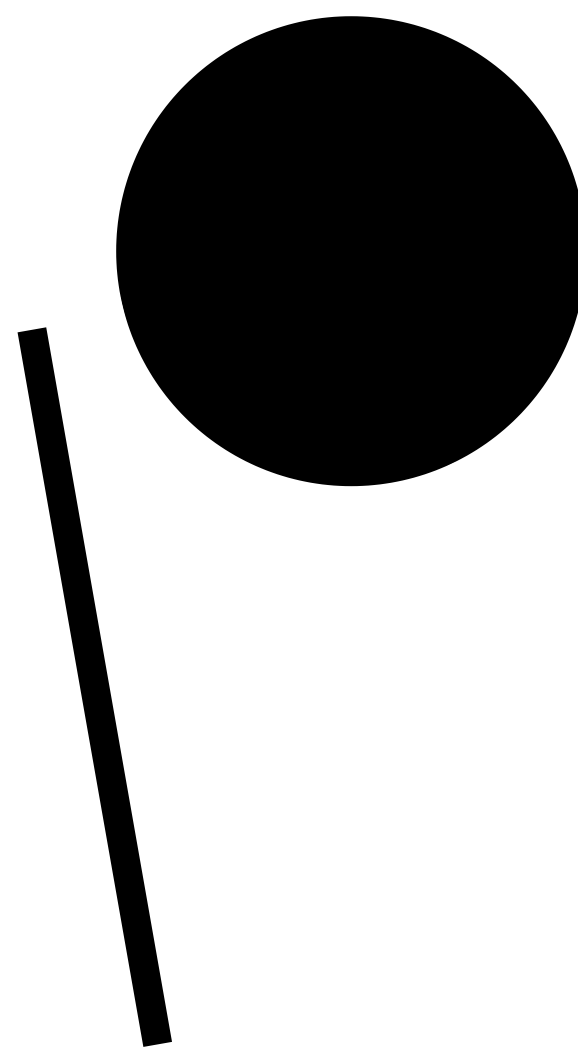


A decorative graphic consisting of a large black circle and a thick black diagonal line extending from the bottom left towards the circle.

Genetické algoritmy a genetické programovanie v dolovaní dát

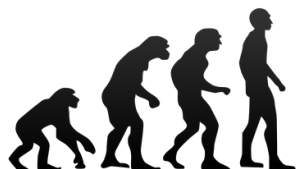
Študenti:

Bc. Mark Chernomorchenko

Bc. Eduard Matovka

Vyučujúci:

Marian Mach, M.S., Ph.D.



Prečo potrebujeme automatizované dolovanie dát?

1

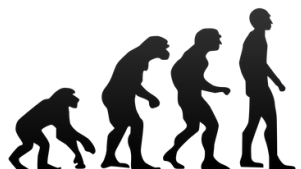
Objem dát rastie oveľa
rýchlejšie ako kapacita
analytikov

2

Skryté znalosti môžu
výrazne zlepšiť
rozhodovanie

3

Ručná analýza je
časovo náročná, drahá
a subjektívna



Od dolovania dát k objavovaniu znalostí (KDD)

Zber a integrácia dát

- databázy,
- súbory,
- senzory...

Čistenie a predspracovanie

- odstraňovanie chýb
- neúplnosť
- normalizácia

Výber a transformácia atribútov

- výber relevantných znakov pre analýzu

Dolovanie dát

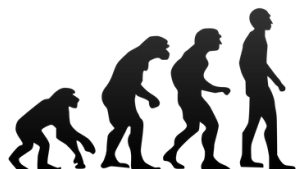
- klasifikácia
- zhlukovanie
- pravidlá...

Vyhodnotenie

- výber najlepších pravidiel
- vizualizácia
- interpretácia

Znalosť

- pochopené
- zvalidované
- použiteľné znalosti



Požadované vlastnosti získaných znalostí

Presnosť

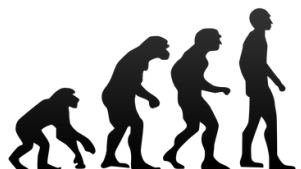
Znalosť by mala byť schopná správne predpovedať výsledky pre nové, neznáme dáta.

Zrozumiteľnosť

Pravidlo by malo byť ľahko interpretovateľné pre človeka – napr. vo forme IF–THEN.

Zaujímavosť

Znalosť by mala byť nová, prekvapivá alebo užitočná pre používateľa – nie triviálna či známa.



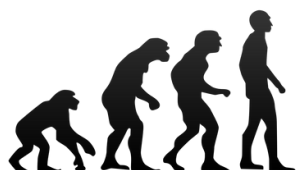
Hlavné úlohy dolovania dát

Klasifikácia: Predpovedá cieľový atribút (napr. „dobrý“ vs. „zlý“ kredit) na základe vstupných údajov.

Zhlukovanie: Zoskupovanie podobných dát bez predchádzajúcich značiek.

Asociačné pravidlá: Hľadanie vzťahov medzi položkami (napr. „Ak si kúpil X, kúpil aj Y“).

