Odkrivanje znanja v podatkih

Marko Bohanec
Institut Jožef Stefan, Ljubljana in
Univerza v Novi Gorici

Kazalo

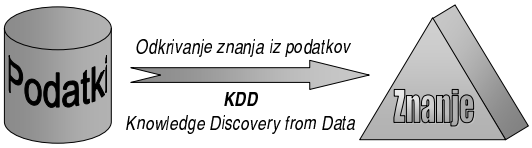
- Uvod
  - Kaj je KDD in kaj DM?
  - Literatura in viri
  - Faze procesa KDD
  - Klasifikacija procesov KDD
  - Področja uporabe
- Faze procesa KDD
- Orodja in praktični primeri

Izhodišča

- Vedno večje zbirke podatkov
  - podatki o poslovanju, transakcijah
  - podatki o poslovnih partnerjih, strankah
  - rezultati anket, analiz, poskusov, ...
- Dosegljivost podatkov
  - javne podatkovne zbirke
  - internet
- Tehnološki napredek
  - informacijska orodja
  - procesna moč računalnikov
  - cene hranjenja in obdelave podatkov
  - metode za analizo podatkov

Opredelitev problema

- Ali se iz podatkov lahko kaj naučimo?
- Ali podatki skrivajo kakšne (doslej neznane) vzorce ali zakonitosti?
- Cilj: Pridobivanje *novega znanja*
  na primer za izboljšanje poslovanja in odločanja v podjetju



Marko Bohanec

Odkrivanje znanja iz podatkov

KDD: *Knowledge Discovery from Data(bases)*

Netrivialen proces odkrivanja implicitnega, doslej neznanega in potencialno uporabnega znanja iz podatkov.

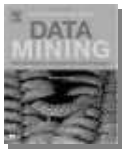
DM: *Data Mining ("Rudarjenje podatkov")*

Faza KDD, v kateri dejansko pride do odkrivanja znanja.

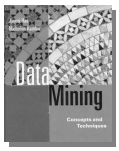
Značilnost: uporaba številnih in raznovrstnih metod.

Marko Bohanec

Viri



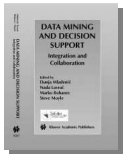
Ian H. Witten, Eibe Frank: Data Mining: *Practical Machine Learning Tools and Techniques, Second Edition*.
Morgan Kaufmann Publishers, 2005.



Jiawei Han, Micheline Kamber:
Data Mining: Concepts and Techniques.
Morgan Kaufmann Publishers, 2001.

Marko Bohanec

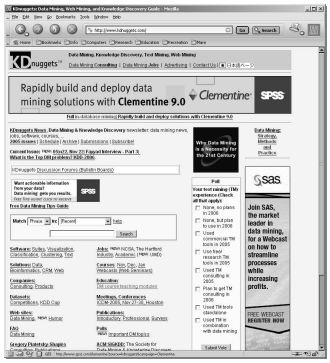
Viri



Dunja Mladenić, Nada Lavrač, Marko Bohanec, Steve Moyle (eds.): *Data Mining and Decision Support: Integration and Collaboration*, Kluwer Academic Publishers, 2003.

Igor Kononenko: *Strojno učenje, druga izdaja*. Založba FE in FRI, 2005.

Viri



http://www.kdnuggets.com/

Odkrivanje znanja iz podatkov

KDD: *Knowledge Discovery from Data(bases)*

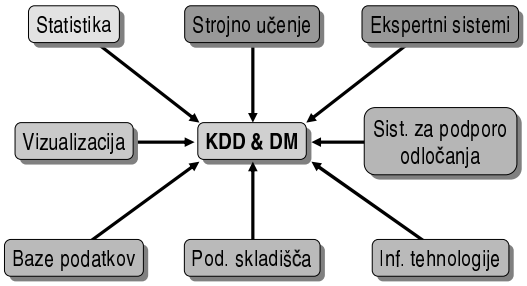
Netrivialen proces odkrivanja implicitnega, doslej neznanega in potencialno uporabnega znanja iz podatkov.

DM: *Data Mining ("Rudarjenje podatkov")*

Faza KDD, v kateri dejansko pride do odkrivanja znanja.

