

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>

Marko Bohanec

Kazalo

Strojno učenje

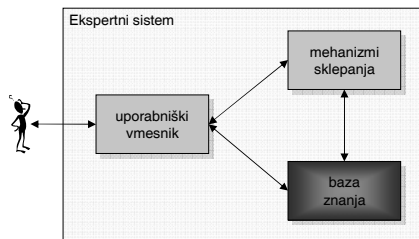
- motivacija in definicija
- metode strojnega učenja
- metoda učenja odločitvenih dreves
- orodja in praktični primeri

Iskanje zakonitosti v podatkih

- faze procesa KDD
- metode KDD
 - statistične metode
 - vizualizacija
 - (strojno učenje)
 - asociacijska (povezovalna) pravila
 - razvrščanje v skupine
- orodja in praktični primeri

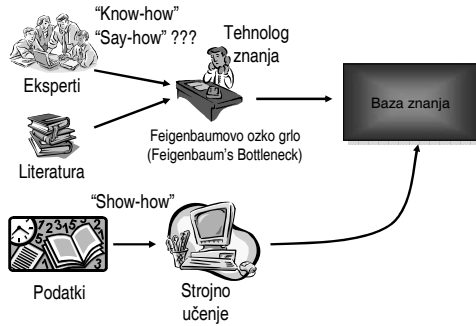
Marko Bohanec

Arhitektura ES



Marko Bohanec

Strojno učenje



prevzeto po: Bojan Cestnik: Strojno učenje

Marko Boharac

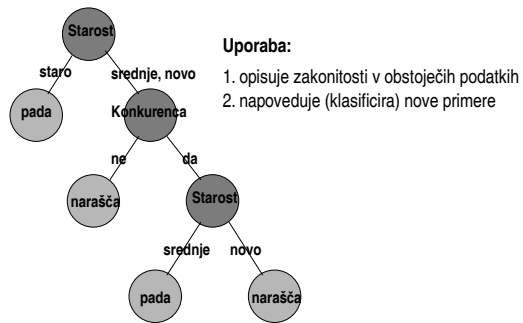
Primer

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

prevzeto po: Bojan Cestnik: Strojno učenje

Marko Boharac

Primer: Odločitveno drevo



prevzeto po: Bojan Cestnik: Strojno učenje

Marko Boharac

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

Marko Bohanec

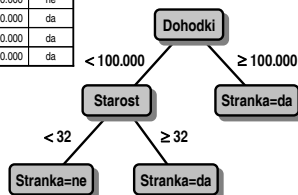
Zahteve pri strojnem učenju

- 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

Marko Bohanec

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



Marko Bohanec

Učenje odločitvenih dreves

KLJUČNI KONCEPTI

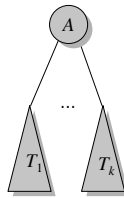
1. Gradnja drevesa
 - algoritem
 - izbiranje atributov
2. Preverjanje kakovosti drevesa
 - učna in testna množica
 - klasifikacijska točnost
3. Rezanje drevesa
 - rezanje naprej
 - rezanje nazaj

Marko Bohanec

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 $T_1..T_k$



prevzeto po: Bojan Cestrik: Strojno učenje

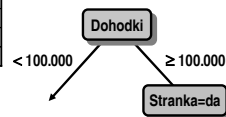
Marko Bohanec

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

Marko Bohanec

Izbiranje atributov (delitev)

ZAHTEVA: Delitev na čimbolj "čiste" podmnožice

HERE "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

prevzeto po: Bojan Cestnik: Šolsko učanje

Marko Boharac

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-] $E=1,00$

Marko Boharac

Primer

(2 od 3)

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

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

$E(S) = E(9+, 5-) = -(9/14) \log_2(9/14) - (5/14) \log_2(5/14) = 0,94$

$E(S_{\text{veter=ne}}) = 0,81$

$E(S_{\text{veter=da}}) = 1,00$

$\text{Gain}(S, \text{Veter}) = 0,94 - (8/14)0,81 - (6/14)1,00 = \mathbf{0,048}$

$\text{Gain}(S, \text{Vreme}) = \mathbf{0,246}$ (max)

