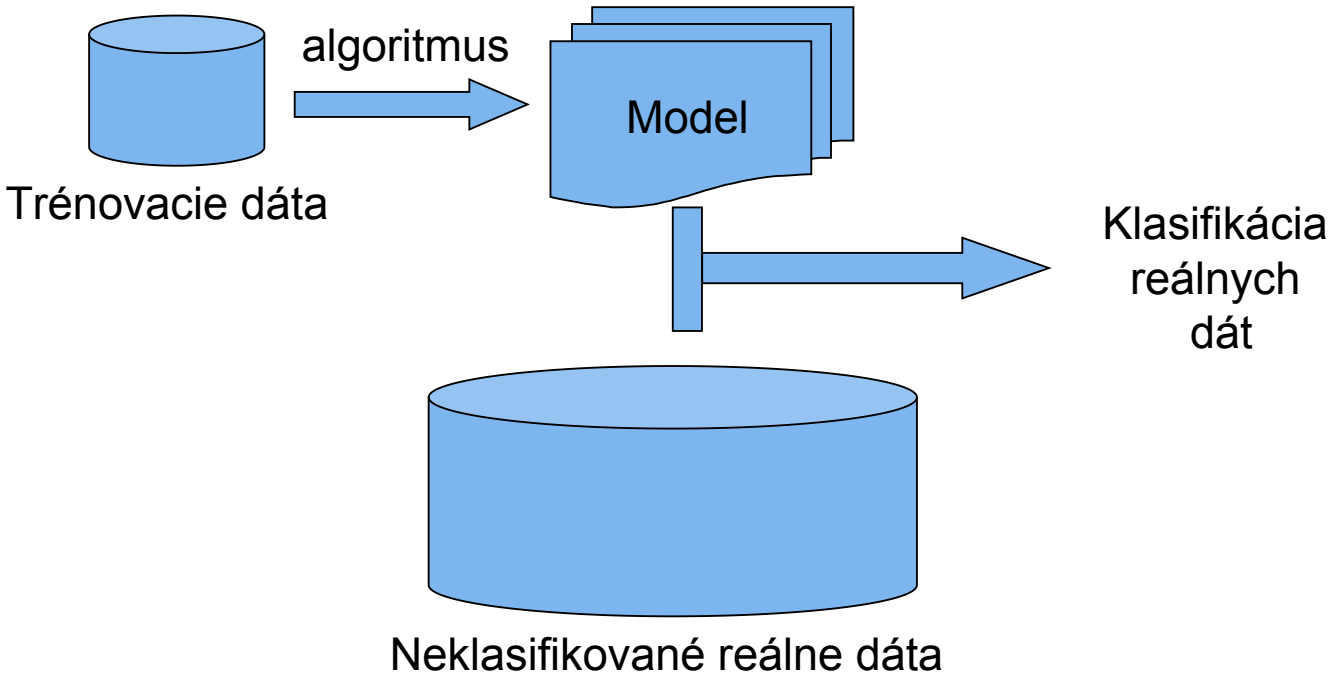
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

