

Eksploracja danych – wybrane metody i zastosowania w analizie danych

Jerzy Stefanowski
Instytut Informatyki
Politechnika Poznańska



Wykład wstępny dla spec. PIESI
Poznań 2007

1 1

Data Mining oraz KDD

1. **Wprowadzenie**
2. **Terminologia**
3. **Proces odkrywanie wiedzy**
4. **Wybrane metody**
5. **Podsumowanie**

2

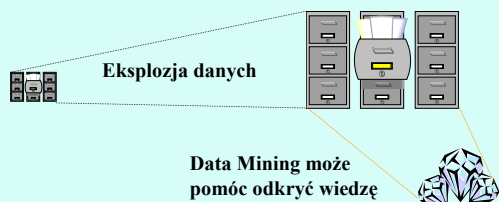
Motywacje



1. Rozwój technologii automatycznego gromadzenia i przechowywania informacji
→ wzrost rozmiarów przechowywanych danych!
Nie tylko w dużych przedsiębiorstwach.
2. Dostęp do informacji nie jest równoznaczny z posiadaniem wartościowej **wiedzy**.
3. Wyzwaniem jest nie tylko efektywne przechowywanie danych, lecz także ich analiza, zdolność interpretacji i wyciągania użytecznych wniosków, które mogą prowadzić do lepszych decyzji!

3

„We are Data Rich but Knowledge Poor”



4

Potrzeba matką wynalazku



- Istnieje zapotrzebowanie na narzędzia pozwalające na automatyczną analizę gromadzonych danych → nowa dziedzina *Odkrywanie Wiedzy* (ang. *Knowledge Discovery*)
- Inne terminy: *Eksploracja Danych* (ang. *Data Mining*), *Zgłębianie Danych*,
- Rozwój historyczny:
 - 1989 Workshop on Knowledge Discovery in Databases (USA)
 - Statystyka i maszynowe uczenie się - wcześniej

5

Czym jest odkrywana Wiedza?



“Wiedza jest uporządkowanym zbiorem *interesujących i użytecznych* regularności”

G. Piatetsky-Shapiro (1991)

- *Regularność* (wzorzec, ang. *pattern*) – zależność między elementami danych,
- *Użyteczny* – może prowadzić do użytecznych działań,
- *Interesujący* – nowy (poprzednio nieznan i nieoczekiwany) nietrywialny i zrozumiały.



6

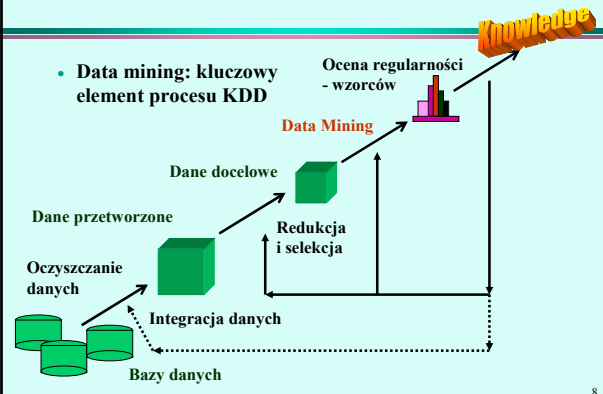
Odkrywanie wiedzy w danych



- Odkrywanie wiedzy to nietrywialny proces poszukiwania nowych (nieoczekiwanych), potencjalnie użytecznych i zrozumiałych regularności z danych.
- Data mining (eksploracja danych, zgłębianie danych) ? Istotny etap wewnątrz procesu odkrywania wiedzy

7

Proces Odkrywania Wiedzy - KDD



8

Etapy procesu odkrywania wiedzy

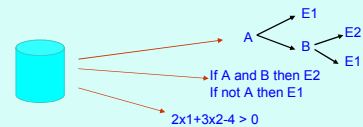


- Analiza i poznanie dziedziny zastosowania, identyfikacja dostępnej wiedzy i celów użytkownika,
- Wybór danych związanych z celami procesu,
- Czyszczenie i wstępne przetwarzanie danych oraz ich redukcja,
- Wybór zadań i algorytmów eksploracji danych,
- Pozyskiwanie wiedzy z danych (krok eksploracji danych),
- Interpretacja i ocena odkrytej wiedzy,
- Przygotowanie wiedzy do użycia.



9

Eksploracja danych



- Data mining** (ang.)
 - Poszukiwanie w zgromadzonych danych nieznanych, użytecznych regularności, związków między elementami danych.
- Eksploracja danych to **etap** w procesie odkrywania wiedzy.
- Wiedza** → uporządkowana i formalna reprezentacja odkrytej regularności między elementami danych.