📌 **V tejto prezentácii sa zameriavame na klasifikáciu, pretože úzko súvisí s objavovaním pravidiel pomocou evolučných algoritmov.**



Čo sú to evolučné algoritmy?

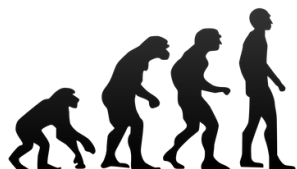
Inšpirované evolúciou v prírode

- Pracujú s populáciou jedincov
- Používajú selekciu, kríženie, mutácie...
- Cieľ: nájsť čo najlepší *organizmus* = riešenie



Príklad:

IF (Vek > 30) AND (Zamestnaný = áno) THEN Kredit = „dobrý“



Genetické algoritmy pre objavovanie pravidiel

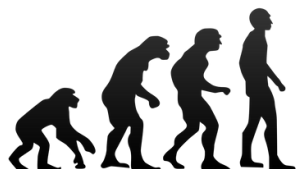
Každý jedinec v GA predstavuje jedno pravidlo:

**IF (Príjem > 3000) AND (Úver = nie) THEN
Kredit = „dobrý“**

- GA generuje počiatočnú populáciu pravidiel náhodne
- Vyhodnocuje ich pomocou fitness funkcie (presnosť, jednoduchosť)
- Aplikuje výber, kríženie a mutáciu na tvorbu lepších pravidiel

Dva hlavné prístupy:

- **Michigan:** každý jedinec = jedno pravidlo
- **Pittsburgh:** každý jedinec = celá množina pravidiel



Genetické programovanie (GP) pre objavovanie pravidiel

GP generuje pravidlá vo forme stromu:

- Každý jedinec = program = strom rozhodovania
- Uzly = operátory (AND, >, =)
- Listy = atribúty alebo hodnoty

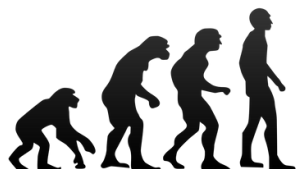
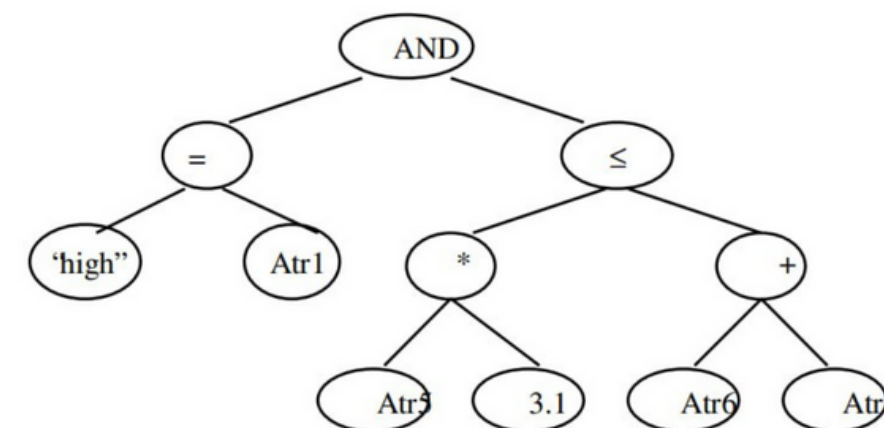
**IF (Vek > 30) AND (Pohlavie = „muž“) THEN
Kredit = „dobrý“**

Toto pravidlo je reprezentované ako strom funkcií a podmienok.

Rozdiel oproti GA:

- **GA:** pracuje s pevnou štruktúrou pravidla
- **GP:** tvorí a modifikuje programovateľné výrazy s rôznou dĺžkou

GP je flexibilnejšie a umožňuje generovať zložitejšie (ale aj zrozumiteľné) pravidlá automaticky.



Fitness funkcie: Presnosť vs. jednoduchosť vs. zaujímavosť

Presnosť (Accuracy)

- Ako dobre pravidlo predpovedá výstupné dáta.
- Metrika: napr. „confidence factor“ = $TP / (TP + FP)$

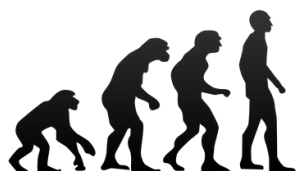
Jednoduchosť (Simplicity)

- Kratšie pravidlá sú ľahšie na pochopenie.
- Počet podmienok alebo uzlov v strome (v GP).
- Príklad penalizácie: čím viac uzlov, tým nižšie skóre.

Zaujímavosť (Interestingness)

- Pravidlo by nemalo byť triviálne alebo známe.
- Môže byť meraná subjektívne alebo cez nové kombinácie atribútov.
- Napr. „IF Pohlavie = muž THEN Kredit = dobrý“ = nezaujímavé.

V praxi je potrebné nájsť kompromis medzi týmito tromi kritériami – veľmi presné pravidlo môže byť zároveň príliš zložité alebo nezaujímavé.



Výber atribútov pomocou genetických algoritmov

Prečo výber atribútov?