Využitie genetických algoritmov pri tvorbe rozhodovacích stromov

Lukáš Šurín

Klasifikovanie

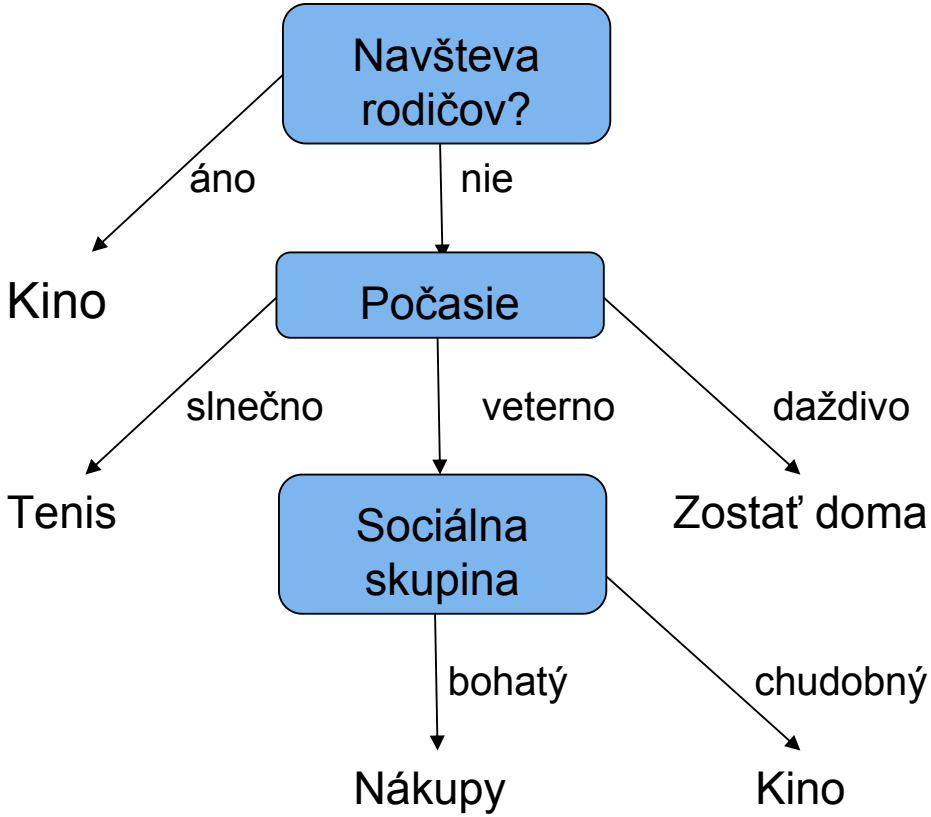


Rozhodovacie stromy

• Rozhodovací strom, ktorý určuje pre akú dennú aktivitu sa rozhodneme

• Príklad klasifikácie:
Bez rodičov, veterno, chudobný
=> **Kino**

• Indukčné algoritmy:
- ID3
- C4.5 (C5.0)
- CART



Výhody/Nevýhody

Výhody:

- robustnosť
- zrozumiteľnosť/jednoduchosť
- rozumná kvalita stromov použitím indukčných techník

Nevýhody:

- indukcia veľkých stromov
- indukcia stromov s komplikovanými kritériami

Cieľ práce

- dooptimalizovanie stromov vytvorených indukciou:
  - optimalizácia veľkosti stromu bez zníženia generalizačných schopností
  - optimalizácia stromu komplikovanými kritériami
- zásuvný modul nástroja Weka

Genetické algoritmy

Inicializácia:

• počiatočná populácia jedincov (indukčné algoritmy)

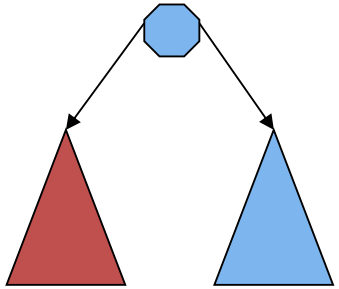
Fitness funkcie:

• matica chybovosti
• veľkosť stromu
• výška stromu
• ...

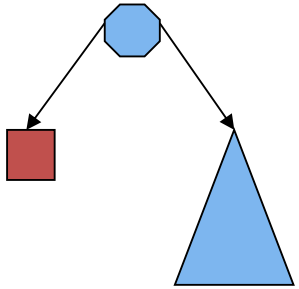
	reálne pozitívne	**reálne negatívne**
Pozitívne klasifikované	skutočne pozitívne	nepravdivo pozitívne
Negatívne klasifikované	nepravdivo negatívne	skutočne negatívne

Operátory:

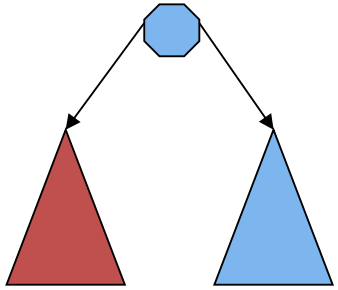
1.