10

Eksploracja danych a inne systemy informatyczne

Informacje / bazy danych	Zapytania SQL i raporty	<ol style="list-style-type: none"> Który klient dokonał największego zakupu? Lista klientów, którzy zakupili produkt A w ostatnim roku?
Hurtownie danych	Wielowymiarowa agregacja danych i podsumowania	<ol style="list-style-type: none"> Jakie są średnie zakupy klientów, którzy kupili produkt A w ostatnim roku, z podziałem na regiony?
Zaawansowane systemy - „Data mining”	Predykcja lub opis	<ol style="list-style-type: none"> Jakie są cechy charakterystyczne klientów, którzy mogą kupić produkt A? Do kogo skierować ofertę reklamową?

11

Przykłady zastosowań eksploracji danych

- Marketing**
 - „Target marketing”, identyfikacja profilu klientów, ocena lojalności klientów, problem koszyka zakupów - asocjacje produktów w sieciach sprzedaży, segmentacja rynków, klientów, itp.
- Analizy finansowe**
 - Analiza ryzyka kredytowego, rekomendacje produktów, przewidywanie trendów i przebiegów czasowych,...
- Wykrywanie nieprawidłowości i anomalii**
 - Analiza defraudacji i nieprawidłowości kart kredytowych, systemy telekomunikacyjne, towarzystwa ubezpieczeniowe, systemy opieki medycznej.
- Text mining oraz Web mining** (zachowania użytkowników w e-serwisach, wspomaganie wyszukiwania informacji), ...
- Wiele innych** (przemysł, nauka, administracja),...

12

Typowe zadania w KDD:

1. Podsumowywanie danych (tzn. znajdowanie zwięzłych opisów lub ogólnych własności pewnych klas obiektów).
2. Odkrywanie asocjacji (zależności lub korelacji między elementami danych).
3. Klasyfikowanie (zmienna wyjściowa jest jakościowa).
4. Predykcja (zmienna wyjściowa jest liczbową).
5. Grupowanie danych (analiza skupień).
6. Poszukiwanie obserwacji osobliwych, anomalii,...
7. Analiza złożonych typów danych (wielo-relacyjne zależności logiczne, multimedialne, przebiegi czasowe, grafy, itp.).
8. Text i Web mining.
9. Analiza danych strumieniowych, ciągle napływających z sieci sensorów.

13

Połączenie wielu dziedzin

- Statystyka
- Maszynowe Uczenie się
- Wizualizacja danych
- Systemy baz danych, hurtownie danych, techniki OLAP
- Inne dziedziny:
 - wyszukiwanie informacji, obliczenia równoległe, przetwarzanie obrazów, itp.

14

Odkrywanie wiedzy a inne dziedziny – cz.1.

- **Statystyka:**
 - Oparta na silnych podstawach teoretycznych i mocnych założeniach co do danych.
 - Ukierunkowana na testowanie hipotez oraz estymację parametrów.
- **Uczenie maszynowe:**
 - Ukierunkowane na polepszanie działania przez uczącego się agenta.
 - Uczenie się pojęć lub zadań (lepiej zdefiniowane niż w KDD).
 - Adaptacja do rzeczywistego i zmiennego środowiska → np. robotów (nie rozważane w KDD).
 - Większa rola metod heurystycznych i wywodzących się ze sztucznej inteligencji.

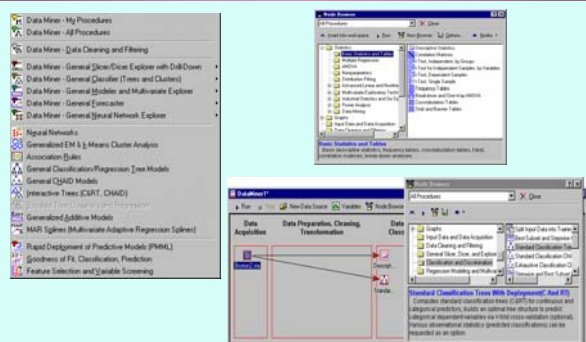
15

Odkrywanie wiedzy a inne dziedziny – cz. 2

- **Odkrywanie wiedzy**
 - Łączy modele teoretyczne i heurystyczne, lecz
 - odmienne cele,
 - Większa różnorodność i złożoność analizowanych danych (nietypowych dla statystycznej analizy danych),
 - Często brak jasnej definicji pojęć do odkrycia.
 - Nacisk na reprezentację wiedzy,
 - Inny niż poprzednio charakter procesu odkrywania wiedzy.
 - Duża rola przygotowania i wstępnego przetwarzania danych.

