

Zaawansowana eksploracja danych



Jerzy Stefanowski
Instytut Informatyki
Politechnika Poznańska

Wykład wstępny. dla spec. TPD
Poznań 2008
Zmiany 2009/10

Kilka uwag wstępnych

- Kont. wcześniejszego wykładu „**Eksploracja danych**” (prof.. Morzy)
- Powstaje WWW poświęcona przedmiotowi
<http://www.cs.put.poznan.pl/jstefanowski/tpd.html>
Będzie powiązana z instrukcjami do wykonywania ćwiczeń laboratoryjnych.
- **Oprogramowanie**
 - **WEKA** – źródło + dokumentacja i podręczniki dostępne w WWW.
 - **RapidMiner** (i inne)
 - **Statsoft Dataminer / Statistica 8.0**
 - **MOA** – projekt dla eksploracji strumieni danych
- **Warto spojrzeć na zasoby internetowe**
- **KDnuggets** → bogaty serwis WWW w j. ang. / także wiele materiałów dydaktycznych.

Wymagania wstępne – wcześniejsze przedmioty

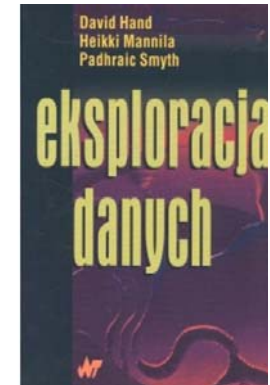
- **Eksploracja danych (TPD)**
 - Reguły asocjacyjne i wzorce sekwencyjne
 - Algorytmy budowy klasyfikatorów
 - Drzewa decyzyjne (C4.5, CART, pruning)
 - Klasyfikacja bayesowska
 - K-NN
 - Algorytmy analizy skupień
 - K-means i AHC
 - Ocena ważności atrybutów
- **Statystyczna analiza danych**
 - Miary opisowe, testy statystyczne, regresja prosta

Literatura anglojęzyczna

- Han Jiawei and Kamber M. Data mining: Concepts and techniques, Morgan Kaufmann, 2001 (1 ed.), there is 2d
- Hand D., Mannila H., Smyth P. Principles of Data Mining, MIT Press, 2001 (tłumaczenie polskie WNT).
- Kononenko I., Kukar M., Machine Learning and Data Mining: Introduction to Principles and Algorithms. Horwood Pub, 2007.
- Maimon O., Rokach L., The data mining and knowledge discovery Handbook, Springer 2005.
- Witten I., Eibe Frank, Data Mining, Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 1999.
- Weiss S., Indurkha N., Predictive data mining, Morgan Kaufmann, 1998.



Polskie ...



- Larose D., Odkrywanie wiedzy z danych. Wprowadzanie do eksploracji danych, PWN, 2006.
- Larose D., Metody i modele eksploracji danych, PWN 2008.
- Hand D., Mannila H., Smyth P. Eksploracja danych, WNT, 2005.

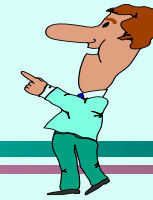
Polskie książki

- **Dobry podręcznik jeszcze nie istnieje ...**
- Koronacki J., Ćwik J., Statystyczne systemy uczące się, WNT 2005 (kolejne wydanie w drodze).
- Krawiec K, Stefanowski J., Uczenie maszynowe i sieci neuronowe, Wyd. PP, 2003.
- Cichosz P., Systemy uczące się. WNT, 2000.
- Lasek M., Data mining: Zastosowanie w ocenach i analizach klientów bankowych. Biblioteka Menadżera, 2003.

Data Mining oraz KDD

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

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!

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)

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 nieznanym i nieoczekiwanym) nietrywialny i zrozumiały.