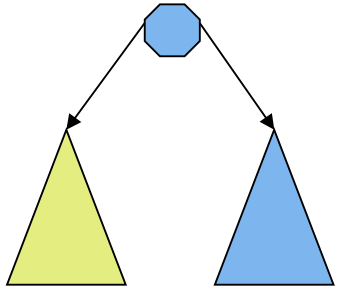
Mutácia podstromu na list
s určitou hodnotou klasifikácie

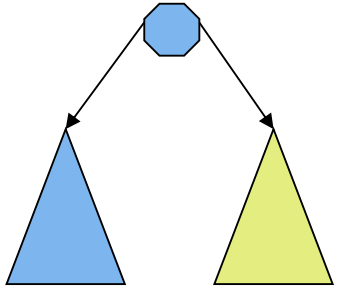


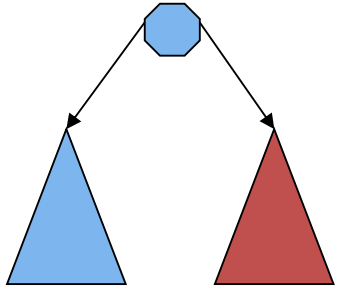
2.



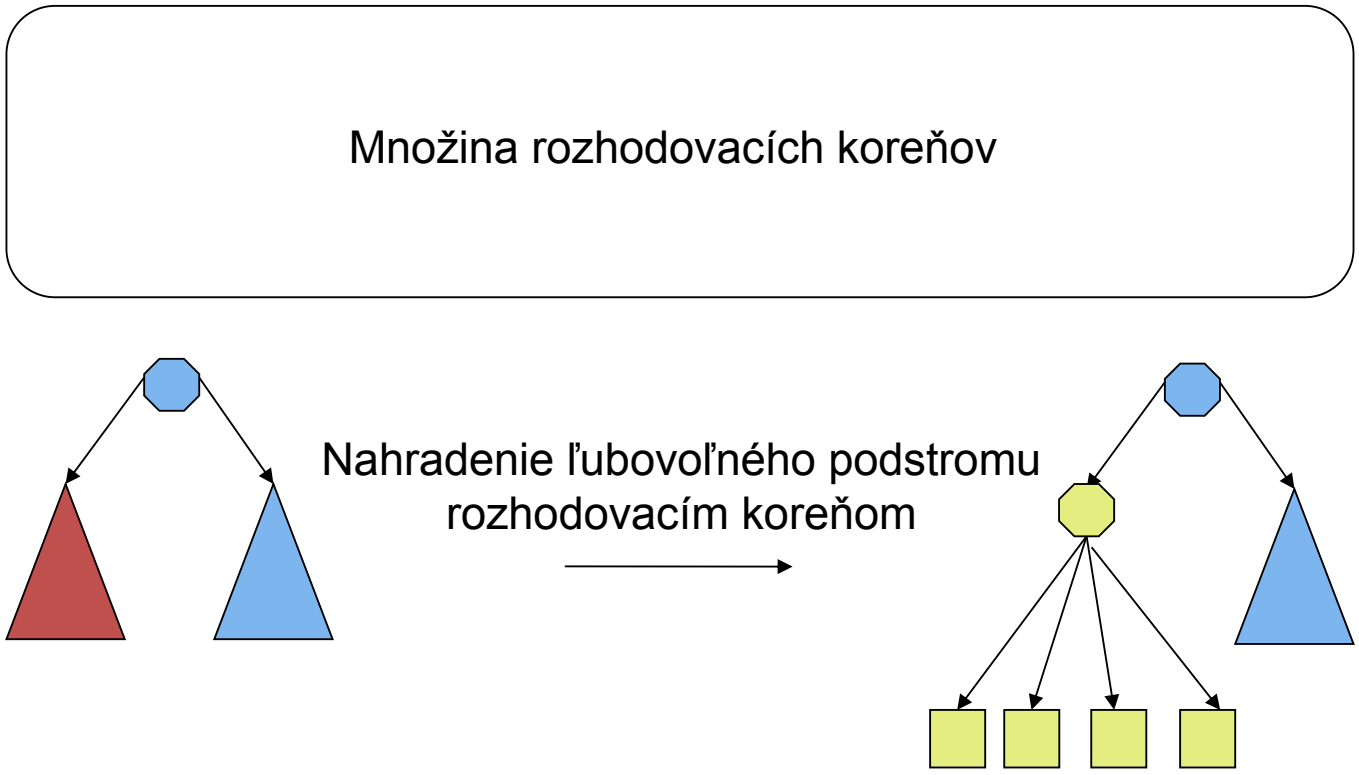
Vzájomné prekríženie
dvoch podstromov







3.


Množina rozhodovacích koreňov
Nahradenie ľubovoľného podstromu
rozhodovacím koreňom

Testovanie

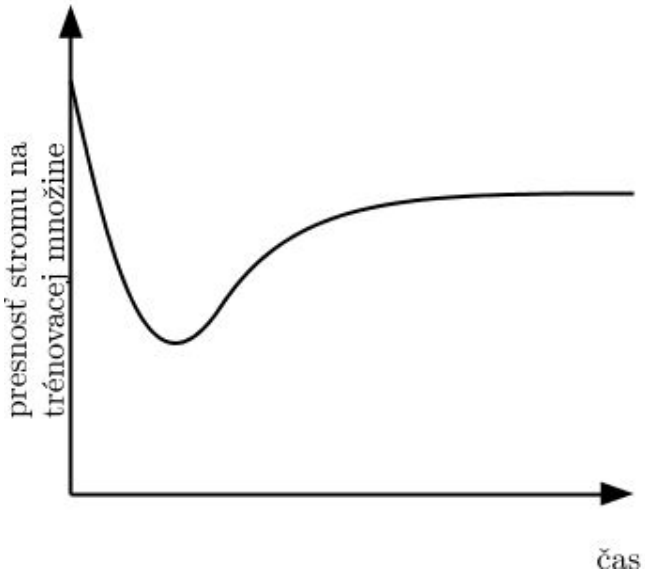
Dátová množina	# inštancií	chýbajúce hodnoty	# num.	# kat.	# tried
Colic	368	áno	7	15	2
Credit-a	690	áno	6	9	2