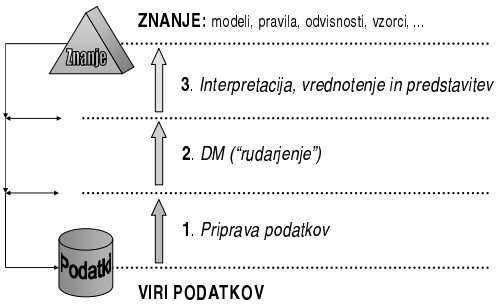
Značilnost: uporaba številnih in raznovrstnih metod.

Interdisciplinarnost KDD in DM



Marko Bohanec

Faze procesa KDD



Marko Bohanec

Uporaba KDD

- Analiza poslovnih partnerjev in strank
- Analiza poslovanja in proizvodnje
- Detekcija in predvidevanje zlorab
- Marketing in odnosi z javnostjo
- Znanstvene raziskave: farmacija, genetika, medicina, ...
- Pri nas:
  - Medicina in biomedicina: diagnostika, prognostika
  - Krmiljenje industrijskih procesov
  - Znanstvene raziskave: genetika, ekologija, ...
  - Obdelave anket in javnomnenjskih raziskav

Marko Bohanec

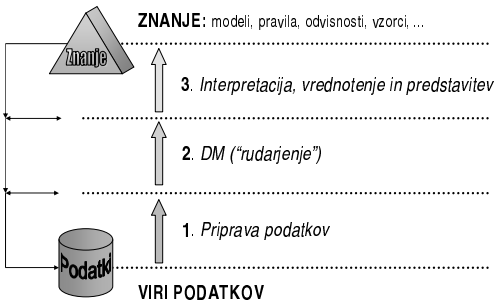
Kazalo

- Uvod
- Faze procesa KDD
  - Priprava podatkov
  - Metode "rudarjenja" in iskanja zakonitosti
  - Interpretacija, vrednotenje in predstavitev rezultatov
- Orodja in praktični primeri

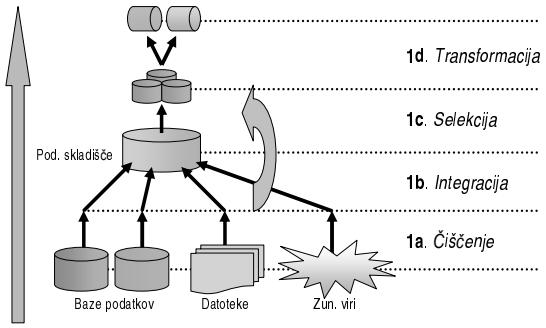
Kazalo

- Uvod
- Faze procesa KDD
  - Priprava podatkov
    - čiščenje, integracija, selekcija, transformacija
    - podatkovno skladišče
  - Metode "rudarjenja" in iskanja zakonitosti
  - Interpretacija, vrednotenje in predstavitev rezultatov
- Orodja in praktični primeri

Faze procesa KDD



1. faza KDD: Priprava podatkov



Marko Bohanec

1. faza KDD: Priprava podatkov

Klient	Priimek	Naslov	Dat.nak.	Artikel
23003	Kranjc	Ptujska 1	1.3.2000	strip
23003	Kranjc	Ptujska 1	1.3.2000	CD
23003	Kranjc	Ptujska 1	6.4.2000	bonboni
23009	Novak	Dunajska 3	1.1.0001	revija
23013	Kralj	Tržaška 4	30.2.2000	revija
23019	Kranjec	Ptujska 1	7.6.2002	CD

Marko Bohanec

1. faza KDD: Priprava podatkov

ČIŠČENJE

Klient	Priimek	Naslov	Dat.nak.	Artikel
23003	Kranjc	Ptujska 1	1.3.2000	strip
23003	Kranjc	Ptujska 1	1.3.2000	CD
23003	Kranjc	Ptujska 1	6.4.2000	bonboni
23009	Novak	Dunajska 3	1.1.0001	revija
23013	Kralj	Tržaška 4	30.2.2000	revija
23019	Kranjec	Ptujska 1	7.6.2002	CD

Marko Bohanec

1. faza KDD: Priprava podatkov

PO ČIŠČENJU

Klient	Priimek	Naslov	Dat.nak.	Artikel
23003	Kranjc	Ptujska 1	1.3.2000	strip
23003	Kranjc	Ptujska 1	1.3.2000	CD
23003	Kranjc	Ptujska 1	6.4.2000	bonboni
23009	Novak	Dunajska 3	NULL	revija
23013	Kralj	Tržaška 4	30.2.2000	revija
23003	Kranjc	Ptujska 1	7.6.2000	CD