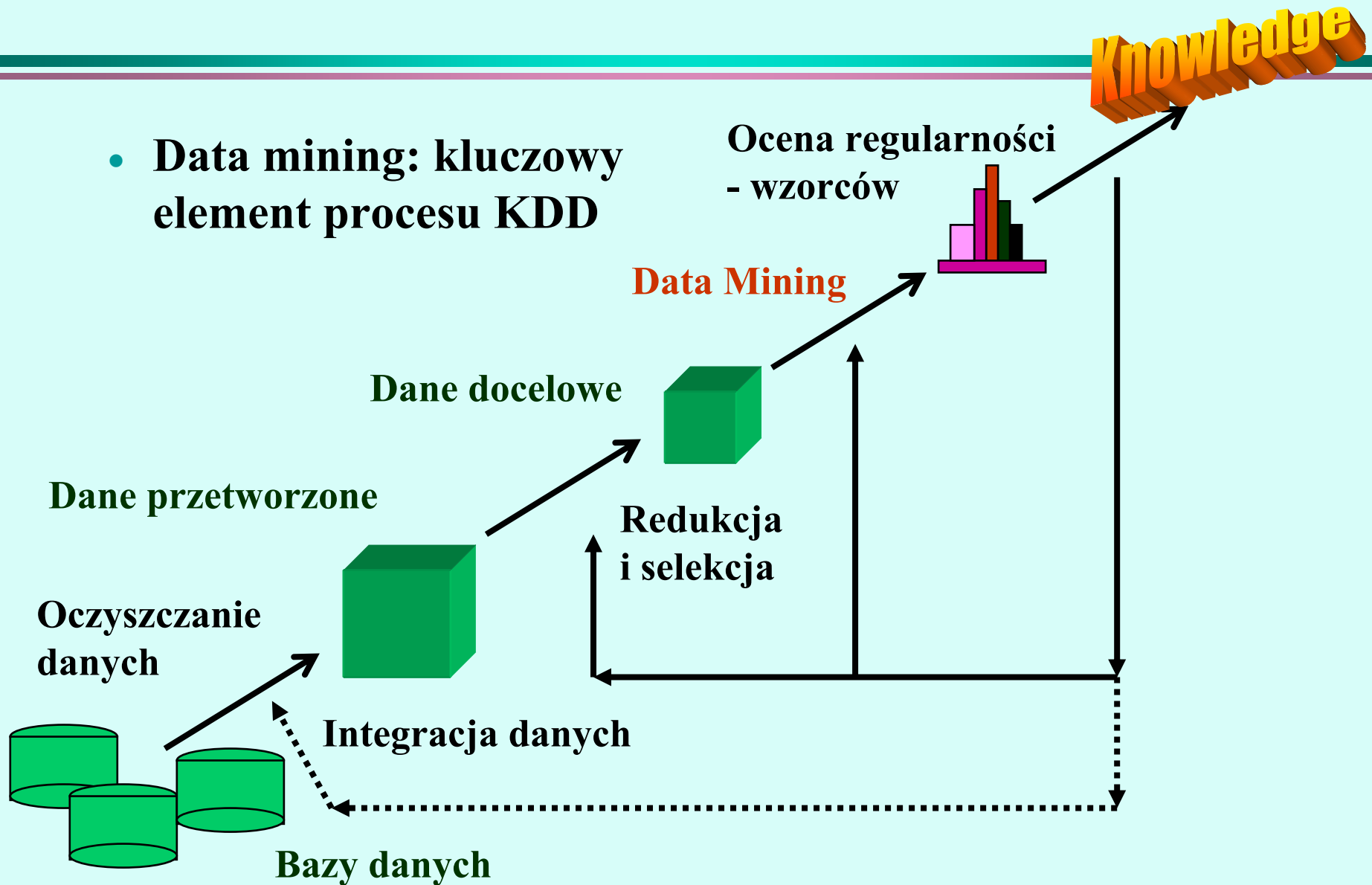
Odkrywanie wiedzy w danych



- 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

Proces Odkrywania Wiedzy - KDD

- **Data mining: kluczowy element procesu KDD**



Etapy procesu odkrywania wiedzy



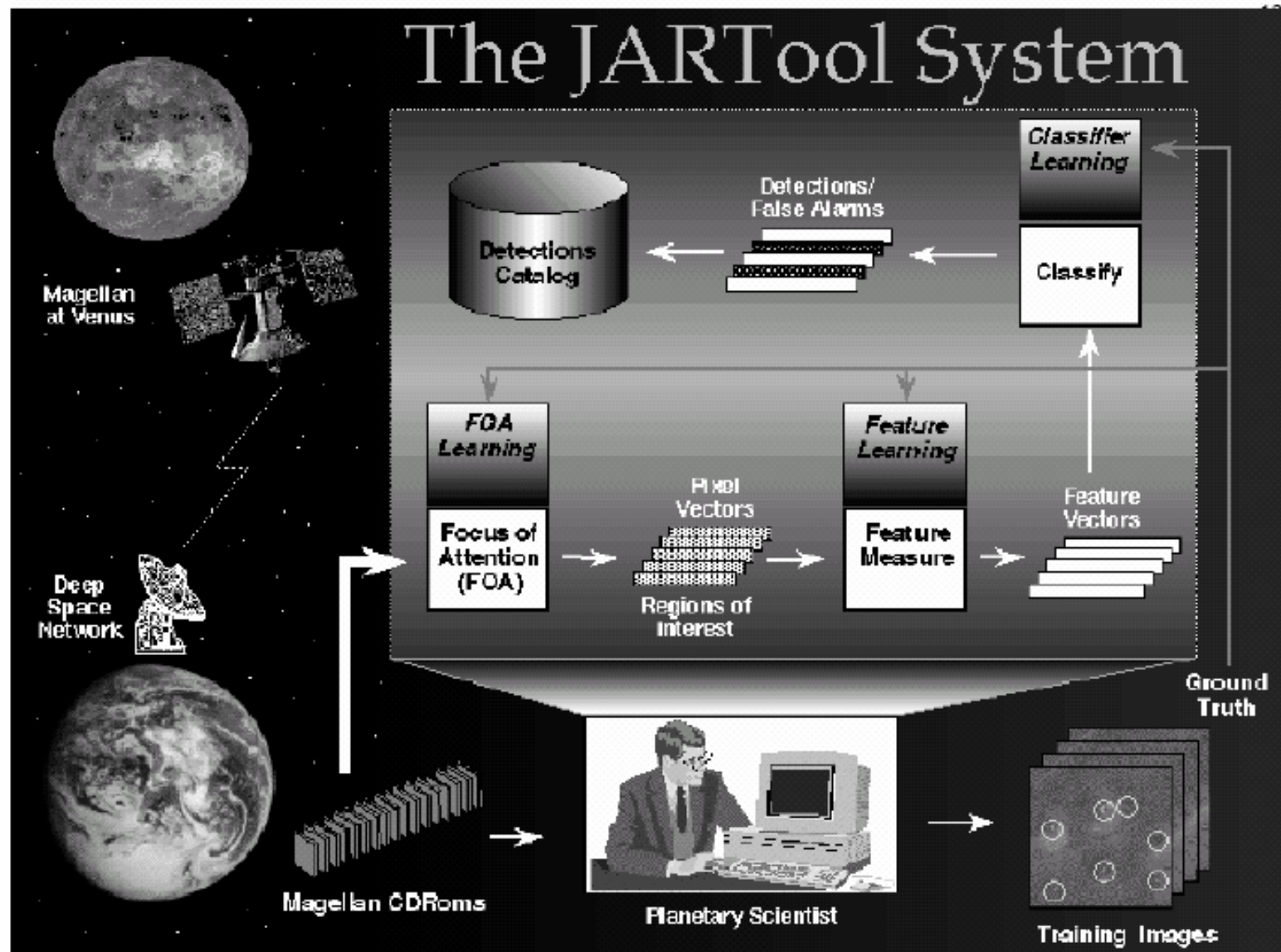
- Analiza i poznanie dziedziny zastosowania, identyfikacja dostępnej wiedzy i celów użytkownika,
- Wybór danych związanych z celami procesu,
- Czyszczenie i wstępne przetwarzanie danych oraz ich redukcja,
- Wybór zadań i algorytmów eksploracji danych,
- Pozyskiwanie wiedzy z danych (krok eksploracji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



