3 Symbolická regrese: Jak získat kvalitní a současně kompaktní řešení?

Optimalizace výstupu geometrického sémantického genetického programování pomocí kartézského genetického programování

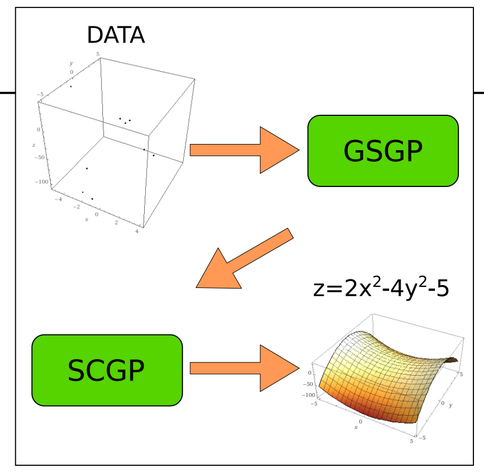
Abstrakt

Geometrické sémantické genetické programování (GSGP) dosahuje kvalitních výsledků při popisu dat složitými matematickými modely. Cenou za přesný popis je ale výsledná velikost řešení. Tento článek se zabývá převodem řešení získaného GSGP na instanci kartézského genetického programování (CGP) a jeho následnou optimalizací. CGP dokáže dobře redukovat velikost již vzniklých řešení a kombinace těchto dvou metod má tak potenciál vytvořit kvalitní a zároveň malý model popisující vstupní data. Úspěšným principem redukce je podstromové (subtree) CGP (SCGP), které je představeno v tomto článku. Využívá možnosti rozdělení řešení na podstromy a následně je upravuje. Na všech testovaných úlohách z oblasti symbolické regrese se podařilo dosáhnout znatelného zmenšení řešení a pouze u jedné úloh ze 4 došlo k přetrénování. Kombinace GSGP a SCGP tak má potenciál vytvořit dostatečně dobrý model, který je i přiměřeně velký a to v rozumném čase.

GSGP

- GSGP upravuje jedince na úrovni sémantiky
- křížení: $T_{XO}=(T_1 \bullet T_R)+((1-T_R) \bullet T_2)$
- mutace: $T_M=T+ms \bullet (T_{R1}-T_{R2})$
- mutace zvětšuje jedince konstantně
- křížení zvětšuje jedince v závislosti na generaci
- velikost populace: 2000; počet generací: 1000 (300 pro P3D); počet náhodných stromů: 500; metoda tvorby stromů: Ramped Half-and-half; max. hloubka stromu: 8; pravděp. křížení: 0; pravděp. mutace: 0,9; velikost turnaje: 8
- ze 40 běhů vybrán nejlepší výsledek

SCGP



- používá výstup GSGP bez křížení
- chromozom CGP se rozdělí na sémanticky spjaté části
- části jsou upravovány pomocí CGP
- části se po úpravě spojují po dvou
- úprava a spojování probíhá tak dlouho, dokud nevznike opět jeden chromozom
- velikost populace: 8; počet generací: 50; pravděp. mutace: 0,08; metoda selekce: Nejméně aktivních uzlů a fitness alespoň tak nízká, jako u GSGP; selekce podstromů: náhodný výběr