Nie všetky atribúty sú dôležité – niektoré môžu byť redundantné alebo zavádzajúce.

GA umožňuje nájsť kombináciu atribútov, ktorá vedie k najlepšiemu pravidlu.

1. Binárne kódovanie (0/1):

Každý gén = jeden atribút

1 = použije sa, 0 = ignoruje sa

Príklad: [1, 0, 1, 1, 0]

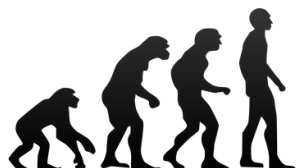
2. Váhové kódovanie:

Každému atribútu sa priradí váha (0 až 1)

Vyjadruje dôležitosť daného znaku

Príklad: [0.8, 0.0, 0.4, 0.9]

GA prehľadáva priestor možných kombinácií a vyhodnocuje, ktoré atribúty prispievajú k najlepším pravidlám.



Konštrukcia nových atribútov pomocou GP

Tradičné pravidlá porovnávajú atribúty s konštantami:

IF Vek > 30 THEN ...

Ale nie: IF Príjem > Výdavky THEN ...

Riešenie pomocou GP:

GP môže vytvoriť nový atribút:

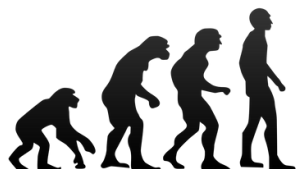
NovýAtribút = (Príjem > Výdavky)

IF NovýAtribút = áno THEN Akcia = „kúpiť akcie“

GP kombinuje viaceré existujúce
znaky do jedného výrazu

Odhaľuje skryté vzťahy medzi atribútmi

Zlepšuje klasifikáciu bez potreby
ručnej analýzy



Post-spracovanie získaných znalostí

Problém:

Po dolovaní získame príliš veľa pravidiel alebo modelov.

Nie všetky sú užitočné alebo spoľahlivé.

Ciele post-spracovania:

- Znížiť počet pravidiel (pruning)
- Zvýšiť zrozumiteľnosť (zjednodušiť podmienky)
- Zvážiť dôležitosť rôznych modelov (váhovanie v ansámblach)



Použitie GA:

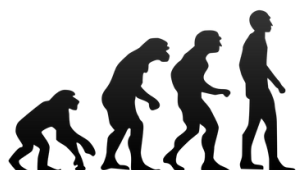
Každý jedinec = vektor váh
pre modely/pravidlá
GA optimalizuje váhy tak,
aby výsledný ansámbel bol
čo najpresnejší



Príklad:

[0.6, 0.0, 0.2, 0.2]

Model s váhou 0.0 je vylúčený z hlasovania



Výzvy a smerovanie ďalšieho výskumu

Ako merať „zaujímavosť“ pravidiel?

- Presnosť a zrozumiteľnosť sa dajú zmerať, ale zaujímavosť je často subjektívna.
- Potrebné sú kombinované metriky alebo interaktívne hodnotenie používateľom.

Prispôsobenie operátorov GA/GP konkrétnym úlohám

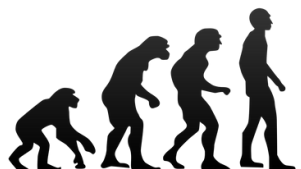
- Nie všetky kríženia a mutácie sú vhodné pre pravidlá → treba „inteligentné“ operátory.

Rozšírenie mimo klasifikácie

- Väčšina výskumu sa sústreďí na klasifikáciu, ale GA/GP majú potenciál aj pre:
 - asociačné pravidlá
 - modelovanie závislostí
 - zhlukovanie

Interaktívne evolučné systémy

- Spojenie GA/GP s ľudskou spätnou väzbou → používateľ pomáha hodnotiť, čo je zaujímavé.



Zhrnutie a hlavné poznatky

Evolučné algoritmy sú účinným nástrojom pre dolovanie znalostí

- Najmä v úlohách, kde je priestor riešení veľmi rozsiahly a zložitý.

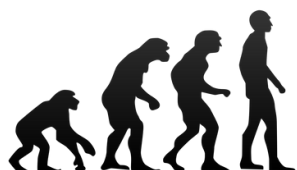
GA a GP dokážu nielen nájsť pravidlá, ale aj:

- Vybrať dôležité atribúty
- Vytvoriť nové (zložené) atribúty
- Optimalizovať modely aj po dolovaní

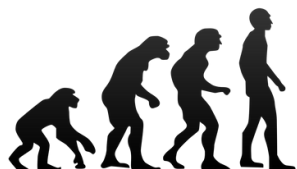
Dobré pravidlá musia byť:

- Presné, zrozumiteľné, a zaujímavé
- GA/GP umožňujú hľadať kompromis medzi týmito cieľmi

Evolučné algoritmy nám neponúkajú „hotové pravdy“, ale výkonný nástroj na ich objavovanie.



Ďakujeme
za pozornosť



Povodny clanok:

<https://fmfi-uk.hq.sk/Informatika/Strojove%20Ucenie/AdvEC-bk.pdf>