SKICAT – przykład procesu KDD

- Sky Image Cataloging and Analysis Tool (SKICAT)
- Developed by NASA's Jet Propulsion Laboratory and the California Institute of Technology in the 90's (Fayyad, Djorgowski, Weir et al.) .
- Aim: a software system to catalog and analyze the estimated half billion sky objects in the second Palomar Observatory sky survey
 - Task 1 – general classes (galaxis, stars, quasars, etc.)
 - Task 2 – find some interesting clusters of objects (quasars with redshift..)
- The survey of the northern sky includes more than 3,000 digitized photographic plates produced by Palomar, located in San Diego.
 - Over 3 terrabytes of images 13000×13000 pixels
- The SKICAT system will produce a comprehensive survey catalog database containing about one-half billion entries by automatically processing about three terabytes (24 trillion bits, 8-bits to a byte) of image data.
- SKICAT is based on state-of-the-art machine learning, high performance database and image processing techniques.
- SKICAT has a correct sky object classification rate of about 94 percent, which exceeds the performance requirement of 90 percent needed for accurate scientific analysis of the data.

Idea systemu



SKICAT – proces KDD

- *Dostępne dane i wiedza początkowa:* kilkadziesiąt tysięcy fotografii o różnych rozdzielczości ręcznie skatalogowanych przez ekspertów.
- *Wybór docelowych danych* – identyfikacja właściwych atrybutów charakteryzujących poszczególne klasy
- konieczność uwzględnienia dodatkowej wiedzy astronomicznej and image analysis FOCAS.
- *Krok czyszczenia* – identyfikacja różnych obserwacji odstających i błędów w danych.
- *Krok redukcji* – wybór tylko części z dostępnych danych.
- *Wybór zadania i algorytmu:* klasyfikowanie – drzewa decyzyjne (modyfikacja C4.5, ale także alternatywne klasyfikatory).
- *Ocena rezultatów* (trafność klasyfikacji – tutaj ponad 94%).
- *Zastosowanie* – wspomaganie tworzenia elektronicznego katalogu gwiazd i galaktyk wraz z ich opisami.

Przetwarzanie wstępne → zmiana reprezentacji

Model wektorowej reprezentacji tekstu

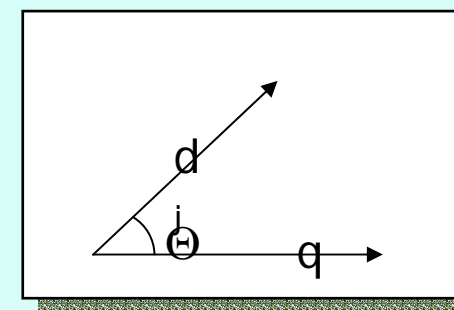
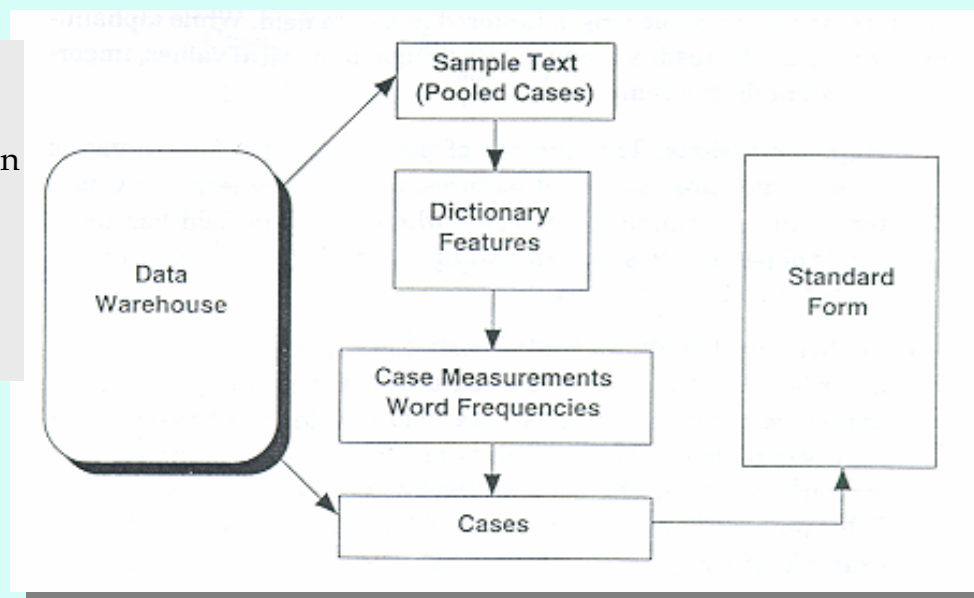
The $d=7$ documents:

- D1: Large Scale Singular Value Computations
- D2: Software for the Sparse Singular Value Decomposition
- D3: Introduction to Modern Information Retrieval
- D4: Linear Algebra for Intelligent Information Retrieval
- D5: Matrix Computations
- D6: Singular Value Analysis of Cryptograms
- D7: Automatic Information Organization

The $t=5$ terms:

- T1: Information
- T2: Singular
- T3: Value
- T4: Computations
- T5: Retrieval

$$A = \begin{pmatrix} 0.00 & 0.00 & 0.71 & 0.71 & 0.00 & 0.00 & 1.00 \\ 0.58 & 0.71 & 0.00 & 0.00 & 0.00 & 0.71 & 0.00 \\ 0.58 & 0.71 & 0.00 & 0.00 & 0.00 & 0.71 & 0.00 \\ 0.58 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.71 & 0.71 & 0.00 & 0.00 & 0.00 \end{pmatrix}$$



Eksploracja danych a inne systemy informatyczne

Informacje / bazy danych	Zapytania SQL i raporty	<ol style="list-style-type: none">1. Który klient dokonał największego zakupu?2. Lista klientów, którzy zakupili produkt A w ostatnim roku?
Hurtownie danych	Wielowymiarowa agregacja danych i podsumowania	<ol style="list-style-type: none">1. 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">1. Jakie są cechy charakterystyczne klientów, którzy mogą kupić produkt A?2. Do kogo skierować ofertę reklamową?

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),...**

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 liczbowa).
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.

Powiązanie zadań i metod

- Za S. Hahn (i S. Weiss, Indurkhia)

Data mining function	Algorithm	Application examples
Association	Statistic, set theory	Market analysis
Classification	Decision trees, neural networks	Target marketing, equality control, risk assessment
Clustering	Neural networks, statistics	Market segmentation, design reuse
Modeling	Linear/nonlinear regression, curve fitting, neural networks	Ranking/scoring customers, pricing models, process control
Times-series forecasting	Statistics ARMA models, Box-Jenkins, neural networks	Sales forecasting, interest rate prediction, inventory control
Sequential patterns	Statistics, set theory	Market basket, analysis over time

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.

Odkrywanie wiedzy a inne dziedziny – cz1.

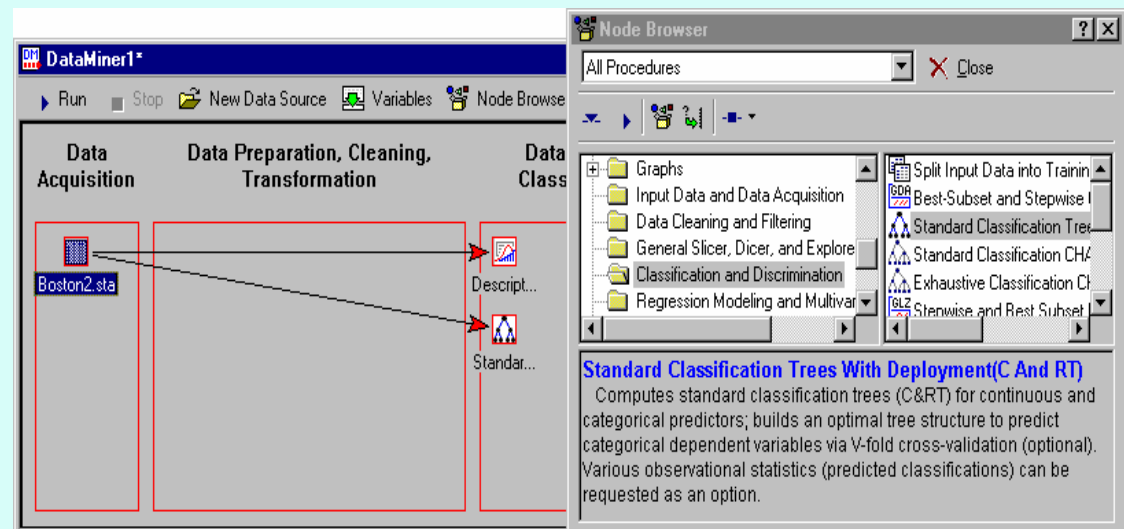
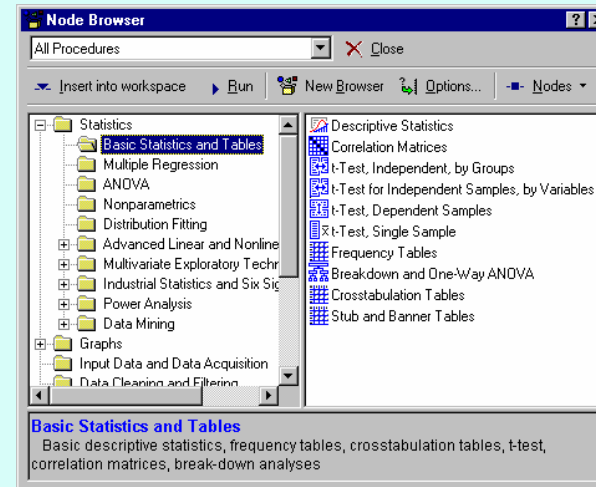
- **Statystyka:**
 - Oparta na silnych podstawach teoretycznych i mocnych założeniach co do danych.
 - Ukierunkowana na testowanie hipotez oraz estymacje parametrów.
- **Uczenie maszynowe:**
 - Ukierunkowane na polepszanie działania przez uczącego się agenta.
 - Uczenie się pojęć lub zadań (lepiej zdefiniowane 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.

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 reprezentacje wiedzy,
 - inny niż poprzednio charakter procesu odkrywania wiedzy.
 - duża rola przygotowania i wstępnego przetwarzanie danych.

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

- Data Miner - My Procedures
- Data Miner - All Procedures
- Data Miner - Data Cleaning and Filtering
- Data Miner - General Slicer/Dicer Explorer with Drill-Down
- Data Miner - General Classifier (Trees and Clusters)
- Data Miner - General Modeler and Multivariate Explorer
- Data Miner - General Forecaster
- Data Miner - General Neural Network Explorer
- Neural Networks
- Generalized EM & k-Means Cluster Analysis
- Association Rules
- General Classification/Regression Tree Models
- General CHAID Models
- Interactive Trees (C&RT, CHAID)
- Boosted Tree Classifiers and Regression
- Generalized Additive Models
- MAR Splines (Multivariate Adaptive Regression Splines)
- Rapid Deployment of Predictive Models (PMML)
- Goodness of Fit, Classification, Prediction
- Feature Selection and Variable Screening



Orange (Slovenia)



[Home](#)
[Screenshots](#)
[Contact & Support](#)
[Acknowledgements](#)

[Download](#)

[Forum](#) (RSS)

[Documentation](#)

[Search](#)

[Visual Programming](#)

[Catalog of Widgets](#)

[Scripting for Beginners](#)

[Class Reference](#)

[Modules](#)

[Example Scripts](#)

[Data Sets](#)

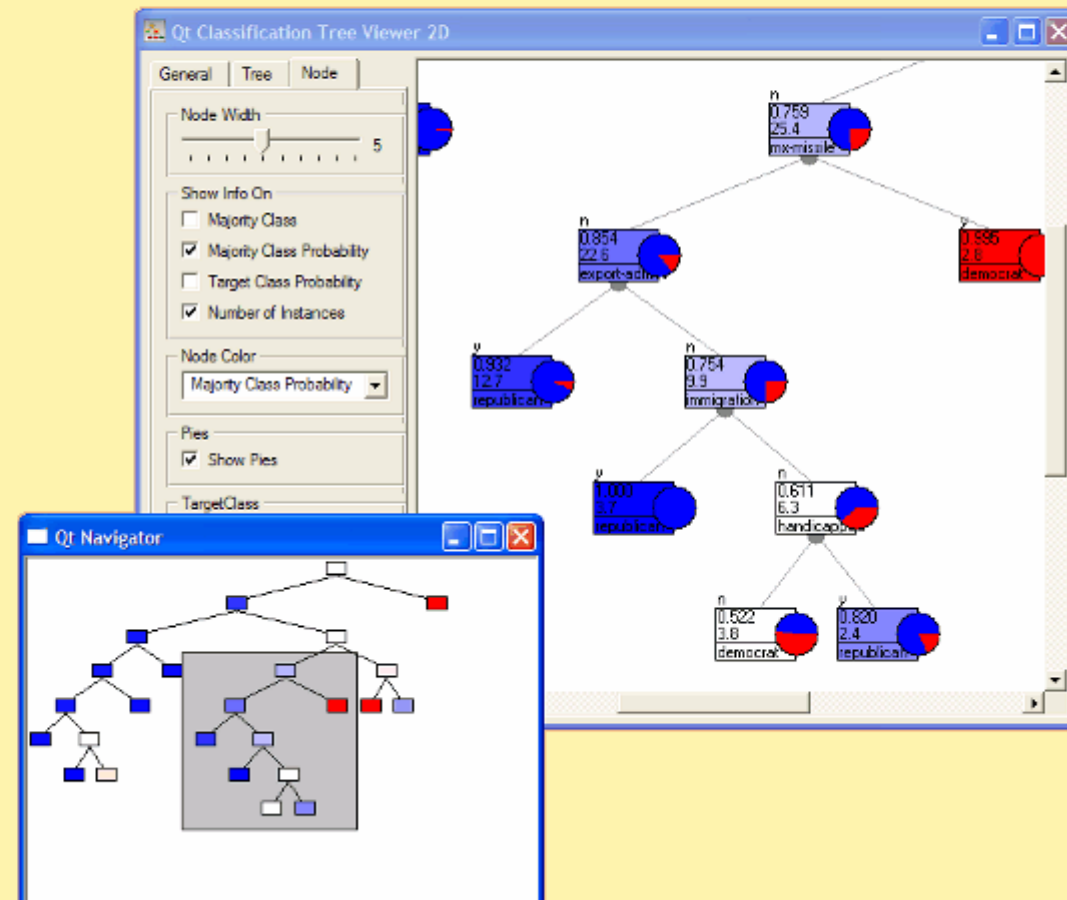
Latest News

Oct 31: The list of [example scripts](#) from documentation works again. For instance, you want to know how to induce random forests in

Orange Screenshots

Following are screenshots of Orange Widgets and Orange's visual programming interface for data mining.

Classification tree viewer with a navigator.



Warto spojrzeć na praktyczne zastosowania

- o Spójrz np. Lavrac et al. Lessons from Data mining ... Machine Learning Journal 2004.
- o Langley, Simon Applications of Machine Learning and Rule Induction
- o Slajdy of Piatetsky Shapiro on KDD

Applications of Machine Learning and Rule Induction

PAT LANGLEY^o

Robotics Laboratory, Computer Science Dept.
Stanford University, Stanford, CA 94305

HERBERT A. SIMON

Department of Psychology
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

An important area of application for machine learning is in automating the acquisition of knowledge bases required for expert systems. In this paper, we review the major paradigms for machine learning, including neural networks, instance-based methods, genetic learning, rule induction, and analytic approaches. We consider rule induction in greater detail and review some of its recent applications, in each case stating the problem, how rule induction was used, and the status of the resulting expert system. In closing, we identify the main stages in fielding an applied learning system and draw some lessons from successful applications.