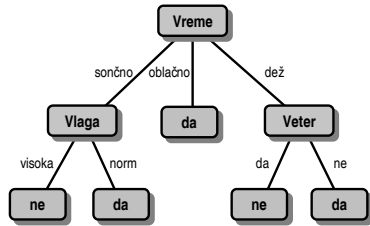
$\text{Gain}(S, \text{Vlaga}) = \mathbf{0,151}$

$\text{Gain}(S, \text{Temp}) = \mathbf{0,029}$

Marko Boharac

Primer

(3 od 3)



Marko Bohanec

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:

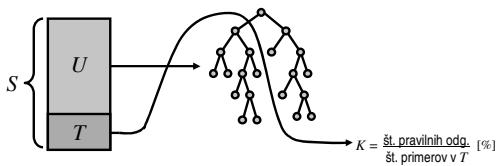
Povezano z razumljivostjo: zaželena čim manjša drevesa!

Marko Bohanec

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



Marko Bohanec

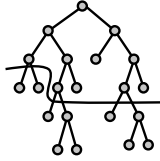
Rezanje dreves

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

Marko Bohanec

Nekaj učnih algoritmov in programov

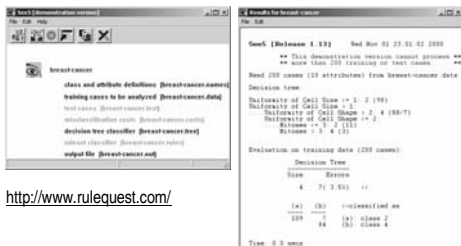
UČENJE ODLOČITVENIH DREVES

- 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)

Marko Bohanec

See5 (RuleQuest)

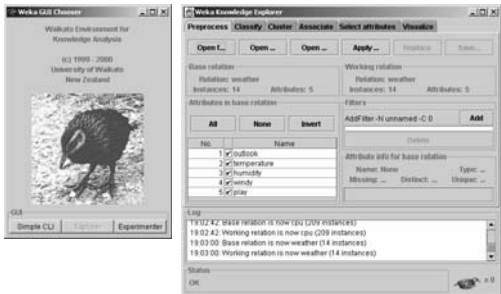


<http://www.rulequest.com/>

Marko Bohanec

Weka

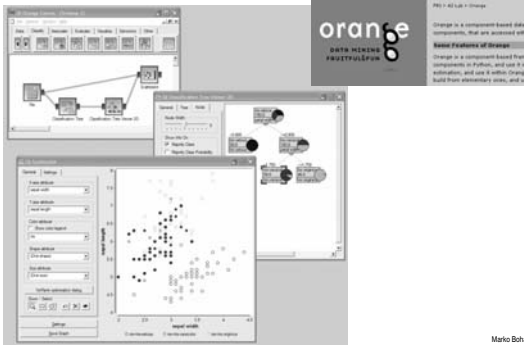
<http://www.cs.waikato.ac.nz/~ml/weka/>



Marko Boharac

Orange

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



Marko Boharac

Kazalo

Strojno učenje

- motivacija in definicija
- metode strojnega učenja
- metoda učenja odločitvenih dreves
- orodja in praktični primeri

Iskanje zakonitosti v podatkih

- faze procesa KDD
- metode KDD
 - statistične metode
 - vizualizacija
 - (strojno učenje)
 - asociacijska (povezovalna) pravila
 - razvrščanje v skupine
- orodja in praktični primeri

Marko Boharac

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

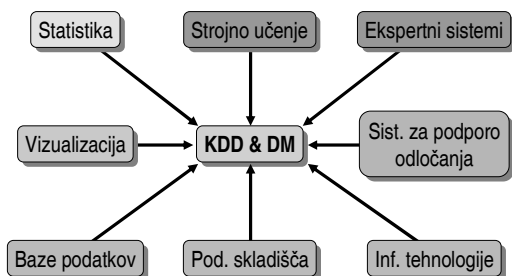
("rudarjenje podatkov", "izkopavanje podatkov", "izkopavanje znanja")

Faza KDD, v kateri dejansko pride do odkrivanja znanja.

