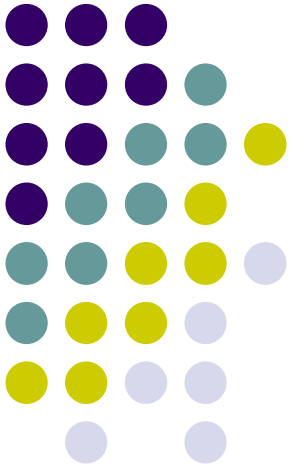
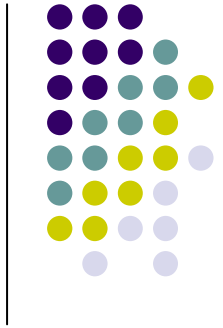
Rozhodovací stromy

semestrální práce z předmětu
Měření a zpracování dat v dopravě



Obsah prezentace



- Teoretická část
  - Data mining
  - Popis rozhodovacích stromů (DT)
  - Tvorba DT
  - Převod stromu na pravidla
  - Prořezávání stromů
  - Reálné aplikace

- Praktická část – klasifikace stupňů dopravy
  - Software na tvorbu DT
  - Předzpracování dat
  - Zpracování softwarem
  - Výstupy práce
  - Shrnutí výsledků

Úvod: Data mining

2.polovina 20.století - „exploze“ objemu dat

Rozvoj analytických metod, tzv. data miningu

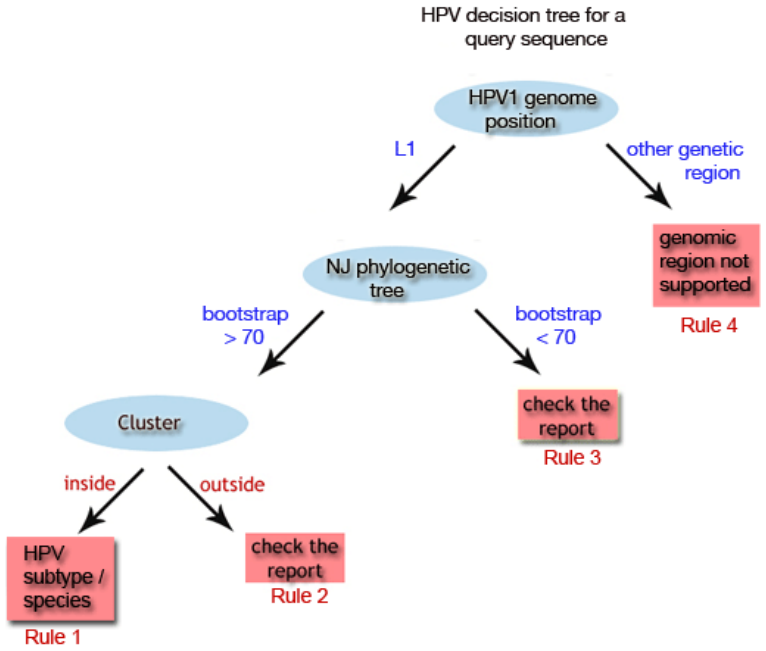
= proces výběru, prohledávání a modelování ve velkých objemech dat sloužící k odhalení dříve neznámých vztahů (+ získání obchodní výhody)

Co jsou rozhodovací stromy?

Některé metody dolování dat:

- Regresní analýza
- Shluková analýza
- Neuronové sítě
- Genetické algoritmy
- Rozhodovací stromy

  ke klasifikaci nebo predikci dat



Popis rozhodovacího stromu

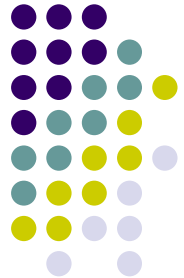
Kořen stromu
(root node)

Klient	Příjem	Konto	Pohlaví	Nezaměstnaný	Úvěr
001	vysoký	vysoké	žena	ne	ano
002	vysoký	vysoké	muž	ne	ano
003	nízký	nízké	muž	ne	ne
004	nízký	vysoké	žena	ano	ano
005	nízký	vysoké	muž	ano	ano
006	nízký	nízké	žena	ano	ne
007	vysoký	nízké	muž	ne	ano
008	vysoký	nízké	žena	ano	ano
009	nízký	střední	muž	ano	ne
010	vysoký	střední	žena	ne	ano
011	nízký	střední	žena	ano	ne

Listový uzel
(leaf node)

Neterminální uzel
(interior node)

Tvorba rozhodovacího stromu



Metoda „rozděl a panuj“ (divide and conquer) – data se postupně rozdělují na menší a menší podmnožiny (uzly stromu) podle hodnoty určitého atributu tak, aby se v těchto podmnožinách nacházely příklady jedné třídy.

→ tzv. metoda shora-dolů (top down induction)

Jak vybrat vhodný atribut pro větvení stromu?

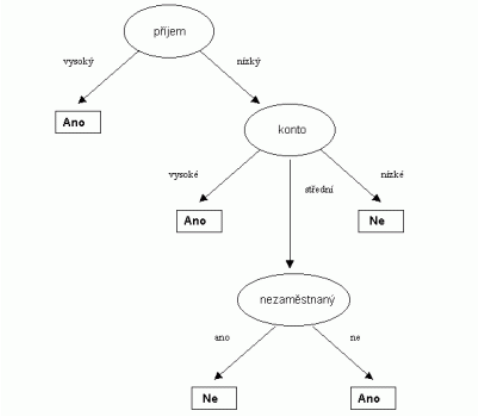
Chceme takový atribut, který od sebe co nejlépe odliší příklady různých tříd. To posuzujeme na základě teorie informace – pomocí charakteristik: entropie, informační zisk, poměrný informační zisk, chí kvadrát, Gini index.

Převod stromu na pravidla

Je výhodný pro automatizované zpracování - každé cestě stromem odpovídá jedno pravidlo.

- Nelistové uzly (atributy) a hodnota pro příslušnou hranu → předpoklad pravidla
- Listový uzel → závěr pravidla

If (prijem=vysoky) then ANO;
If (prijem=nizky & konto=vysoke) then ANO;
If (prijem=nizky & konto=nizke) then NE;
If (prijem=nizky & konto=stredni & nezamestnany=ano) then NE;
If (prijem=nizky & konto=stredni & nezamestnany=ne) then ANO;



Prořezávání stromů

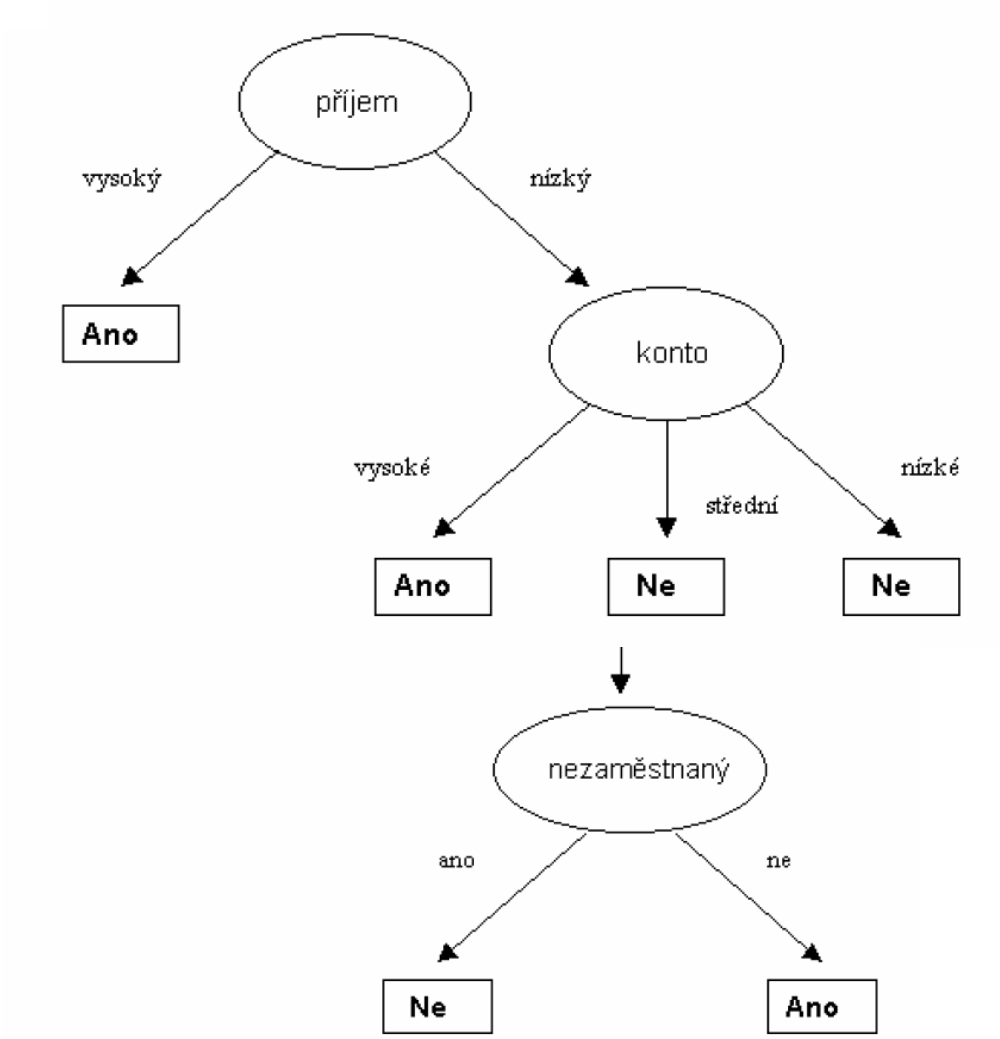
Příliš „košatý“ strom může být nesrozumitelný, navíc v případě dat zatížených šumem není bezchybná klasifikace možná.

Požadujeme tedy, aby v listovém uzlu „převažovaly“ příklady jedné třídy.

+ menší, srozumitelnější strom
- zhoršené schopnosti klasifikace nových dat

Pro jednotlivé nelistové uzly původního stromu posuzujeme, do jaké míry se strom zhorší náhradou tohoto uzlu (a jeho podstromu) listem.

