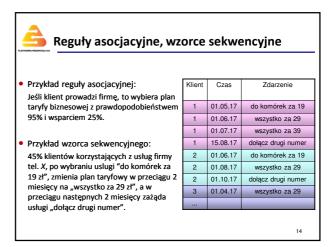
Inne zastosowania klasyfikacji

- 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...).











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 wystąpił razem 151 razy, a 2 zdarzenia z części IF reguły wystąpiły 168 razy.



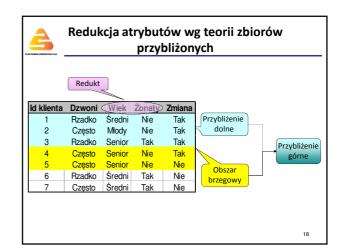
Odkrywanie zależności funkcyjnych i przybliżonych w dużych bazach danych...

- Ustalenie zależności funkcyjnych jest jednym z niezbędnych, podstawowych zadań w procesie projektowania bazy danych.
- Przykład zależności funkcyjnej:
- $\mathsf{pesel} \to \mathsf{imie},\, \mathsf{nazwisko};$
- $\begin{array}{l} \mathsf{kod}\;\mathsf{pocztowy} \to \mathsf{miasto;}\\ \mathsf{miasto,}\;\mathsf{ulica} \to \mathsf{kod}\;\mathsf{pocztowy.} \end{array}$
- Próby automatycznego odkrywania zależności funkcyjnych napotykały na następujące problemy:
- $\mbox{mało danych} \rightarrow \mbox{prawdziwe, ale i nieprawdziwe zależności funkcyjne;}$
- dużo danych → brak efektywnych, skalowalnych metod; przekłamania 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.





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

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