Marko Bohanec

1. faza KDD: Priprava podatkov

INTEGRACIJA

Klient	Priimek	Naslov	Dat.nak.	Artikel
23003	Kranjc	Ptujska 1	1.3.2000	strip
23003	Kranjc	Ptujska 1	1.3.2000	CD
23003	Kranjc	Ptujska 1	6.4.2000	bonboni
23009	Novak	Dunajska 3	NULL	revija
23013	Kralj	Tržaška 4	30.2.2000	revija
23003	Kranjc	Ptujska 1	7.6.2000	CD

Priimek	Dat.roj.	Promet	Avto
Kranjc	7.7.1965	102.413,00	ne
Novak	6.6.1971	13.666,00	da

Klient	Priimek	Naslov	Dat.roj.	Dat.nak.	Artikel	Promet	Avto
23003	Kranjc	Ptujska 1	7.7.1965	1.3.2000	strip	102.413,00	ne
23003	Kranjc	Ptujska 1	7.7.1965	1.3.2000	CD	102.413,00	ne
23003	Kranjc	Ptujska 1	7.7.1965	6.4.2000	bonboni	102.413,00	ne
23009	Novak	Dunajska 3	6.6.1971	NULL	revija	13.666,00	da
23013	Kralj	Tržaška 4	NULL	30.2.2000	revija	NULL	NULL
23003	Kranjc	Ptujska 1	7.7.1965	7.6.2000	CD	102.413,00	ne

Marko Bohanec

1. faza KDD: Priprava podatkov

SELEKCIJA

Klient	Priimek	Naslov	Dat.roj.	Dat.nak.	Artikel	Promet	Avto
23003	Kranjc	Ptujska 1	7.7.1965	1.3.2000	strip	102.413,00	ne
23003	Kranjc	Ptujska 1	7.7.1965	1.3.2000	CD	102.413,00	ne
23003	Kranjc	Ptujska 1	7.7.1965	6.4.2000	bonboni	102.413,00	ne
23009	Novak	Dunajska 3	6.6.1971	NULL	revija	13.666,00	da
23013	Kralj	Tržaška 4	NULL	30.2.2000	revija	NULL	NULL
23003	Kranjc	Ptujska 1	7.7.1965	7.6.2000	CD	102.413,00	ne

Marko Bohanec

1. faza KDD: Priprava podatkov

PO SELEKCIJI

Klient	Naslov	Dat.roj.	Artikel	Promet	Avto
23003	Ptujska 1	7.7.1965	strip	102.413,00	ne
23003	Ptujska 1	7.7.1965	CD	102.413,00	ne
23003	Ptujska 1	7.7.1965	bonboni	102.413,00	ne
23009	Dunajska 3	6.6.1971	revija	13.666,00	da
23003	Ptujska 1	7.7.1965	CD	102.413,00	ne

Marko Bohanec

1. faza KDD: Priprava podatkov

TRANSFORMACIJA

Klient	Regija	Dat.roj.	Artikel	Promet	Avto	Starost
23003	02	7.7.1965	strip	visok	0	35
23003	02	7.7.1965	CD	visok	0	35
23003	02	7.7.1965	bonboni	visok	0	35
23009	01	6.6.1971	revija	nizek	1	29
23003	02	7.7.1965	CD	visok	0	35

Marko Bohanec

1. faza KDD: Priprava podatkov

TRANSFORMACIJA

Klient	Regija	Dat.roj.	Starost	Promet	Avto	strip	CD	bonb.	revija
23003	02	7.7.1965	35	visok	0	1	2	1	0
23009	01	6.6.1971	29	nizek	1	0	0	0	1

Marko Bohanec

Podatkovno skladišče

Podatkovno skladišče (*Data Warehouse*)
je zbirka podatkov,
namenjena podpori odločanja (pri upravljanju podjetij).

