Drążenie danych i inne tematy

Materiały do wykładu Komputerowe Przetwarzanie Wiedzy
(opracowane na podstawie materiałów dostępnych w sieci Internet)
Tomasz Kubik

Drążenie danych (data mining)

- „Proces odkrywania istotnych zależności (korelacji), wzorców i tendencji poprzez przesiewanie dużych ilości danych przechowywanych w repozytoriach za pomocą technik rozpoznawania wzorców oraz technik statystycznych i matematycznych” [Gartner Group].
- „Automatyczne lub półautomatyczne badanie dużych ilości danych w celu odkrycia istotnych wzorców i reguł” [M. J. A. Berry i G. Linoff w książce „Data Mining Techniques”]
- W polskim tłumaczeniu:
  - drążenie danych,
  - zgłębianie danych,
  - eksploracja danych, itp.

Odkrywanie wiedzy a drążenie danych

- **Odkrywanie wiedzy** jest postrzegane jako złożony proces selekcji i transformacji danych, ich eksploracji, a następnie interpretacji uzyskanych wyników. W procesie tym dochodzi do wysoce intensywnego współdziałania człowieka i maszyny.
- **Systemem odkrywania wiedzy** nazywane jest specjalizowane oprogramowanie wspomagające człowieka w odkrywaniu wiedzy, które zazwyczaj współpracuje z systemem zarządzania bazą danych.
- **Drążenie danych** to nietrywialny proces odkrywania nowych (nieoczekiwanych), potencjalnie użytecznych regularności w danych. Drążenie danych jest jednym z elementów odkrywania wiedzy.

Drążenie danych

- Wybrane cechy
  - Dokładność często nie jest najważniejsza
  - Dane bywają niepełne i niepewne
  - Dopuszcza się m.in. stosowanie metody „czarnej skrzynki”
- Słowa kluczowe
  - Automatyzacja
  - Predykcja
  - Ukryte modele
- Dyscypliny związane z drążeniem danych
  - Systemy baz danych
  - Sztuczna inteligencja
  - Optymalizacja
  - Statystyka
  - Obliczenia równoległe

Otoczenie drążenia danych

- Dążenie do uproszczenie i automatyzacja procesów statystycznych prowadzących od źródła danych do zastosowania modelu
- Istnieje wiele różnych, dostępnych algorytmów i narzędzi
- Dobór najlepszego rozwiązania wymaga użycia statystyki
- Algorytmy posiadają wbudowaną sztuczną inteligencję

Algorytmy drążenia danych

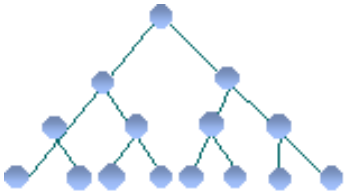
- klasyfikacja
  - to określenie wartości jednego konkretnego atrybutu – klasy, na podstawie pozostałych atrybutów, a zwłaszcza ich podzbioru, może to być dokonane w postaci jawnej np. drzewa decyzyjne, bądź niejawnej, np. sieci neuronowych, który jest następnie wykorzystywane do klasyfikowania nowych obiektów w bazie danych.
- analiza skupień (klasteryzacja)
  - Podział danych na odpowiednią, nieznaną przed rozpoczęciem procesu, liczbę grup. Elementy w grupie muszą być „bliskie” sobie, sąsiadujące grupy powinny wykazywać pewne podobieństwa (podobieństwo mierzone jest funkcją odległości).
  - może być przeprowadzona również w oparciu o reguły logiczne (zbudowne na podstawie prezentacji przykładów pozytywnych i negatywnych)

Algorytmy drążenia danych

- predykcja
  - umożliwia estymację przyszłych wartości badanych danych
  - wykorzystuje metody statystycznej prognozy, systemy ekspertowe bazujące na mechanizmach sztucznej inteligencji, metody analizy szeregów czasowych, teorię chaosu
- dyskryminacja
  - polega na znajdowaniu cech, które odróżniają wskazaną klasę obiektów (*target class*) od innych klas (*contrasting classes*)
- odkrywanie asocjacji
  - polega na znajdowaniu związków pomiędzy występowaniem grup elementów w zadanych zbiorach danych
- regresja

Przykłady metod drążenie danych

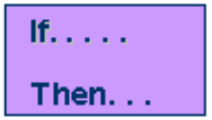
- Drzewa decyzyjne



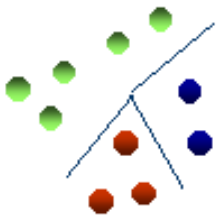
- Klasyfikator najbliższego sąsiedztwa

- Sieci neuronowe

- Indukcja reguł



- Analiza k-skupień



Czym drążenie danych nie jest

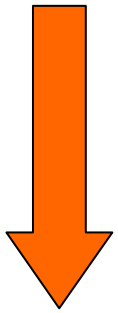
- Data warehousing
- SQL / Zapytania wprost / Raportowanie
- Agenci programowi
- Online Analytical Processing (OLAP)
- Wizualizacja danych