16

Data Miner (Statistica Statsoft) – przykład metod dostępnych w systemie



17

SAS – podstawowe algorytmy

- Przykładowe algorytmy eksploracji danych (dostępne w tzw. węzłach SAS Enterprise Miner)
 - wiele metod statystyki opisowej,
 - metody przekształceń danych (przeskalowania, uwzględnianie nieznanymi wartości, wykrywanie nietypowych obserwacji),
 - poszukiwanie reguł asocjacyjnych,
 - analiza skupień (k-średnich, sieci SOM Kohonen'a)
 - modele predykcyjne (liniowa, nieliniowa, logistyczna regresja, drzewa regresji)
 - drzewa klasyfikacyjne (CART, CHAID, C4.5like)
 - sztuczne sieci neuronowe (liniowe/nieliniowe sieci wielowarstwowe, różne wersje RBF).
 - modele złożonych klasyfikatorów (bagging, boosting, combiners,...)
 - modele k-NN
 - modele szeregów czasowych.
- Oferuje przetwarzanie danych za pomocą specjalnego języka oraz interfejsy graficzne

18

Wiedza klasyfikacyjna

- Problem określania zasad przydziału obiektów do znanych wstępnie klas na podstawie analizy danych o przykładach klasyfikacji.

Wiek	Zawód	dochód	...	Decyzja
21	Prac. fiz.	1220	...	Nie kupi
26	Menadżer	2900	...	Kupuje
44	Inżynier	2600	...	Kupuje
23	Student	1100	...	Kupuje
36	Nauczytel	1700	...	Nie kupi
...
45	Lekarz	2200	...	Nie kupi
25	Student	800	...	Kupuje

Przykłady uczące

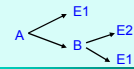
Algorytm eksploracji



Reprezentacja wiedzy:
np. reguły
R1. Jeżeli student to kupuje komputer
R2. Jeżeli dochód > 2400 ...

19

Drzewa decyzyjne



- Najpopularniejsza reprezentacja wiedzy klasyfikacyjnej.
- Rozwinięta metodologicznie → algorytmy C4.5 lub CART.
- Liczne implementacje (dostępne w wielu systemach).
- Typowe zastosowania:
 - Klasyfikacja typu klienta, np.:**
 - Ocena zdolności kredytowej klienta w inst. finansowych.
 - Podziału klientów na tych, którzy przynoszą zyski i tych, którzy przynoszą straty.
 - Identyfikacja potrzeb klienta.
 - Analiza efektywności kampanii sprzedaży.
 - Ukierunkowane kampanie reklamowe (ang. target marketing).
 - Poszukiwanie charakterystyki lojalnych klientów.

20

Indukcja reguł decyzyjnych

- Podstawowa idea - reguły poszukuje się bezpośrednio z danych
 - potencjalnie większa zrozumiałość wiedzy
 - ale więcej różnych podejść
 - opisy rzeczywistych danych minimalnym zbiorem reguł dyskryminujących/ klasyfikujących
 - poszukiwanie bardziej wyczerpujących zbiorów reguł o dobrych własnościach interpretacyjnych
 - więcej parametrów do sterowania w algorytmach indukcji reguł.
- Więcej → J.Stefanowski strona WWW z publikacjami

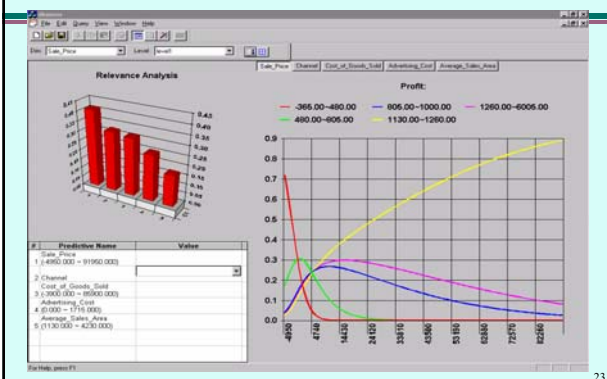
21

Inne metody klasyfikacyjne

- Sztuczne sieci neuronowe.
- Klasyfikacja bayesowska (statystyczna teoria decyzji)
 - analiza dyskryminacyjna
 - statystyczna klasyfikacja – metoda wektorów wspierających (SVM).
- Metody k-najbliższych sąsiadów.
- Regresja logistyczna.
- Klasyfikatory genetyczne.
- Metody wywodzące się z teorii zbiorów przybliżonych
- Metody oparte na logice → ILP
- Logika rozmyta.
- Inne ...