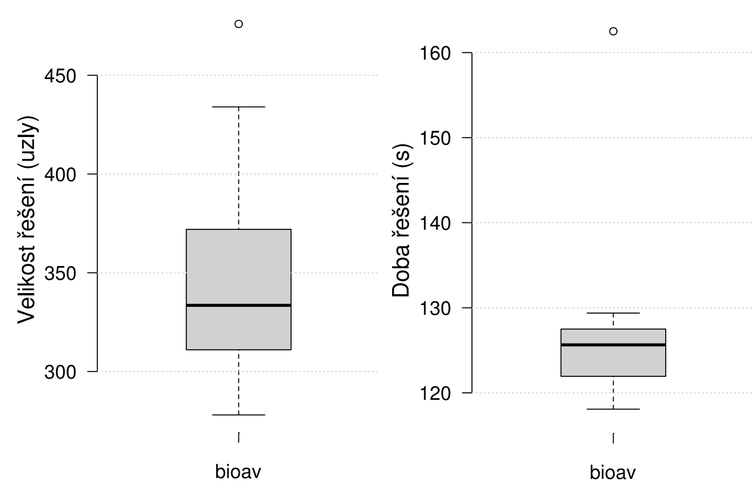
Experimenty

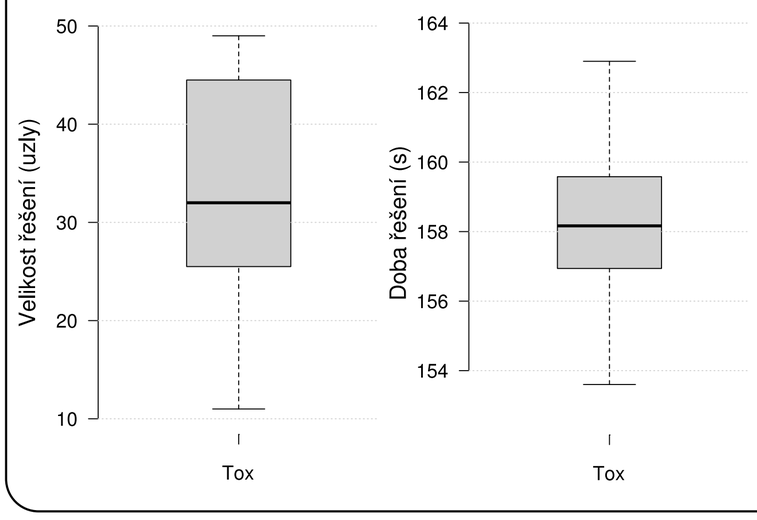
- úlohy z oblasti farmakokinetiky - lidská orální biodostupnost (*bioav*); medián smrtící dávky (*Tox*); 3D struktura proteinu (*P3D*); úroveň vazby plazmatických bílkovin (*PPB*)
- použita data z 20-ti nezávislých běhů

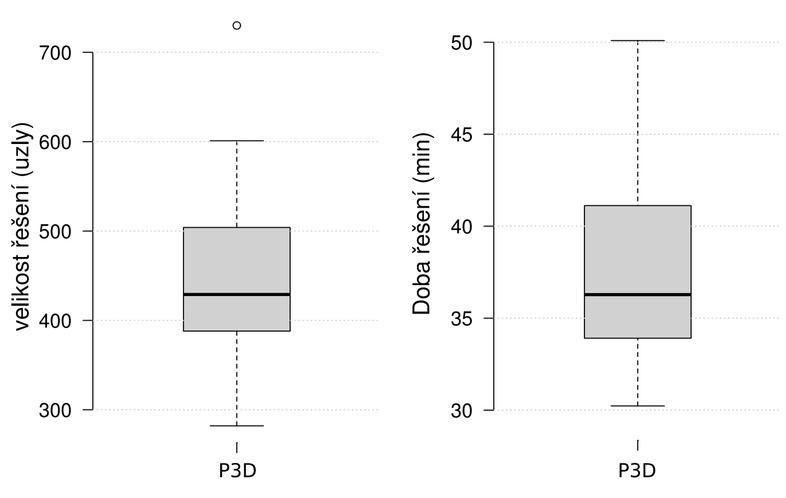
Výsledky

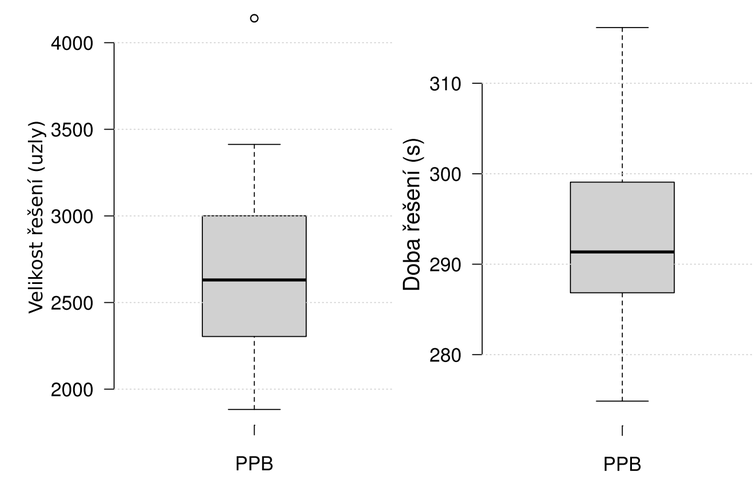
	počet uzlů GSGP	nejkratší SCGP
bioav	22285	330
Tox	21568	11
P3D	5764	282
PPB	22907	2722

- u *bioav*, *Tox*, *P3D* je zmenšení větší jak 95%
- zmenšení u *PPB* činí 88%, ale nastává přetrénování









Bc. Ondřej Končal, xkonca00@stud.fit.vutbr.cz
Vedoucí: prof. Ing. Lukáš Sekanina, Ph.D.

Excel@FIT 2018

BRNO UNIVERSITY OF TECHNOLOGY FACULTY OF INFORMATION TECHNOLOGY