Credit-g	1000	áno	7	14	2
Hepatitis	155	áno	6	13	2
Iris	150	nie	4	0	3
Labor	57	nie	7	9	2
Lymph	148	nie	0	18	4
Breast-cancer	286	áno	0	9	2

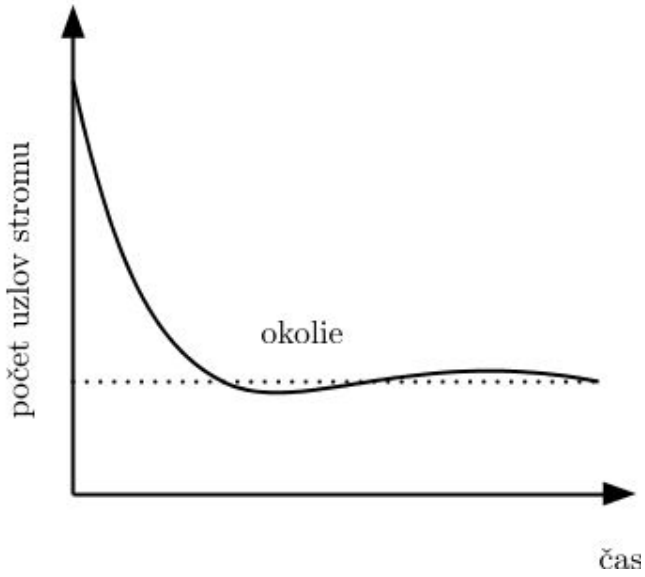
Testovanie

Parameter	Hodnota
kríženia	kríženie podstromov
mutácie	•mutácia vrcholu na list •mutácia rozhodovacími koreňmi
výber jedincov	turnaj
fitness funkcie	•presnosť modelu •počet uzlov stromu
optimalizovanie k počtu uzlom	4,6,10,15
váhy fitness funkcií	1,0.5
veľkosť populácie	100
pravdepodobnosti operátorov	0%,4%,40%,80%,90%
elitizmus	0.15