Lastnosti:
- vključuje podatke iz različnih virov
- namenjeno podrobni analizi velike količine podatkov
- urejeno po:
  - predmetu obravnave (kupec, proizvod, prodaja, dobavitelj)
  - času
- relativno statično (občasna ažuriranja, sicer poizvedovanje)

Marko Bohanec

Baza : Skladišče

BAZA PODATKOV
- podpira delo s podatki
- vnos in branje podatkov
- dinamično spreminjanje vsebine
- struktura se redko spreminja
- veliko uporabnikov
- transakcijske obdelave
- vnaprej določeni izpisi ali poizvedovanja s SQL

SKLADIŠČE PODATKOV
- podpira analizo podatkov
- branje podatkov
- podatki so statični, le občasno ažuriranje
- strukturo prilagajamo potrebam
- malo uporabnikov
- analitične in sintetične obdelave
- ad-hoc analize, korelacije, statistike, OLAP

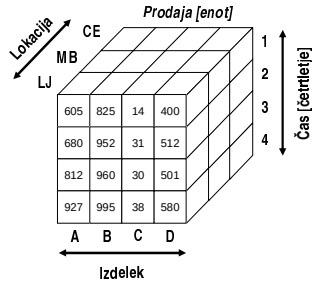
Marko Bohanec

Organizacija podatkovnih skladišč

Osnovni koncepti:
- podatkovna kocka (*Data Cube*)
- hierarhično urejene dimenzije (*Concept Hierarchy*)
- analize tipa OLAP (*On Line Analytical Processing*)

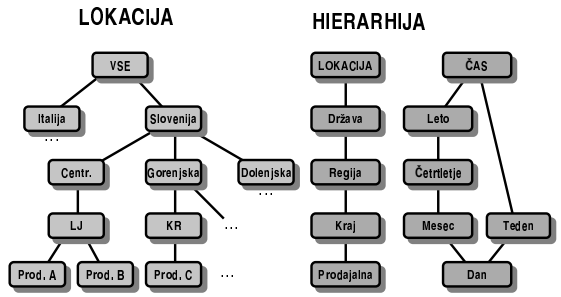
Marko Bohanec

Podatkovna kocka (*Data Cube*)



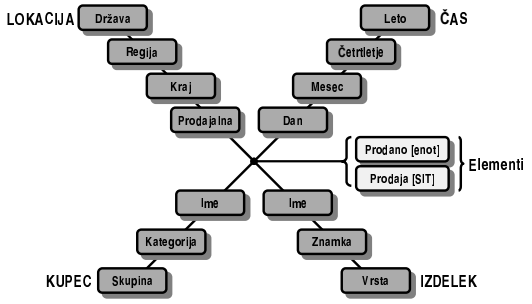
Marko Bohanec

Hierarhija dimenzij (*Concept Hierarchy*)



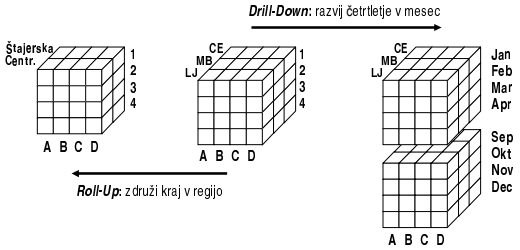
Marko Bohanec

Dimenzije in elementi skladišča



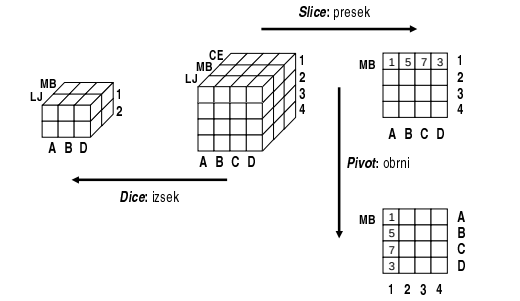
Marko Bohanec

Operacije OLAP



Marko Bohanec

Operacije OLAP



Marko Bohanec

Kazalo

- Uvod
- Faze procesa KDD
  - Priprava podatkov
  - Metode "rudarjenja" in iskanja zakonitosti
    - statistične metode
    - vizualizacija
    - metode strojnega učenja
    - asociacijska (povezovalna) pravila
    - razvrščanje v skupine
  - Interpretacija, vrednotenje in predstavitev rezultatov
- Orodja in praktični primeri

Marko Bohanec

Statistične metode

Priprava podatkov:
- odkrivanje in glajenje "šuma", napak v podatkih

Začetne faze KDD
- osnovne statistike: srednja vrednost, standardni odklon
- vizualizacija: histogrami, razpršitveni diagrami

Osrednje faze KDD
- korelacije
- regresijske metode
- diskriminantna analiza
- analiza osnovnih komponent (*PCA*)