22

Predykcja: metody regresji (analityczne modele); drzewa regresji, sztuczne sieci neuronowe



23

Reguły asocjacyjne



- Reprezentują lokalne zależności między elementami danych transakcyjnych.
- Problem **koszyka zakupów**
- Znalezienie wzorców zachowań klientów poprzez wskazanie tych produktów, które są kupowane najczęściej łącznie

TID	Produce
1	MLEKO, CHLEB, JAJA
2	CHLEB, CUKIER
3	CHLEB, SER
4	MLEKO, CHLEB, CUKIER
5	MLEKO, SER
6	CHLEB, SER
7	MLEKO, SER
8	MLEKO, CHLEB, SER, JAJA
9	MLEKO, CHLEB, SER

- {SER, MLEKO} → CHLEB [sup=5%, conf=80%]
- Przykład reguły:
„80% klientów, którzy kupują ser i mleko także kupuje chleb oraz 5% klientów kupuje te produkty łącznie”

24

Wzorce sekwencyjne

- Wzorzec symboliczny opisujący zależności występujące między zdarzeniami zachodzącymi w pewnym przedziale czasu.
- Wzorzec: $A \rightarrow B \rightarrow C$, mówi że po zdarzeniu A wystąpi zdarzenie B, a po nim zdarzenie C z pewną częstością.
- Przykład:
 - 30% klientów sklepu wysyłkowego najpierw zamówiło płytę Anny Marii Jopek, później Diany Krall, a następnie Stacey Kent.

25

Reguły asocjacyjne i wzorce sekwencji

- Typowe zastosowania:
 - poszukiwania typowych wzorców zachowań klientów lub użytkowników np. w sieciach sprzedaży, telekomunikacji, ubezpieczeniach, bankowości, analizie serwerów WWW.
 - eksploracja zawartości logów serwera WWW prowadzi do zreorganizowania sposobu nawigacji po stronach, prezentacji treści, dynamicznego dostosowania wyglądu strony do profilu użytkownika.

26

Grupowanie, Analiza Skupień

- Grupowanie** (ang. clustering) jest to proces podziału zbioru danych (obiektów) na podzbiory, nazywane klasami lub skupieniami
 - Proces powinien wyodrębnić jednorodne grupy obiektów podobnych do siebie
 - Podział powinien posiadać pewne znaczenie dla użytkownika i pomóc mu zrozumieć strukturę danych
- Skupienie (Klasa)**: zbiór obiektów które są podobne do siebie i mogą być traktowane zbiorczo jako jednorodna grupa
- Grupowanie: nienadzorowana klasyfikacja - brak predefiniowanych klas decyzyjnych.

27

Podstawowe metody grupowania

- Iteracyjno-optymalizacyjne
 - Hierarchiczne (aglomeracyjne, podziału)
 - oparte na funkcjach gęstości
- Inne aspekty:
- Tradycyjne metody – numeryczne (k-średnich, AHC)
 - Uczenie Maszynowe - nacisk na symboliczną klasyfikację, automatyczne tworzenie opisu charakteryzującego tworzone klasy (conceptual clustering)
 - Eksploracja baz danych - problemy dużych rozmiarów (obiekty, atrybuty), złożone typy danych (np. grafy, teksty, obrazy)

28

Podsumowanie

- Problemem nie jest elektroniczne gromadzenie danych ale ich właściwa analiza i wyciąganie użytecznych wniosków.
- Metody statystyczne i uczenia maszynowego mogą być podstawą do odkrywania wiedzy z danych.
- Należy zwracać uwagę na wcześniejsze etapy procesu odkrywania wiedzy, np. integracji danych z różnych źródeł, czyszczenia danych, przetwarzania wstępnego oraz redukcji rozmiarów danych.
- Istnieje oprogramowanie wspierające proces odkrywania wiedzy.
- Integracja z hurtowniami danych i biznesowymi systemami wspomaganie decyzji.

29

Wybrana literatura

- Larose D.T., Odkrywanie wiedzy z danych: Wprowadzenie do eksploracji danych. (tłumaczenie z ang.) PWN 2006.
- Hand D., Mannila H., Smyth P. Principles of Data Mining, MIT Press, 2001. Tłumaczenie polskie Eksploracja danych, WNT 2005.
- Han Jiawei and Kamber M. Data mining: Concepts and techniques, Morgan Kaufmann, 2001. (Slajdy dostępne w Internecie)
- Krawiec K, Stefanowski J., Uczenie maszynowe i sieci neuronowe, Wydawnictwo PP, 2004.
- WEKA – open source project = źródło + dokumentacja i podręczniki dostępne w WWW.
- KDnuggets → bogaty serwis WWW w j. ang. / także wiele materiałów dydaktycznych.

Data mining, by J.Stefanowski



30