Testovanie

Typický vývoj fitness funkcií v rámci GA





ilustračne

Výsledky testov

Dátová množina	Presnosť stromu			Počet uzlov		
	C4.5	GA6	GA10	C4.5	GA6	GA10
Colic	85.16% (5.91%)	85.86% (5.59%)	85.51% (5.95%)	8.8 (2.69)	6.0 (0.0)	10.0 (0.0)
Credit-a	85.57% (3.96%)	85.39% (3.81%)	84.99% (4.24%)	32.82 (9.9)	6.0 (0.0)	10.07 (0.29)
Credit-g	71.25% (3.17%)	71.70% (2.12%)	70.79% (3.14%)	126.85 (20.66)	6.0 (0.0)	10.0 (0.0)
hepatitis	79.22% (9.57%)	80.18% (8.28%)	78.78% (9.05%)	17.66 (4.75)	6.82 (0.58)	10.94 (0.34)
iris	94.73% (5.30%)	95.80% (4.41%)	95.0% (5.14%)	8.28 (1.19)	7.0 (0.0)	9.98 (1.0)
labor	78.60% (16.58%)	84.13% (15.68%)	84.7% (15.8%)	6.92 (2.53)	6.84 (0.53)	10.02 (0.2)
lymph	75.84% (11.05%)	72.70% (9.9%)	76.61% (9.07%)	28.0 (4.56)	6.73 (0.45)	10.0 (0.0)
breast-cancer	74.28% (6.05%)	75.02% (5.22%)	73.24% (6.07%)	12.78 (9.37)	6.0 (0.0)	10.0 (0.0)

Výsledky testov

Dátová množina	Presnosť stromu			Počet uzlov		
	C4.5	GA6	GA10	C4.5	GA6	GA10
Colic	85.16% (5.91%)	85.86% (5.59%)	85.51% (5.95%)	8.8 (2.69)	6.0 (0.0)	10.0 (0.0)
Credit-a	85.57% (3.96%)	85.39% (3.81%)	84.99% (4.24%)	32.82 (9.9)	6.0 (0.0)	10.07 (0.29)
Credit-g	**71.25% (3.17%)**	**71.70% (2.12%)**	**70.79% (3.14%)**	**126.85 (20.66)**	**6.0 (0.0)**	**10.0 (0.0)**
hepatitis	79.22% (9.57%)	80.18% (8.28%)	78.78% (9.05%)	17.66 (4.75)	6.82 (0.58)	10.94 (0.34)
iris	94.73% (5.30%)	95.80% (4.41%)	95.0% (5.14%)	8.28 (1.19)	7.0 (0.0)	9.98 (1.0)
labor	78.60% (16.58%)	84.13% (15.68%)	84.7% (15.8%)	6.92 (2.53)	6.84 (0.53)	10.02 (0.2)
lymph	75.84% (11.05%)	72.70% (9.9%)	76.61% (9.07%)	28.0 (4.56)	6.73 (0.45)	10.0 (0.0)
breast-cancer	74.28% (6.05%)	75.02% (5.22%)	73.24% (6.07%)	12.78 (9.37)	6.0 (0.0)	10.0 (0.0)

Záver práce

- ciele práce splnené
- modul do Weky (dostupný z GitHub)
- stromy sú výrazne menšie, kvalita stromov neporušená
- komplikované kritéria ako kombinácia fitness funkcií

Budúce rozširenia:

- lepšia paralelizácia
- zastavovacie kritéria
- dynamická populácia

Ďakujem za pozornosť

Vyjadrenie k posudkom

- Výber najlepšieho modelu pomocou testovacej množiny

proces učenia v nástroji Weka: