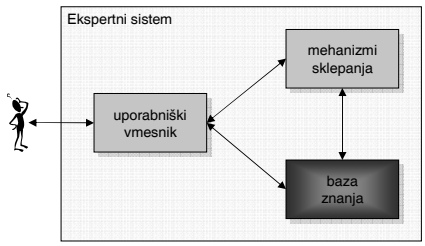
Strojno učenje in iskanje zakonitosti v podatkih

Marko Bohanec
Institut Jožef Stefan, Ljubljana
http://www-ai.ijs.si/MarkoBohanec/mare.html

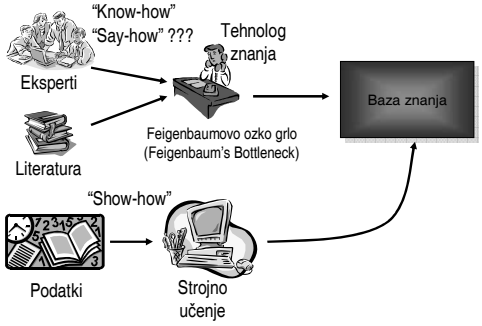
Gradivo:
http://www-ai.ijs.si/MarkoBohanec/PES/ML-KDD-3.pdf
http://www-ai.ijs.si/MarkoBohanec/PES/ML-KDD-6.pdf

Kazalo

Arhitektura ES



Strojno učenje



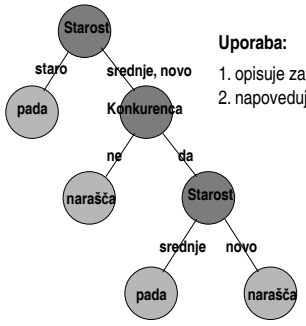
povzeto po: Bojan Cestnik: Strojno učenje

Primer

Dobiček	Starost	Konkurenca	Vrsta
pada	staro	ne	SW
pada	srednje	da	SW
narašča	srednje	ne	HW
pada	staro	ne	HW
narašča	novo	ne	HW
narašča	novo	ne	SW
narašča	srednje	ne	SW
narašča	novo	da	SW
pada	srednje	da	HW
pada	staro	da	SW

povzeto po: Bojan Cestnik: Strojno učenje

Primer: Odločitveno drevo



Uporaba:

1. opisuje zakonitosti v obstoječih podatkih
2. napoveduje (klasificira) nove primere

povzeto po: Bojan Cestnik: Strojno učenje

Metode strojnega učenja

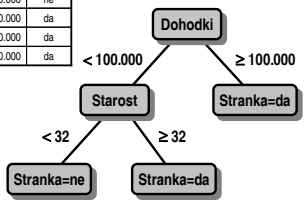
- Statistične metode
  - Bayesov klasifikator
  - *k*-najbližjih sosedov (*k-Nearest Neighbors, k-NN*)
  - diskriminantna analiza
- Simbolično induktivno učenje
  - odločitvena drevesa (*Decision Trees*)
  - odločitvena pravila (*Decision Rules*)
  - učenje konceptov (*Concept Learning*)
  - indukcija logičnih programov (*ILP: Inductive Logic Programming*)
- Umetne nevronske mreže

Zahteve pri strojnem učenju

- Zanesljivost delovanja
  - velika klasifikacijska točnost
- Transparentnost naučenega znanja
  - eksplicitna simbolična predstavitev, razumljiva ekspertom
- Sposobnost pojasnjevanja
  - argumentiranje in podpora ekspertnim odločitvam
- Odpornost na "šum" v podatkih
  - delovanje ob manjkajočih, nepopolnih ali nenatančnih podatkih
  - problemi iz realnega sveta

Učenje odločitvenih dreves

Oseba	Starost	Spol	Dohodki	Stranka
Ana Kranjc	32	Ž	10.000	da
Micka Kovač	53	Ž	1.000.000	da
Meta Novak	27	Ž	20.000	ne
Jana Bevc	55	Ž	20.000	da
Peter Dolenc	26	M	100.000	da
Janez Gorenc	50	M	200.000	da



Učenje odločitvenih dreves

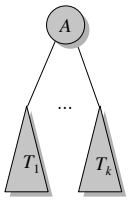
KLJUČNI KONCEPTI

1. Gradnja drevesa
   - algoritem
   - izbiranje atributov
2. Preverjanje kakovosti drevesa
   - učna in testna množica
   - klasifikacijska točnost
3. Rezanje drevesa
   - rezanje naprej
   - rezanje nazaj

Gradnja klasifikacijskega drevesa

ALGORITEM

- Če vsi učni primeri pripadajo istemu razredu C, potem je rezultat list C
- Sicer
  - Izberi *najboljši* atribut A (ali *najboljšo* delitev po A)
  - Razdeli učno množico glede na vrednosti A
  - Rekurzivno zgradi poddrevesa $T_1..T_k$ za vsako podmnožico
  - Rezultat je drevo z vozliščem A in poddrevesi $T_1..T_k$



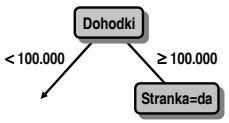
povzeto po: Bojan Cestnik: Strojno učenje

Primer

Oseba	Starost	Spol	Dohodki	Stranka
Ana Kranjc	32	Ž	10.000	da
Micka Kovač	53	Ž	1.000.000	da
Meta Novak	27	Ž	20.000	ne
Jana Bevc	55	Ž	20.000	da
Peter Dolenc	26	M	100.000	da
Janez Gorenc	50	M	200.000	da

Vsi primeri v istem razredu?

***Najboljša* delitev?**



Oseba	Starost	Spol	Dohodki	Stranka
Ana Kranjc	32	Ž	10.000	da
Meta Novak	27	Ž	20.000	ne
Jana Bevc	55	Ž	20.000	da

Oseba	Starost	Spol	Dohodki	Stranka
Micka Kovač	53	Ž	1.000.000	da
Peter Dolenc	26	M	100.000	da
Janez Gorenc	50	M	200.000	da

Izbiranje atributov (delitev)

ZAHTEVA: Delitev na čimbolj "čiste" podmnožice

MERE "NEČISTOČE"

za dva razreda, $p(C_1)=p_1$, $p(C_2)=p_2$

- Entropija E: $-p_1\log_2 p_1 - p_2\log_2 p_2$
- Napaka prevladujočega razreda: $1-\max(p_1,p_2)$
- Indeks Gini: $1-(p_1^2+p_2^2)$

INFORMACIJSKI PRISPEVEK: Koliko pridobimo ob delitvi?

- $\text{Gain}(S,A) = E(S) - \sum_v |S_v|/|S|\ E(S_v)$
- maksimiziramo Gain

povzeto po: Bojan Cestnik: Strojno učenje

Primer (1 od 3)

Vreme	Temp	Vlaga	Veter	Tenis
sončno	vroče	visoka	ne	ne
sončno	vroče	visoka	da	ne
oblačno	vroče	visoka	ne	da
dež	zmerno	visoka	ne	da
dež	hladno	norm	ne	da
dež	hladno	norm	da	ne
oblačno	hladno	norm	da	da
sončno	zmerno	visoka	ne	ne
sončno	hladno	norm	ne	da
dež	zmerno	norm	ne	da
sončno	zmerno	norm	da	da
oblačno	zmerno	visoka	da	da
oblačno	vroče	norm	ne	da
dež	zmerno	visoka	da	ne

Vreme? sončno [2+, 3–] E=0,97
oblačno [4+, 0–] E=0
dež [3+, 2–] E=0,97

Vlaga? visoka [3+, 4–] E=0,99
norm [6+, 1–] E=0,59

Veter? ne [6+, 2–] E=0,81
da [3+, 3–] E=1,00

Primer (2 od 3)

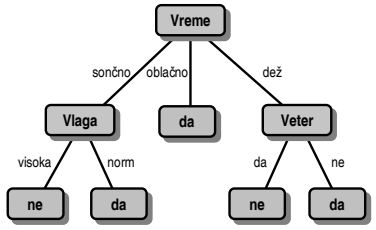
Veter? ne [6+, 2–] E=0,81
da [3+, 3–] E=1,00

$\text{Gain}(S,A) = E(S) - \sum_v |S_v|/|S|\ E(S_v)$

$E(S) = E(9+, 5–) = -(9/14)\log_2(9/14) - (5/14)\log_2(5/14) = 0{,}94$
$E(S_{\text{Veter=ne}}) = 0{,}81$
$E(S_{\text{Veter=da}}) = 1{,}00$
$\text{Gain}(S,\text{Veter}) = 0{,}94 - (8/14)0{,}81 - (6/14)1{,}00 =$ **0,048**

$\text{Gain}(S,\text{Vreme}) =$ **0,246 (max)**
$\text{Gain}(S,\text{Vlaga}) =$ **0,151**
$\text{Gain}(S,\text{Temp}) =$ **0,029**

Primer (3 od 3)



Mere kvalitete odločitvenih dreves

Klasifikacijska točnost:
Kako točno drevo klasificira nove primere?
Kakšna je točnost v primerjavi z *apriorno* ("*naivni klasifikator*")?

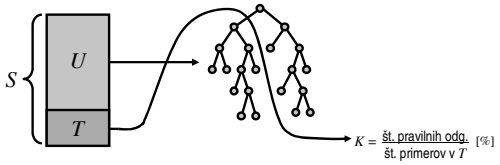
Razumljivost:
Ali ekspert razume drevo in njegovo vsebino?
Ali ga lahko interpretira, utemelji?

Velikost:
Povezano z razumljivostjo: zaželena *čim manjša* drevesa!

Klasifikacijska točnost

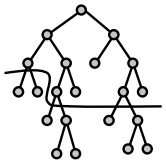
Postopek gradnje in preverjanja klasifikatorja:

- množico primerov S razdelimo na:
  - ***učno*** množico U (npr. 70%) in
  - ***testno*** množico T (30%)
- zgradimo drevo upoštevajoč samo U
- na T preverimo ***točnost klasifikacije*** = delež pravilnih klasifikacij



Rezanje dreves

Spodnji deli drevesa (okrog listov) so manj zanesljivi:
- manjše število učnih primerov
- prevelika prilagoditev podatkom (*overfitting*)



Rezanje (*pruning*):
- **Naprej**: predčasno ustavimo gradnjo
- **Nazaj**: drevo zgradimo do konca, nato režemo manj zanesljive dele (bolje!)

Pridobimo:
- Manjše drevo – večja preglednost in razumljivost
- Večja točnost na testni množici primerov

Nekaj učnih algoritmov in programov

UČENJE ODLOČITVENIH DREVES

- ID3 (Quinlan 1979)
- CART (Breiman et al. 1984)
- Assistant (Cestnik et al. 1987)
- C4.5 (Quinlan 1993)

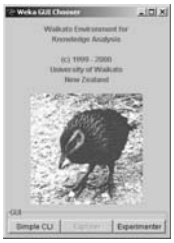
- C5.0, See5 (RuleQuest)
- Weka (Waikato University, NZ)

See5 (RuleQuest)

http://www.rulequest.com/

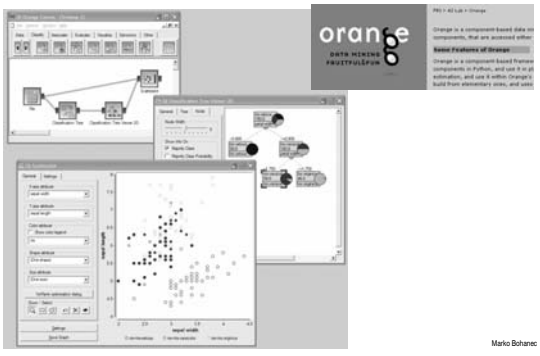
Weka

http://www.cs.waikato.ac.nz/~ml/weka/



Orange

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



Kazalo

Opredelitev problema

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

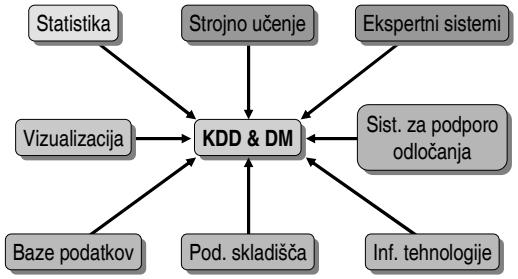
DM: *Data Mining*
(*"rudarjenje podatkov", "izkopavanje podatkov", "izkopavanje znanja"*)

Faza KDD, v kateri dejansko pride do odkrivanja znanja.

Značilnost: uporaba številnih in raznovrstnih metod.

Marko Bohanec

Interdisciplinarnost KDD in DM



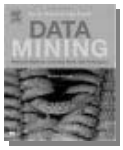
Marko Bohanec

Uporaba KDD

- Detekcija in predvidevanje zlorab
- Analiza poslovnih partnerjev in strank
- Analiza poslovanja in proizvodnje
- 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

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

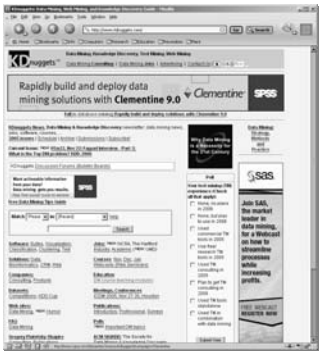
Viri

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.

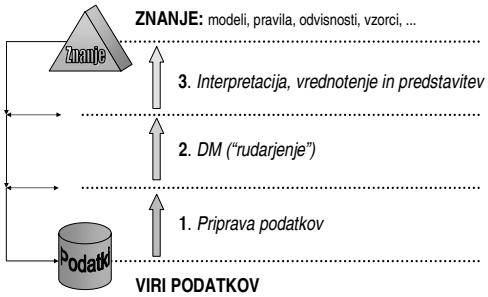
Marko Bohanec

Viri



http://www.kdnuggets.com/

Faze procesa KDD



Kazalo

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

Primer: Osnovne statistike

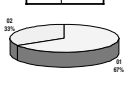
Klient	Regija	Dat.roj.	Starost	Promet	Avto	strip	CD	bonb.	revija
23003	02	7.7.1965	35	visok	0	1	2	1	0
23009	01	6.6.1971	29	nizek	1	0	0	0	1
23011	01	5.5.1931	69	nizek	0	1	0	3	1
23013	01	3.3.1980	20	visok	1	0	0	0	2
23015	02	1.1.1981	19	nizek	0	1	0	0	0
23020	01	8.8.1966	34	nizek	0	0	0	4	0

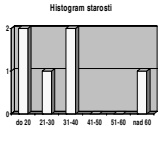
Vsota				2	3	2	8	4
Povpr.	$\bar{X} = \frac{\sum X}{n}$	34,33		0,33	0,50	0,33	1,33	0,67
Std.od.	$\sigma_X = \sqrt{\frac{\sum (X - \bar{X})^2}{n-1}}$	18.28		0,52	0,55	0,82	1,75	0,82

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





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*)

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%]`

Razvrščanje v skupine (*Clustering*)

Klasifikacija:

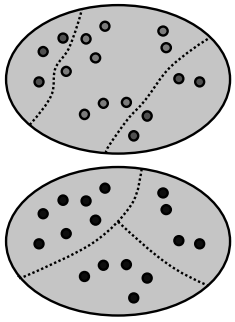
- razvrščanje primerov, katerih razred *je* znan

Razvrščanje v skupine:

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

Definirati je potrebno:

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

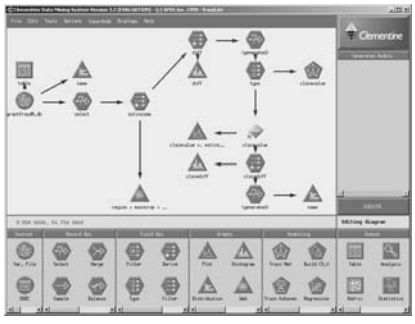


Sistemi in orodja za KDD

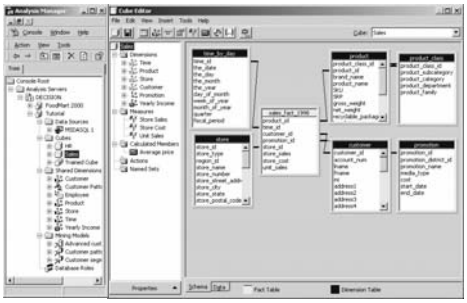
- Izjemno hiter razvoj
- Na internetu je dostopnih precej primerjalnih študij
- Delne rešitve so poceni ali brezplačne, dobre pa razmeroma drage

- 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 2003, Analysis Services, Microsoft
  - Orange (Fakulteta za računalništvo in informatiko)

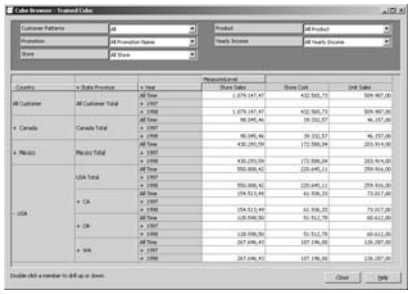
Clementine (SPSS Inc.)



MS Analysis Services: Kocke podatkov

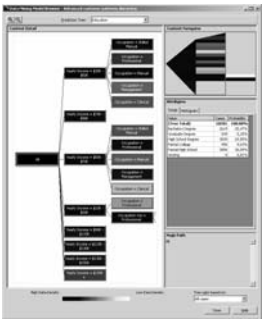


MS Analysis Services: OLAP



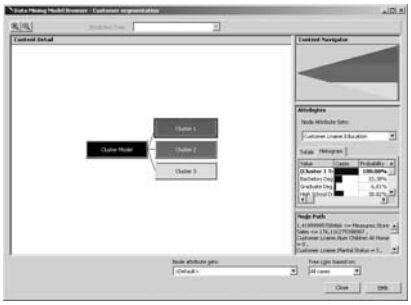
Marko Bohanec

MS Analysis Services: Odločitveno drevo



Marko Bohanec

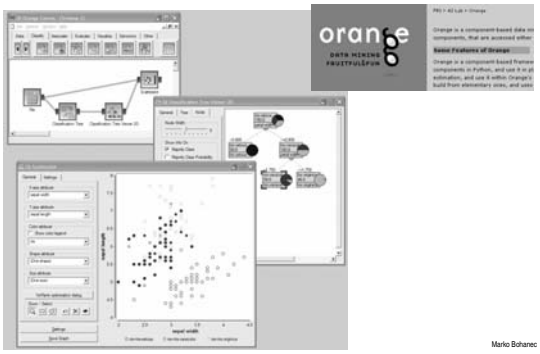
MS Analysis Services: Skupine



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