OLAP (*Online Analytical Processing)*

- Środowiska tego typu pozwalają na prostą analizę baz danych, w której możliwa jest wielowymiarowa obserwacja agregowanych wartości wybranych atrybutów jednej lub wielu połączonych relacji.
- Metodologia OLAP zakłada, że użytkownik przygotowuje pewną hipotezę, której poprawność weryfikuje korzystając z narzędzi OLAP (np. *Oracle Express Server*).
- Narzędzia OLAP pozwalają wyznaczać wartości pewnych parametrów (np. sum sprzedaży w przedziałach czasowych). Ich obserwacja ułatwia zauważenie pewnych zależności (np. sezonowego zwiększenia popytu).
- Ograniczenia OLAP wynikają z koniecznością przygotowywania hipotez, które podlegają późniejszej weryfikacji. Stąd dużą rolę odgrywa kreatywność i wyobraźnia eksperta.
- W podejściu tym istnieje niebezpieczeństwo akceptacji hipotez fałszywych.

Modele stosowane w drążeniu danych

- Przykładowe kryteria oceny modeli
  - dokładność
  - zrozumiałość
- Można je uszeregować od „zrozumiałych” do „niezrozumiałych”
  - Drzewa decyzyjne
  - Indukcja reguł
  - Modele regresji
  - Sieci neuronowe



zrozumiałe

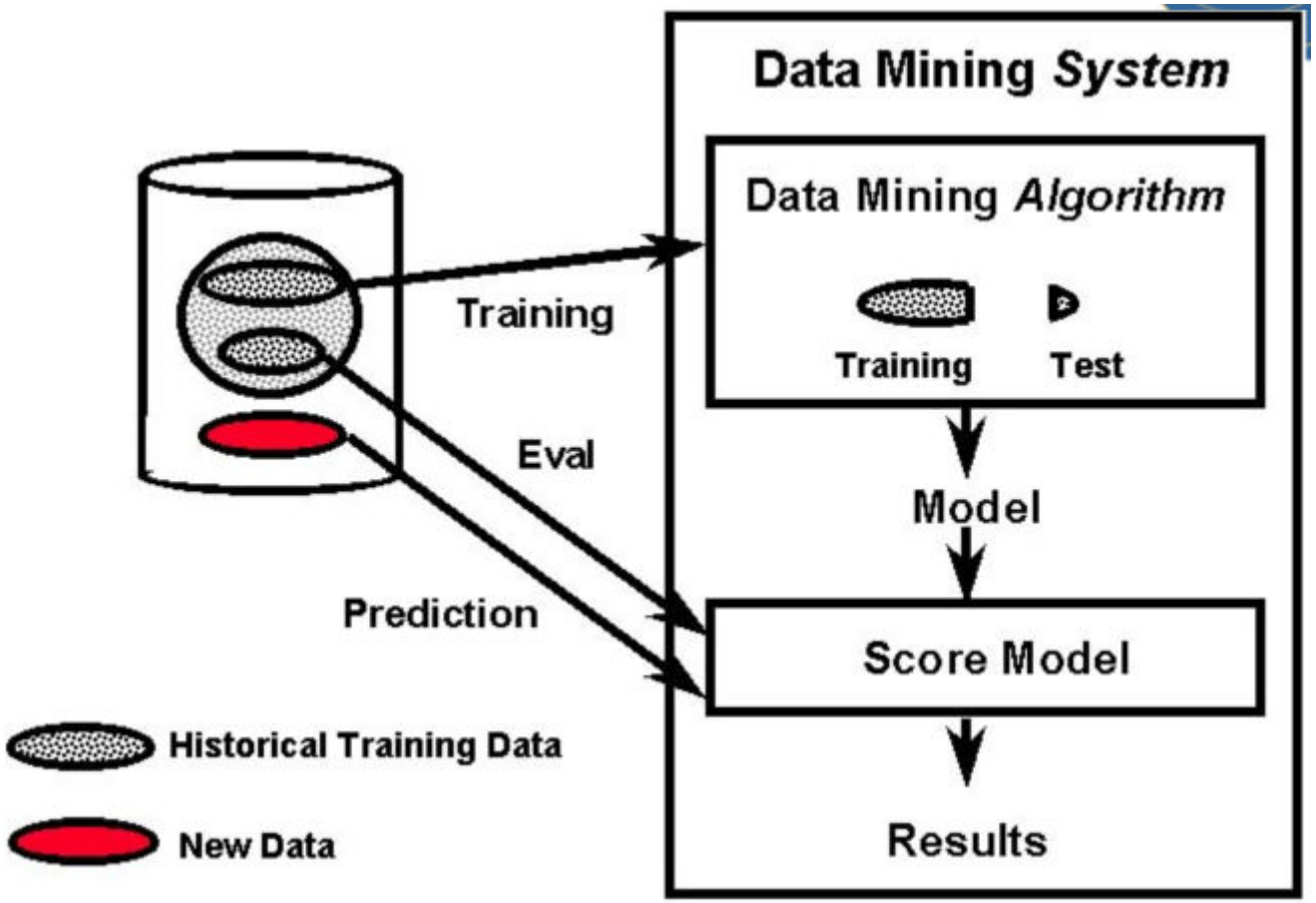
niezrozumiałe

- Problemy
  - Uczenie z nauczycielem i bez nauczyciela
  - Generalizacja i przetrenowanie

Etapy procesy odkrywania wiedzy w bazach danych

- **Selekcja danych** - wybór relacji i krotek, które będą eksplorowane, definicja sposobu łączenia relacji,
- **Transformacja danych** - konwersja typów atrybutów, definicja atrybutów wywiedzionych, dyskretyzacja wartości ciągłych,
- **Eksploracja** - ekstrakcja wiedzy z danych: generowanie reguł, drzew decyzyjnych, sieci neuronowych itp.
- **Interpretacja wyników** - wybór najbardziej interesującej wiedzy, logiczna i graficzna wizualizacja wyników.

Przykładowy proces drążenia danych



Ocena modelu

- Walidacja krzyżowa
  - Podziel dane na N podzbiorów

  - Zbuduj model na bazie N-1 podzbiorów (podzbiór niewykorzystany służyć ma do testowania modelu)

  - Powtórz budowę i testowanie modelu N razy (budując model i testując go dla kolejnych zestawów składających się z N-1 podzbiorów i podzbioru testowego)

Inteligencja

- Inteligencja (łac. intelligentia = pojętność)
  - zdolność rozumienia otaczających sytuacji i znajdowania na nie właściwych, celowych reakcji
    - do zachowań inteligentnych należy postrzeganie, rozpoznawanie, uczenie się, operowanie symbolami, posługiwanie się językiem, rozwiązywanie problemów, twórczość
  - rodzaj szczególnej sprawności umysłowej,
    - inteligencja płynna
    - inteligencja skrystalizowana
  - nazwa warstwy społecznej żyjącej z pracy twórczej umysłu, lub/i będąca głównym nośnikiem kultury narodowej.
  - inteligencja (gra towarzyska)
  - Inteligencja sztuczna (ang. Artificial Intelligence)
    - dział informatyki, którego przedmiotem jest badanie reguł rządzących tzw. inteligentnymi zachowaniami człowieka, tworzenie modeli formalnych tych zachowań i programów komputerowych symulujących te zachowania

Inteligencja

- Typy inteligencji:
  - inteligencja kognitywna (abstrakcyjna)
  - inteligencja werbalna
  - inteligencja emocjonalna
  - inteligencja społeczna
  - inteligencja twórcza
- Składowe ludzkiej inteligencji
  - logika
  - Przeczucia, system wierzeń, irracjonalność i przypadkowość.
  - Sprzężenia zwrotne czyli zmysły.
- Przetwarzanie z góry do dołu
  - Rozwiązując jakiś problem należy zatem najpierw dokonać analizy całości, rozłożyć go na proste elementy i zbudować jego model.
- Przetwarzanie z dołu do góry
  - Zaczyna się od podstaw problemu i stopniowo dochodzi do rozwiązania.
  - Nie jest to tylko techniką rozwiązywania problemów. **Jest to istota funkcjonowania i zachowania żywych organizmów.**

Czy można stworzyć życie na komputerze?

- Życie – posiadanie zdolności do rozmnażania, zróżnicowanie międzyosobnicze, przekazywanie cech następnym generacjom drogą dziedziczenia.
- Żadna struktura, której intuicyjnie nie uznajemy za żywą nie ma tych cech **z wyjątkiem programów komputerowych**. Programy genetyczne ewoluują w czasie, rozmnażają się, dziedziczą zachowanie i podlegają mutacjom.

Sztuczne życie

- Istotą badań nad sztucznym życiem jest prostota.
- Sztuczne życie rządzi się trzema zasadami:
  - Przetwarzanie następuje z dołu do góry
  - Lokalne oddziaływania prowadzą do globalnego (wytwórczego) zachowania
  - Z prostoty powstaje złożoność.

Automat komórkowy Johna von Neumanna

- Czy można stworzyć pojedynczy automat należący do klasy rozwiązującej wszystkie skończone problemy logiczne? – **tak**, maszyna Turinga
- Czy automat może zbudować Taki sam automat?
- Czy automat może ewoluować do bardziej złożonego i stać się bardziej efektywny?
- Von Neumann stworzył pięć modeli samopowielających się automatów przyjmujących jeden z 29 stanów zgodnie z regułami oraz **ze stanem sąsiednich komórek automatu.**
- Idea ta została uproszczona przez Codda – osiem stanów, przestrzeń pięciu komórek – oraz symulacja dwustanowej przestrzeni von Neumanna
- Model ten stał się podstawą gry „Life” Johna Conwaya

Gra „Life”

Reguły przeżycia:

Stan komórki	*Sąsiednie żywe komórki*	*Wynik*
żywa	mniej niż dwie	martwa
żywa	dokładnie dwie	żywa
żywa	dokładnie trzy	żywa
żywa	więcej niż trzy	martwa
martwa	dokładnie trzy	żywa