Zaključne faze KDD
- dokazovanje hipotez

Marko Bohanec

Primer: Osnovne statistike

Klient	Regija	Dat.roj.	Starost	Promet	Avto	strip	CD	bonb.	revija
23003	02	7.7.1965	35	visok	0	1	2	1	0
23009	01	6.6.1971	29	nizek	1	0	0	0	1
23011	01	5.5.1931	69	nizek	0	1	0	3	1
23013	01	3.3.1980	20	visok	1	0	0	0	2
23015	02	1.1.1981	19	nizek	0	1	0	0	0
23020	01	8.8.1966	34	nizek	0	0	0	4	0

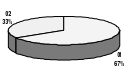
Vsota				2	3	2	8	4
Povpr.	$\bar{x} = \frac{\sum x}{n}$	34,33		0,33	0,50	0,33	1,33	0,67
Std.od.	$\sigma_x = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$	18,28		0,52	0,55	0,82	1,75	0,82

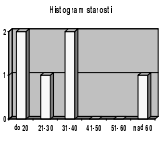
Marko Bohanec

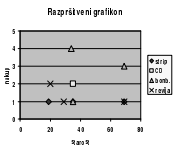
Primer: Osnovne porazdelitve, grafikoni

Klient	Regija	Dat.roj.	Starost	Promet	Avto	strip	CD	bonb.	revija
23003	02	7.7.1965	35	visok	0	1	2	1	0
23009	01	6.6.1971	29	nizek	1	0	0	0	1
23011	01	5.5.1931	69	nizek	0	1	0	3	1
23013	01	3.3.1980	20	visok	1	0	0	0	2
23015	02	1.1.1981	19	nizek	0	1	0	0	0
23020	01	8.8.1966	34	nizek	0	0	0	4	0

Regija	Klientov
01	4
02	2







Marko Bohanec

Vizualizacija

"Slika nadomesti tisoč besed"!

Uporaba:
- osnovno pregledovanje podatkov
- prikaz rezultatov KDD

Osnovna grafična orodja:
- "pogače", histogrami, razpršitveni diagrami, radarski grafikoni
- preprosta
- lahko dostopna (MS Excel)

Zahtevnejše tehnike:
- številne, zanimive, atraktivne, hiter razvoj novih
- specializacija
- zahtevnejša uporaba (usposabljanje)
- težje dostopna orodja

Marko Bohanec

Strojno učenje

Glejte: Posebno predavanje o strojnem učenju

V povezavi s KDD se uporabljajo predvsem:
- metode učenja odločitvenih (klasifikacijskih in regresijskih) dreves
  - izvedenke algoritmov: C4.5, M5, CART
  - preprostejša izvedba
- umetne nevronske mreže

Marko Bohanec

Kazalo

Marko Bohanec

Asociacijska (povezovalna) pravila

Tipični problem: analiza nakupovalnih košaric

Košarica	Artikel
1	mleko
1	maslo
2	mleko
2	med
2	maslo
3	mleko
3	kruh
3	maslo
4	mleko
4	kruh
4	med

Naloga: Poiskati "zanimiva" pravila oblike

če kupi mleko, **potem** kupi tudi maslo

`mleko ⇒ maslo`

Meri "zanimivosti":

- podpora (*support*)
- zaupanje (*confidence*)

Marko Bohanec

Asociacijska (povezovalna) pravila

Košarica	Artikel
1	mleko
1	maslo
2	mleko
2	med
2	maslo
3	mleko
3	kruh
3	maslo
4	mleko
4	kruh
4	med

Podpora: $\text{support}\ (A \Rightarrow B) = P(A \cup B)$

Zaupanje: $\text{confidence}\ (A \Rightarrow B) = P(B \mid A)$

Podpora	Nekatere podmnožice artiklov
4/4=100%	{mleko}
3/4=75%	{maslo}, {mleko, maslo}
2/4=50%	{med}, {kruh}, {med, mleko}, {kruh, mleko}
1/4=25%	{med, kruh}, {med, maslo}, {kruh, maslo}

Tri pravila:

`mleko ⇒ maslo` `[sup 75%, conf 75%]`

`maslo ⇒ mleko` `[sup 75%, conf 100%]`

`med ⇒ mleko` `[sup 50%, conf 100%]`

Marko Bohanec

Razvrščanje v skupine (*Clustering*)

Klasifikacija:

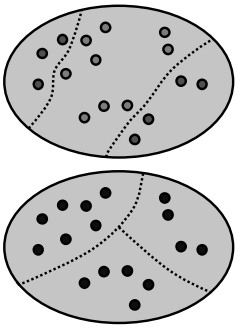
- razvrščanje primerov, katerih razred *je* znan