Značilnost: uporaba številnih in raznovrstnih metod.

Marko Bohanec

Interdisciplinarnost KDD in DM



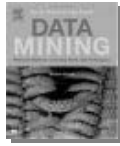
Marko Bohanec

Uporaba KDD

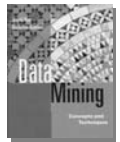
- Detekcija in predvidevanje zlorab
- Analiza poslovnih partnerjev in strank
- Analiza poslovanja in proizvodnje
- Marketing in odnosi z javnostjo
- Znanstvene raziskave: farmacija, genetika, medicina, ...
- Pri nas:
 - Medicina in biomedicina: diagnostika, prognostika
 - Krmiljenje industrijskih procesov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

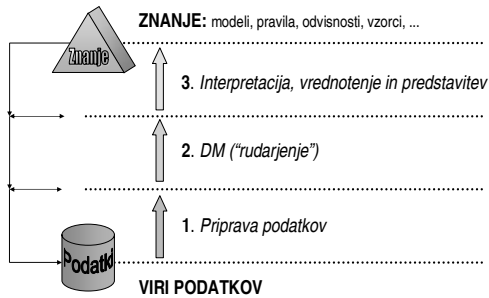
Viri



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

Marko Boharac

Faze procesa KDD



Marko Boharac

Kazalo

Strojno učenje

- motivacija in definicija
- metode strojnega učenja
- metoda učenja odločitvenih dreves
- orodja in praktični primeri

Iskanje zakonitosti v podatkih

- faze procesa KDD
- metode KDD
 - statistične metode
 - vizualizacija
 - (strojno učenje)
 - asociacijska (povezovalna) pravila
 - razvrščanje v skupine
- orodja in praktični primeri

Marko Boharac

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 Boharac

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

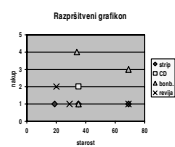
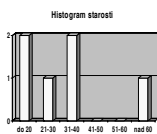
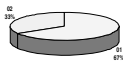
Vsota					2	3	2	8	4
Povpr.	$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$		34,33		0,33	0,50	0,33	1,33	0,67
Std.od.	$\sigma_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$		18,28		0,52	0,55	0,82	1,75	0,82

Marko Boharac

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	Klijentov
01	4
02	2



Marko Boharac

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
 $mleko \Rightarrow maslo$

Meri "zanimivosti":

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

Marko Bohanec

Asociacijska (povezovalna) pravila

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

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

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	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:

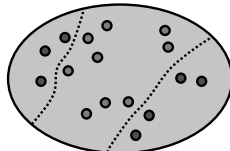
$mleko \Rightarrow maslo$ [sup 75%, conf 75%]
 $maslo \Rightarrow mleko$ [sup 75%, conf 100%]
 $med \Rightarrow mleko$ [sup 50%, conf 100%]

Marko Bohanec

Razvrščanje v skupine (*Clustering*)

Klasifikacija:

- razvrščanje primerov, katerih razred je znan

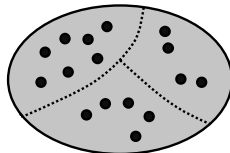


Razvrščanje v skupine:

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

Definirati je potrebno:

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



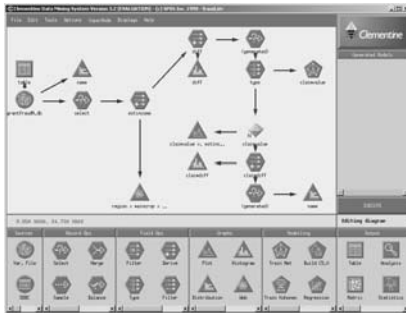
Marko Bohanec

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)

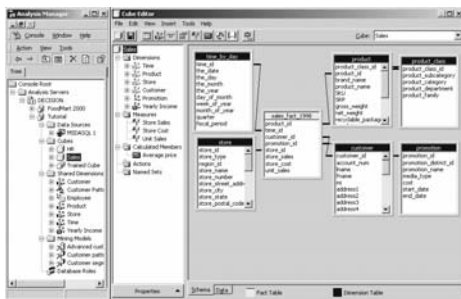
Marko Boharac

Clementine (SPSS Inc.)



Marko Boharac

MS Analysis Services: Kocke podatkov



Marko Boharac

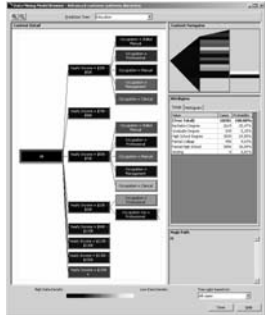
MS Analysis Services: OLAP

Customer Filter: All Product: All Product
 Member: All Member Name Months: 12 Months
 Slice: All Slice

			PivotTable1		
			State Sales	State Totals	Local Sales
Country	State Province	All Totals			
All Customer	All Customer Total	All Totals	3,878,347,417	452,785,711	684,407,391
		All Totals	3,878,347,417	452,785,711	684,407,391
Canada	Canada Total	All Totals	98,548,461	36,332,837	46,457,330
		All Totals	98,548,461	36,332,837	46,457,330
Mexico	Mexico Total	All Totals	430,000,000	173,688,000	253,914,000
		All Totals	430,000,000	173,688,000	253,914,000
USA	USA Total	All Totals	3,350,800,000	383,765,000	584,036,000
		All Totals	3,350,800,000	383,765,000	584,036,000
CA	CA	All Totals	98,548,461	36,332,837	46,457,330
		All Totals	98,548,461	36,332,837	46,457,330
		All Totals	98,548,461	36,332,837	46,457,330
		All Totals	98,548,461	36,332,837	46,457,330
US	US	All Totals	3,350,800,000	383,765,000	584,036,000
		All Totals	3,350,800,000	383,765,000	584,036,000
		All Totals	3,350,800,000	383,765,000	584,036,000
		All Totals	3,350,800,000	383,765,000	584,036,000
WA	WA	All Totals	247,188,411	107,198,000	138,397,101
		All Totals	247,188,411	107,198,000	138,397,101
		All Totals	247,188,411	107,198,000	138,397,101
		All Totals	247,188,411	107,198,000	138,397,101

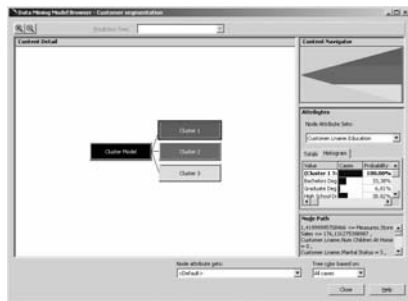
Marko Bohanc

MS Analysis Services: Odločitveno drevo

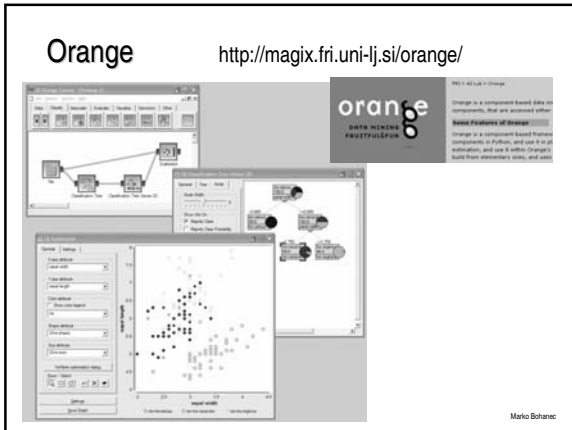


Marko Bohanc

MS Analysis Services: Skupine



Marko Bohanc



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