Prořezávání stromů



Reálné aplikace DT

- Simulátor F16, analýza pilotova chování

cyber.felk.cvut.cz/gerstner/teaching/kui/sbirka/3_StrojU.doc

- Součást učících algoritmů umělé inteligence

http://www.automatizace.cz/article.php?a=681

- Direct marketing

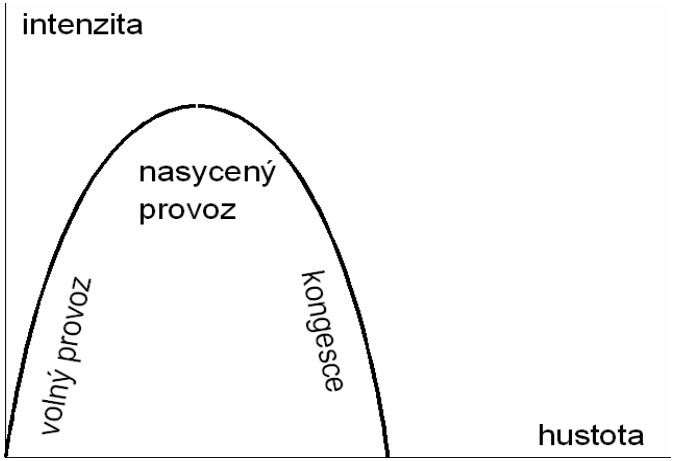
http://www.spss.cz/files/ruzne/at.pdf

Praktická část – klasifikace LOS

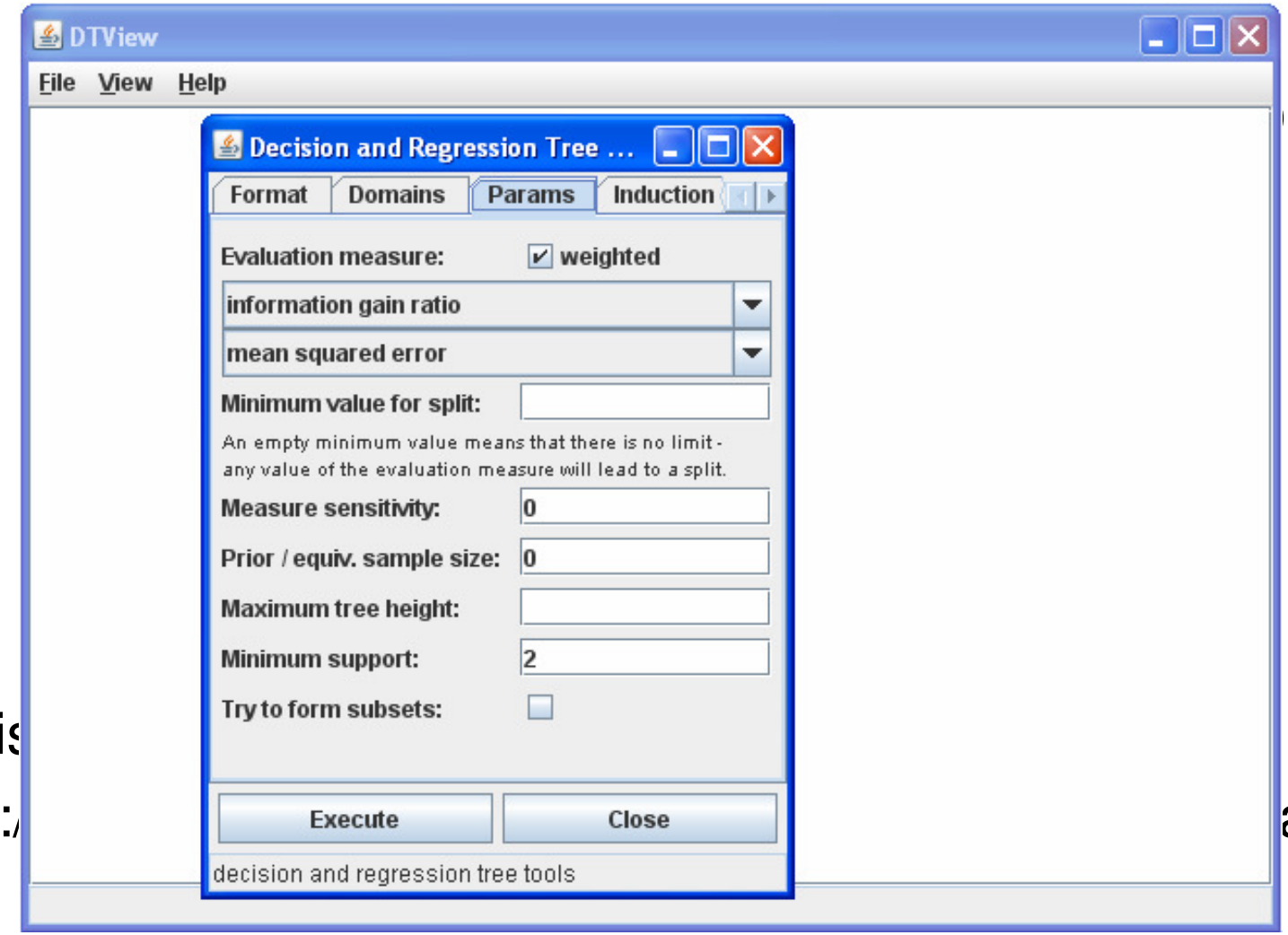
LOS = Level of Service (dopravní stupeň) se určuje expertně na základě obsazenosti a intenzity.

Cíl: Vytvořit strom(y) určující LOS podle vstupních dat. Porovnat „úspěšnost“ stromů různých parametrů.

Data naměřená optickým detektorem v ulici K Barrandovu a na Jižní spojce - přes 64 tisíc měření v době od 24.2.2006 do 18.3.2006



Programy pro tvorbu DT



Chris

http:/ ad

Postup - předzpracování

poradi;den;cas;detektor;intenzita;obsazenost;LOS

5992;pátek;24.02.2006;5:36:25;136;2;1

5993;pátek;24.02.2006;6:48:26;148;2;1

5994;pátek;24.02.2006;7:8:27;;8;0;1

5995;pátek;24.02.2006;8:0:28;;0;0;1

5996;pátek;24.02.2006;9:23:29;123;2;1

5997;pátek;24.02.2006;10:24:2;124;2;1

5998;pátek;24.02.2006;1:32:4;132;4;1

5999;pátek;24.02.2006;2:20:2;120;2;1

6000;pátek;24.02.2006;3:35:3;135;9;1

6001;pátek;24.02.2006;4:14:4;114;1;1

……

Výchozí data z detektorů

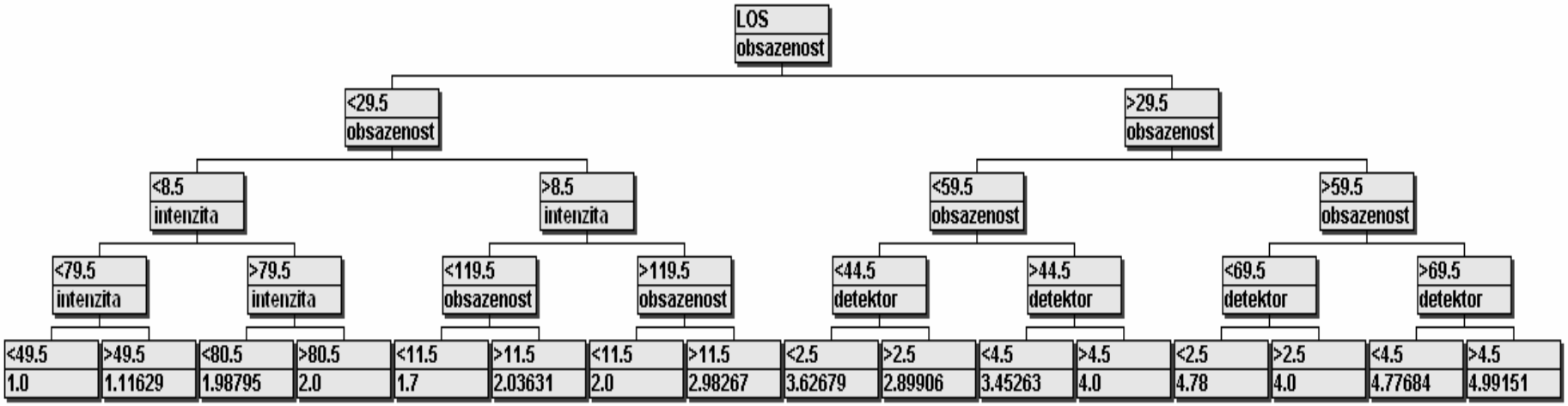
Rozdělení dat na dvě části:

- Trénovací (asi 70%)
- Testovací (zbytek)

Postup - Vytvořené stromy

1) *„úplný strom“* – při tvorbě nebyl omezen počtem úrovní. Má 10 úrovní.
2) *„prořezaný strom“* – strom vzniklý z úplného stromu prořezáním, s cílovým počtem úrovní 7
3) *„strom 7úrovní“* – strom omezený při tvorbě na 7 úrovní
4) *„strom 5úrovní“* – strom omezený na 5 úrovní

Příklad výstupu



poradi;den;cas;detektor;intenzita;obsazenost;LOS;dt

50989;sobota;16:54;4;82;16;2;2.03631

50990;sobota;16:57;8;79;53;4;4

50991;sobota;16:57;9;81;56;4;4

50992;sobota;16:57;10;85;56;4;4

Klasifikace testovacích dat

Výsledky – úspěšnost stromů

Strom_uplny

bez zaokrouhlovani

pocet spatnych vyhodnoceni	5	
uspesnost	99.974	procent

se zaokrouhlovanim

pocet spatnych vyhodnoceni	4	
uspesnost	99.979	procent

Strom_prorezany

bez zaokrouhlovani

pocet spatnych vyhodnoceni	1983	
uspesnost	89.656	procent

se zaokrouhlovanim

pocet spatnych vyhodnoceni	379	
uspesnost	98.023	procent

Strom_7 urovni

bez zaokrouhlovani

pocet spatnych vyhodnoceni	1281	
uspesnost	93.318	procent

se zaokrouhlovanim

pocet spatnych vyhodnoceni	379	
uspesnost	98.023	procent

Strom_5 urovni

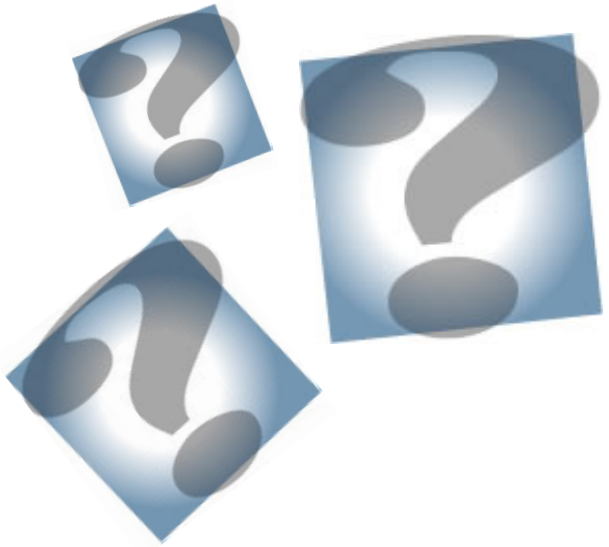
bez zaokrouhlovani

pocet spatnych vyhodnoceni	9164	
uspesnost	52.199	procent

se zaokrouhlovanim

pocet spatnych vyhodnoceni	1269	
uspesnost	93.381	procent

Dotazy?

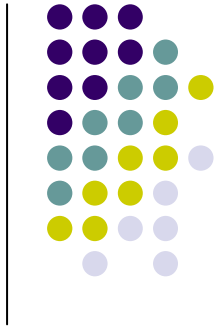


Děkuji za pozornost!

Petr Endel, 17.10.2007

petr.e@centrum.cz

Další použité zdroje:



Berka, P.: Dobývání znalostí z databází. Academia (2003)

Šarmanová, J.: Metody dolování znalostí z dat. FEI-VŠB TU Ostrava (2002)

CIspace, Decision trees:
http://www.cs.ubc.ca/nest/lci/CIspace/Version4/dTree/

Informační bulletin ČSS:
http://statspol.cz/bulletiny/ib-05-3.pdf

EuroMISE centrum:
http://euromise.vse.cz/kdd/index.php?print=metody

Nagy, I., Kratochvílová, J.: Model dopravní mikrooblasti
http://as.utia.cas.cz/doprava/files/publikace/automatizace.pdf