Razvrščanje v skupine:

- razred *ni* znan
- razvrščanje po "podobnosti"

Definirati je potrebno:

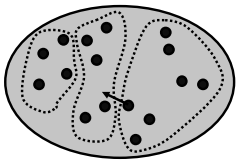
- mero razdalje med primeri
- pri nekaterih metodah tudi število skupin k



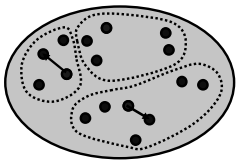
Marko Bohanec

Razvrščanje: Nehierarhične metode

Metoda prestavljanj:
- določi *k* začetnih skupin
- dokler lahko izboljšuješ razvrstitve:
  - izberi element
  - premakni ga v sosednjo skupino.



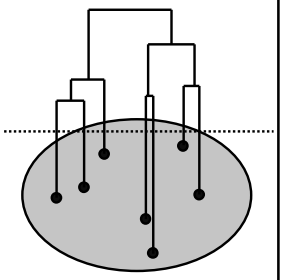
Metoda voditeljev:
- določi *k* voditeljev
- dokler se voditelji ne ustalijo:
  - primere priredi najbližjim voditeljem
  - izračunaj središča skupin in
  - jih določi za nove voditelje.



Marko Bohanec

Razvrščanje: Hierarhične metode

Hierarhično združevanje:
- postopno združevanje najbližjih elementov oziroma skupin
- rezultat je drevo združevanja – *dendrogram*



Marko Bohanec

Kazalo

- Uvod
- Faze procesa KDD
  - Priprava podatkov
  - Metode "rudarjenja" in iskanja zakonitosti
  - Interpretacija, vrednotenje in predstavitev rezultatov
- Orodja in praktični primeri

Marko Bohanec

Interpretacija, vrednotenje, predstavitev

- Rezultate praviloma vrednotimo sproti
- Mere kvalitete:
  - klasifikacijska točnost
  - podpora, zaupanje
  - razumljivost, velikost modelov
  - "zanimivost", novost, uporabnost
- Metode:
  - uporaba različnih metod na istih podatkih
  - uporaba različnih (pod)množic podatkov
  - prečno preverjanje (*cross validation*)
  - statistično dokazovanje hipotez
  - vizualizacija
  - ekspertno mnenje
  - "zdrava pamet"

Marko Bohanec

Kazalo

- Uvod
- Faze procesa KDD
  - Priprava podatkov
  - Metode "rudarjenja" in iskanja zakonitosti
  - Interpretacija, vrednotenje in predstavitev rezultatov
- Orodja in praktični primeri

Marko Bohanec

Sistemi in orodja za KDD

- Izjemno hiter razvoj
- Na internetu je dostopnih precej primerjalnih študij
- Delne rešitve so poceni ali brezplačne, dobre pa razmeroma drage
- Nekaj znanih sistemov:
  - IBM DB2 Intelligent Miner, IBM
  - XpertRuleMiner, Attar Software Ltd.
  - MineSet, Silicon Graphics Inc.
  - Clementine, SPSS Inc.
  - SAS Enterprise Miner, SAS Institute Inc.
  - Weka, University of Waikato
  - SQL Server 2008, Analysis Services, Microsoft
  - Orange (Fakulteta za računalništvo in informatiko)

Marko Bohanec

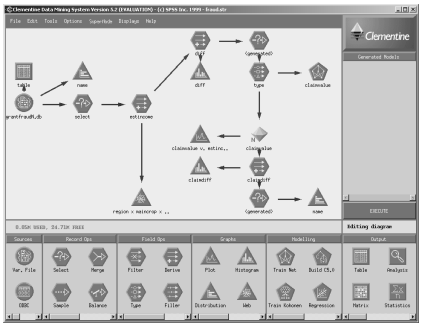
Clementine (SPSS Inc.)

Funkcije
- Preprosto grafično urejanje in prikazovanje procesa KDD
- Doseganje do podatkov in podatkovnih zbirk
- Transformacije podatkov
- Vizualizacija podatkov
- Tekstovni prikazi modelov in delovanja le-teh
- Izvoz modelov: programska koda v jeziku C

Metode
- Nevronske mreže
- Kohonenove mreže
- Učenje odločitvenih dreves in pravil (C5.0)
- Asociacijska pravila (GRI, Apriori)
- Regresija
- Razvrščanje v skupine (Kmeans)

Marko Bohanec

Clementine (SPSS Inc.)



Marko Bohanec

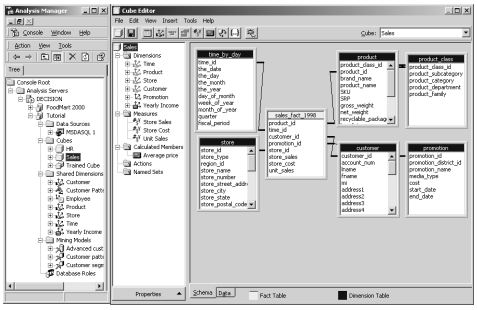
Microsoft SQL Server 2003

Analysis Services / Analysis Manager

Funkcije:
- Izdelava, vzdrževanje in upravljanje podatkovnih skladišč
- Definicija podatkovnih kock
- Analize tipa OLAP
- Odločitvena drevesa
- Razvrščanje v skupine

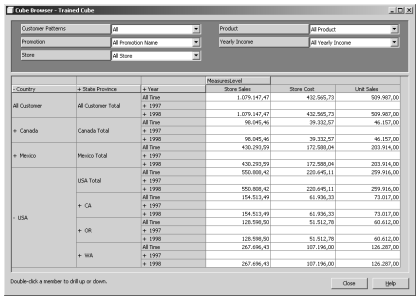
Marko Bohanec

MS Analysis Services: Kocke podatkov



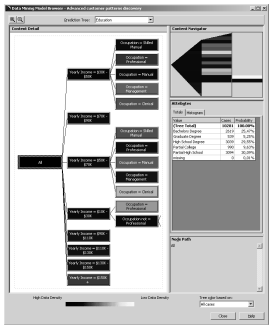
Marko Bohanec

MS Analysis Services: OLAP



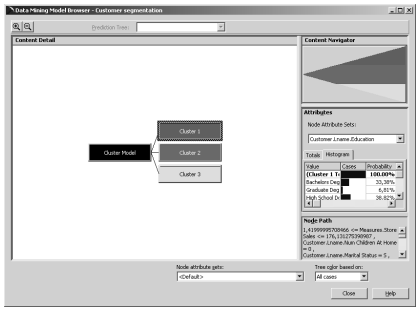
Marko Bohanec

MS Analysis Services: Odločitveno drevo



Marko Bohanec

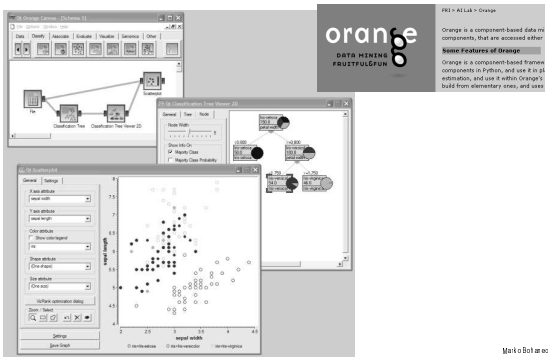
MS Analysis Services: Skupine



Marko Bohanec

Orange

http://magix.fri.uni-lj.si/orange/



Marko Bohanec

KDD: Povzetek

- Cilj: izkoristiti podatke kot vir znanja za boljše odločanje in delovanje
- Interdisciplinarnost:
  - baze podatkov, skladišča podatkov, sistemi za podporo odločanja
  - umetna inteligenca: ekspertni sistemi, strojno učenje, nevronske mreže
  - matematika, statistika, operacijske raziskave, optimizacija
- Faze KDD:
  1. priprava podatkov: integracija, čiščenje, selekcija, transformacija
  2. "rudarjenje" (*Data Mining*): uporaba številnih in raznovrstnih metod
  3. interpretacija, vrednotenje, predstavitev
- Najpomembnejše metode rudarjenja:
  - statistične: osnovne, korelacije, diskriminantne in regresijske analize
  - strojno učenje: odločitvena drevesa, pravila, nevronske mreže, genetski alg.
  - razvrščanje v skupine
  - asociacijska pravila
  - vizualizacija
- Kvaliteta rezultatov:
  - objektivne mere: točnost, zaupanje, podpora
  - razumljivost
  - novost in uporabnost

Marko Bohanec