Machine Learning, 37, 13–34, 2004
© 2004 Kluwer Academic Publishers. Manufactured in The Netherlands.

Introduction: Lessons Learned from Data Mining Applications and Collaborative Problem Solving

NADIA LAVRAC^o nadia.lavrac@ijcl.jussieu.fr
Institut Informatique, Avenue 38, 2000 Leuven, Belgium; Nova Gorica Polytechnic, Vipavski 13, 5000 Nova Gorica, Slovenia

HEROHI MOTODA motoda@rankin.com.tsc.ac.jp
Osaka University, 8-7 Mihayashi, Itanabi, Osaka 565-0847, Japan

TOM FAWCETT tom.fawcett@hp.com
Merwin-Richard Labs, 230 Page Mill Rd., Palo Alto, CA 94304, USA

ROBERT HOLTE holte@cs.ualberta.ca
Computing Science Department, University of Alberta, Edmonton, Alberta Canada T6G 2E8

PAT LANGLEY langley@csli.stanford.edu
Computational Learning Laboratory, Center for the Study of Language & Information, Stanford University, Stanford, CA 94305, USA

PETER ADRIANS pietera@science.ru.nl
Netherlands Institute for Language Logic and Computation, Plantage Muidergracht 24, 1018 TV, Amsterdam, The Netherlands

Abstract. This introductory paper to the special issue on Data Mining Lessons Learned presents lessons from data mining applications, including experience from science, business, and knowledge management in a collaborative data mining setting.

Keywords: data mining, machine learning, scientific discovery, lessons learned, applications, collaborative data mining, knowledge management, future data mining challenges

1. Introduction

This paper reports on experiences gained from a wide variety of applications of machine learning, data mining and scientific discovery. Lessons are drawn from both success and from failures, from the engineering of representations for practical problems, and from expert evaluations of solutions.

In drawing lessons from two different types of data mining and machine learning applications, fielded commercial applications and applications in scientific discovery, we focus on lessons that are new or have been under-emphasized in earlier articles (Langley & Simon, 1995; Brodley & Smyth, 1995; Fayyad, Piatetsky-Shapiro, & Smyth, 1996; Saita & Neri, 1998). However, we also outline the most important lessons reported

www.KDnuggets.com oraz inne serwisy

Address <http://www.kdnuggets.com>



Data Mining, Knowledge Discovery, Genomic Mining, Web Mining

[Data Mining Consulting](#) | [Data Mining Jobs](#) | [Advertising](#) | [Site Map](#)

CLEMENTINE 7.0 = POWER, PREDICTION, PRODUCTIVITY

[SPSS Clementine 7.0 - The next generation of Data Mining](#)

[Free Webinar: Why Use Predictive Analytics?](#)

[KDnuggets News](#), the Data Mining & Knowledge Discovery newsletter: data mining news, jobs, software, courses, ...
[2003 issues](#) | [Schedule](#) | [Archive](#) | [Submit](#) | [Subscribe!](#)

Current Issue: [NEW! 03:19, Oct 14, 2003: Data preparation; NSF deadline; ICDM-2003, Nov 19-22 ... \(29 items\)](#)

Match in: [help](#)

Software:

[Classification](#), [Suites](#), [Text](#)

Jobs:

[Industry](#), [Academic](#)

Solutions:

[Bioinformatics](#), [CRM](#), [Web](#)

Courses:

[Oct](#), [Nov](#), [Dec](#)

Companies:

Meetings

Insightful Miner

Easy to Use & Extensible Data Mining

- Build predictive models easily
- Modern visual interface
- Advanced analytic methods
- Scalable capabilities

Free Webcast & Whitepaper!

[Insightful Miner](#)

Easy to Use & Extensible Data Mining

Poll

How frequently do you do a separate feature selection in classification (rather than have a learning algorithm do selection)

- Always
 Most of the time
 Frequently
 Rarely
 Never

[View Results](#)

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 wspomaganiami decyzji.**