Zpracování výsledků pedagogického výzkumu z pohledu data miningu

Hana Havelková
PF JČU České Budějovice

Proč?

Data mining (DM)

- Data mining je netriviální proces získávání platných, dříve neznámých a potenciálně užitečných informací (znalostí) z dat
- Zpracování velkých objemů dat pomocí statistických, analytických metod

⇩

Použití pro zpracování dat získaných v rámci pedagogického výzkumu?

Cíl

- Využít open-source software pro DM
- Nabídnout další pohledy na data (klasifikace, asociace, vytváření modelů, ...)
- Realizovat testy ověřování hypotéz
- Vizualizace výsledků
- Vše pomocí jediného SW

Zaměření na dílčí úlohy z oblasti zpracování výzkumných dat

Východiska

- Data
  - skutečná (data studentů 1. ročníku IT)
  - fiktivní
- Rozumně zvolený DM software

Etapy

Data mining

- Specifikace problému
- Získání dat
- Výběr metod
- Předzpracování dat
- Skutečné dolování dat
- Interpretace výsledků

Pedagogický výzkum

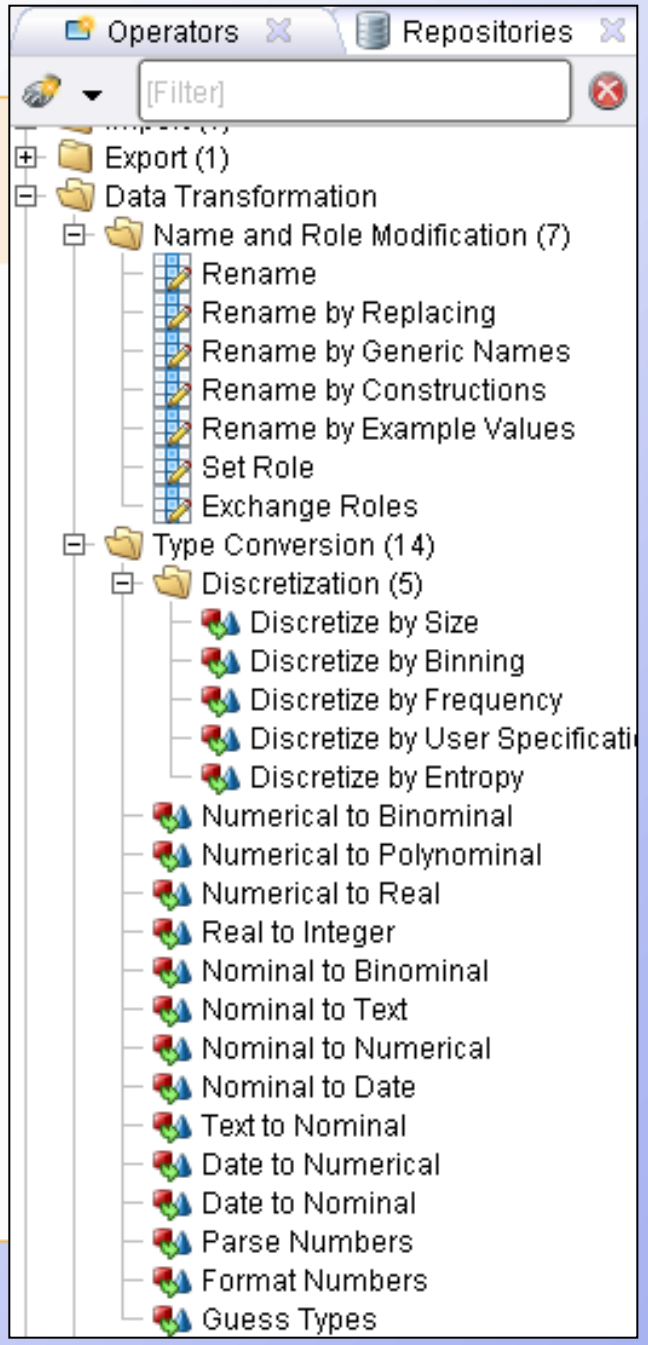
- Stanovení problému
- Formulace hypotéz
- Ověřování hypotéz
- Vyvození závěrů, jejich prezentace

Modelování

- Klasifikace dat
  - Naivní bayesovský model
  - Rozhodovací stromy
  - Rozhodovací pravidla
  - Neuronové sítě
  - …
- Regrese
- Asociace
- Shlukování

Předzpracování dat

- Typová konverze, diskretizace
- Vytváření, redukce a transformace atributů
- Modifikace hodnot
- Filtrování
- Třídění
- Čištění
- Agregace
- ...



Popisné statistiky

ExampleSet (79 examples, 0 special attributes, 7 regular attributes)

Role	Name	Type	Statistics	Range
regular	Skola	nominal	mode = ISS (33), least = G (18)	ISS (33), SOS (28), G (18)
regular	Pohlavi	nominal	mode = zena (47), least = muz (32)	muz (32), zena (47)
regular	IT zam	nominal	mode = NE (58), least = ANO (21)	NE (58), ANO (21)
regular	Obor	nominal	mode = hum (40), least = it (13)	hum (40), pri (26), it (13)
regular	Hodiny	integer	avg = 4.430 +/- 2.188	[2.000 ; 10.000]
regular	Body	integer	avg = 16.570 +/- 5.546	[6.000 ; 25.000]
regular	Splneno	nominal	mode = ano (54), least = ne (25)	ano (54), ne (25)

Attribute	Gini	Distribution			
		Values	Count	Percent	Histogram
Škola	0,6480	ISS	33	41,77 %	
		SOS	28	35,44 %	
		G	18	22,78 %	
Pohlavi	0,4820	Values	Count	Percen	
		muz	32	40	
		zena	47	59	
IT zam	0,3903	Values	Count	Percen	
		NE	58	73	
		ANO	21	26	

Parameters

Attributes : 2
Examples : 79

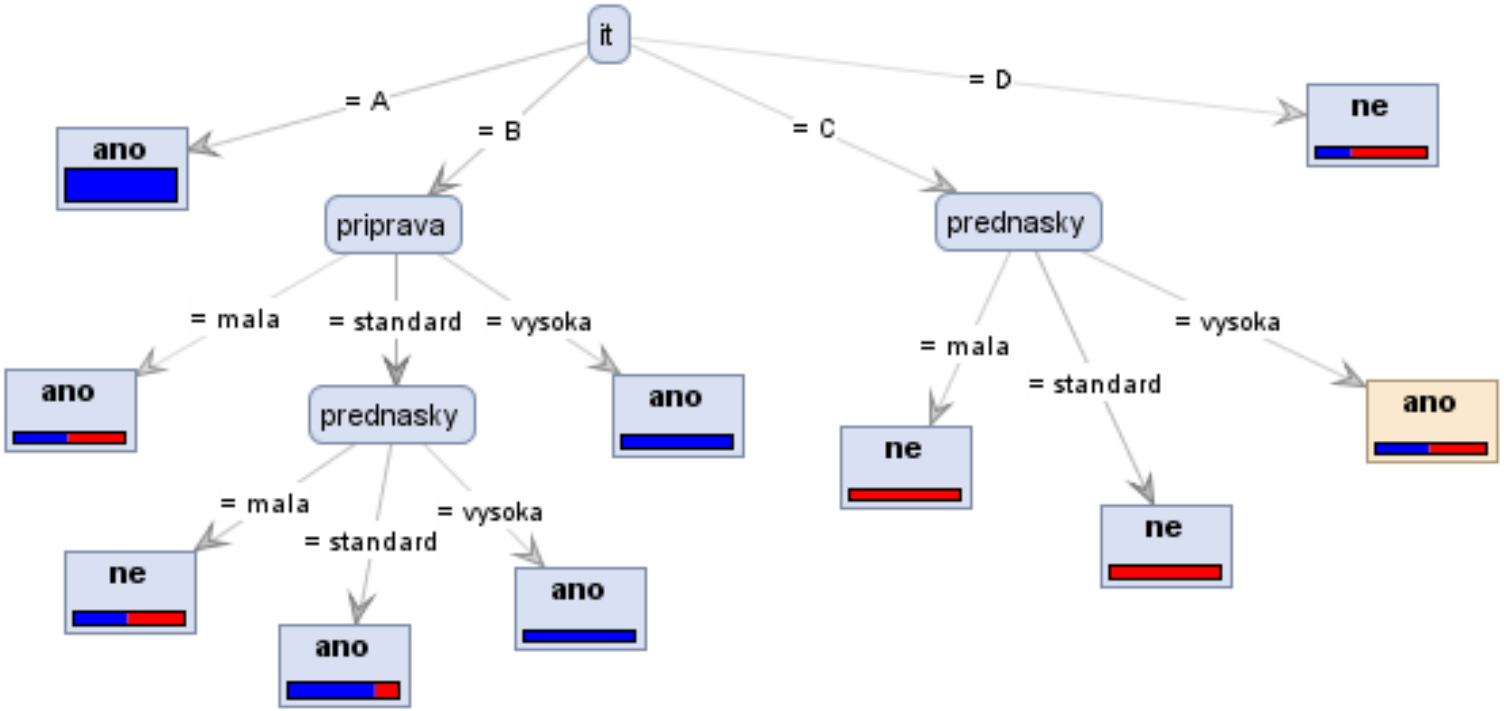
Results

Attribute	Min	Max	Average	Std-dev	Std-dev/avg
Hodiny	2	10	4,4304	2,1879	0,4938
Body	6	25	16,5696	5,5463	0,3347

Klasifikace - rozhodovací strom

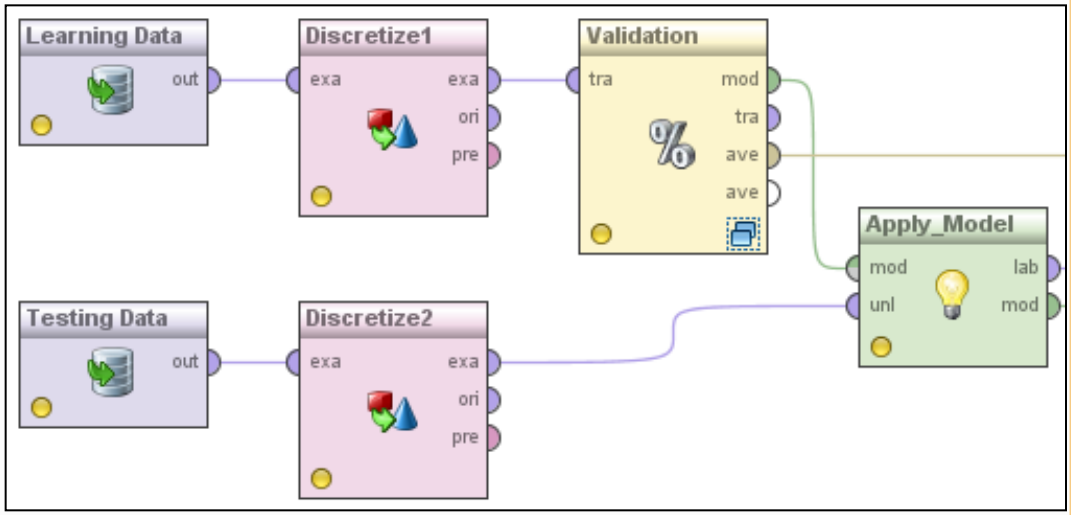
Úspěšnost studentů v předmětu PRG

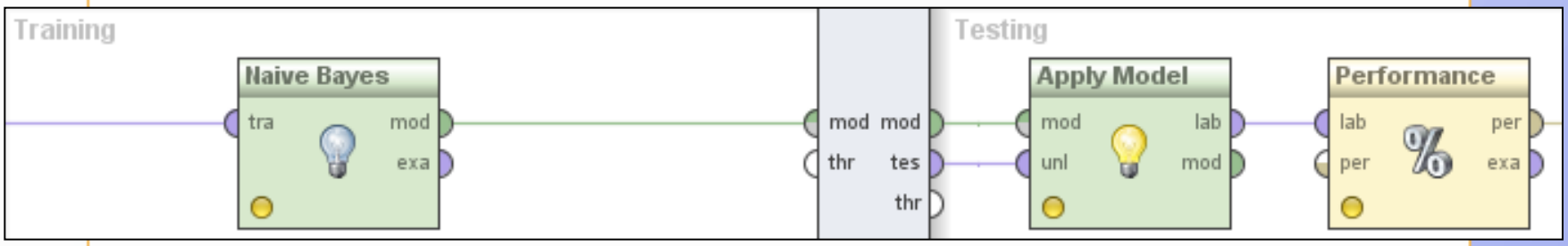
Data: vstupní hodnocení z matematiky a IT, počet hodin přípravy, prezence na přednáškách



Modelování

Data:
upravené údaje
z přijímacího řízení





Klasifikace – naivní Bayes. klas.

Meta Data View | Data View | Plot View

ExampleSet (60 examples, 4 special attributes, 4 regular attributes)

Row No.	Prijat	confidence(ano)	confidence(ne)	prediction(Prijat)	Mat	Inf	Prospech	MI
1	ano	0.591	0.409	ano	dobre	vyborne	dobre	ne
2	ano	0.790	0.210	ano	spatne	vyborne	dobre	ano
3	ano	0.790	0.210	ano	spatne	vyborne	dobre	ano
4	ano	0.956	0.044	ano	dobre	vyborne	dobre	ano
5	ano	0.956	0.044	ano	dobre	vyborne	dobre	ano
6	ano	0.999	0.001	ano	dobre	vyborne	vyborne	ne
7	ano	0.956	0.044	ano	dobre	vyborne	dobre	ano
8	ano	0.790	0.210	ano	spatne	vyborne	dobre	ano
9	ne	0.084	0.916	ne	spatne	dobre	dobre	ne
10	ne	0.063	0.937	ne	spatne	dobre	spatne	ne

Table / Plot View | Text View

Criterion Selector

- accuracy
- precision
- recall
- AUC (optimistic)
- AUC (neutral)
- AUC (pessimistic)

Table View | Plot View

accuracy: 88.33% +/- 16.75% (mikro: 88.33%)

	true ano	true ne	class precision
pred. ano	37	4	90.24%
pred. ne	3	16	84.21%
class recall	92.50%	80.00%	

MBA analýza (asociační pravidla)

Conjunction Type:

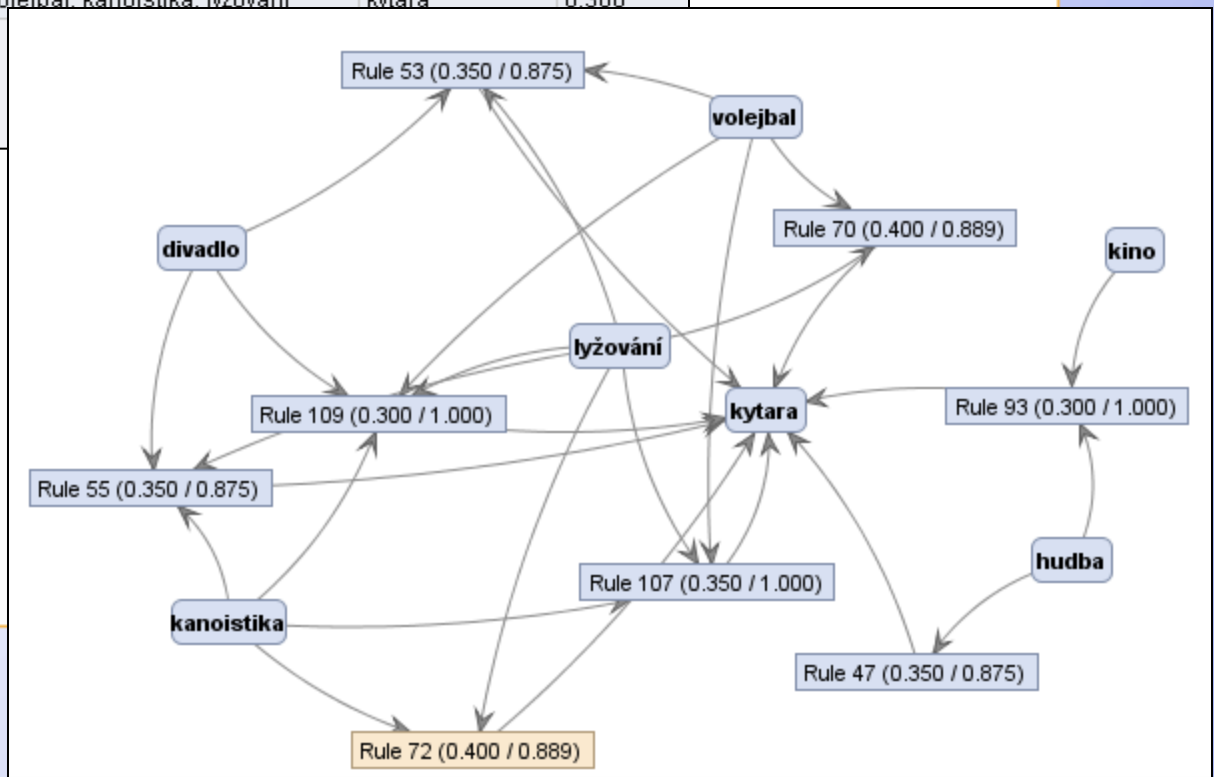
And

Conclusions:

- divadlo
- volejbal
- kanoistika
- lyžování
- kytara
- kino
- cyklistika
- běžkování
- četba

No.	Premises	Conclusion	Support
47	hudba	kytara	0.350
53	divadlo, volejbal, lyžování	kytara	0.350
55	divadlo, kanoistika, lyžování	kytara	0.350
70	volejbal, lyžování	kytara	0.400
72	kanoistika, lyžování	kytara	0.400
93	kino, hudba	kytara	0.300
107	volejbal, kanoistika, lyžování	kytara	0.350
109	divadlo, volejbal, kanoistika, lyžování	kytara	0.300

Data:
oblasti zájmu
vyučujících



Ověřování hypotéz

Testování na základě nasbíraných dat

- chí-kvadrát test dobré shody
- Fisherův test
- znaménkový test
- Wilcoxonův test
- u-test (Mann – Whitney test)
- Kolmogorovův – Smirnovův test
- Kruskalův – Wallisův test
- t-test
- ...

Dostupné testy v "Tanagře"

- Bartlett's test
- Box's M Test
- Brown - Forsythe's test
- Fisher's test
- Linear correlation
- More Univariate cont stat
- Normality Test
- One-way ANOVA
- One-way MANOVA
- Paired T-Test
- Paired V-Test
- Partial Correlation
- Semi-partial Correlation
- T-Test
- Ansari-Bradley Scale Test
- Categorical r
- Cochran's Q-test
- Contingency Chi-Square
- Friedman's ANOVA by Ranks
- Goodman Kruskal Gamma
- Goodman-Kruskal Lambda
- Goodman-Kruskal Tau
- Kendall's tau
- Klotz Scale Test
- Kruskal-Wallis 1-way ANOVA
- Van der Waerden 1-way ANOVA
- Wald-Wolfowitz Runs Test
- Wilcoxon Signed Ranks Test
- K-S 2-sample test
- Mann-Whitney Comparison
- Median test
- Mood Runs Test
- Mood Scale Test
- Sign Test
- Sommers d
- Spearman's rho

Závěr

- Import dat z nejrůznějších zdrojů
- Pohodlné předzpracování dat
- Predikce – učení na základě klasifikace, modelování, ověřování modelů
- Asociační pravidla, MBA
- Netradiční vizualizace dat a výsledků

- Otázka výběru vhodného SW ze stávající nabídky
- Možný nedostatek operátorů pro testování hypotéz
- Netriviální ovládání
- Dokonalejší práce s grafy v tabulkových kalkulátorech

Doporučení...

Zdroje

[1] BERKA, Petr: *Dobývání znalostí z databází*, 1. vydání,Praha:, Academia, 2003, 366 s. ISBN 80-200-1062-9

[2] CHRÁSKA, Miroslav: *Metody pedagogického výzkumu*, Praha, Grada Publishing a.s., 2007, 265 s. ISBN 978-80-247-1369-4

[3] PUNCH, Keith F.: *Základy kvantitativního šetření*, Praha: Portál s.r.o, 2008, ISBN 978-80-7367-381-9

[4] FAYYAD U., PIATETSKY-SHAPIRO G., SMYTH P., UTHURUSAMY R.: *Advances in Knowledge Discovery and Data Mining,* Chapter 1, AAAI/MIT Press 1996, ISBN 0-262-56097-6

[5] MOORE A.: *Statistical Data Mining Tutorials* [online]. [cit. 2010-05-23] Accessible from <http://www.autolab.org/tutorials/lists.html>

[6] OTT T.: *Neural Market Trends Tutorial* [online]. [cit. 2010-05-23] Accessible from <http://www.neuralmarkettrends.com/tutorials>

[7] http://rapid-i.com/content/view/181/196/