Fakulta Informatiky a Informačných Technológií
Slovenská Technická Univerzita
Bratislava

Optimalizačný algoritmus využívajúci genetické výpočty pre dolovanie v údajoch

Peter Krammer
peter.krammer@savba.sk

Seminár z Umelej Inteligencie, 07. 03. 2011

Optimalizácia

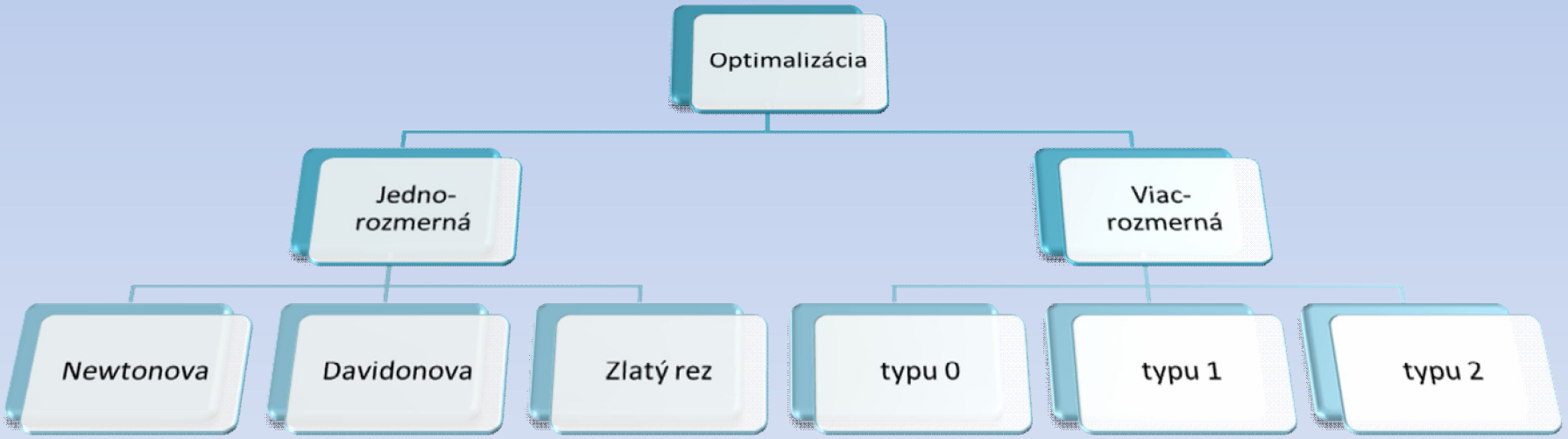
- *zostavenie portfólia tak, aby bola maximalizovaná návratnosť investícií, pre danú úroveň rizika, alebo je minimalizované riziko, pre danú úroveň očakávanej návratnosti (www.investorwords.com).*

- *hľadanie najlepších riešení medzi alternatívami, alebo hľadanie extrému hodnoty premennej alebo funkcie (www.esse.ou.edu).*

- *matematická disciplína, ktorá zaoberá získavaním maxima a minima funkcií, s ohľadom na obmedzenia (www.friartuck.net).*

Optimalizácia

- praktické využitie (ekonomika, priemysel, doprava, ...)

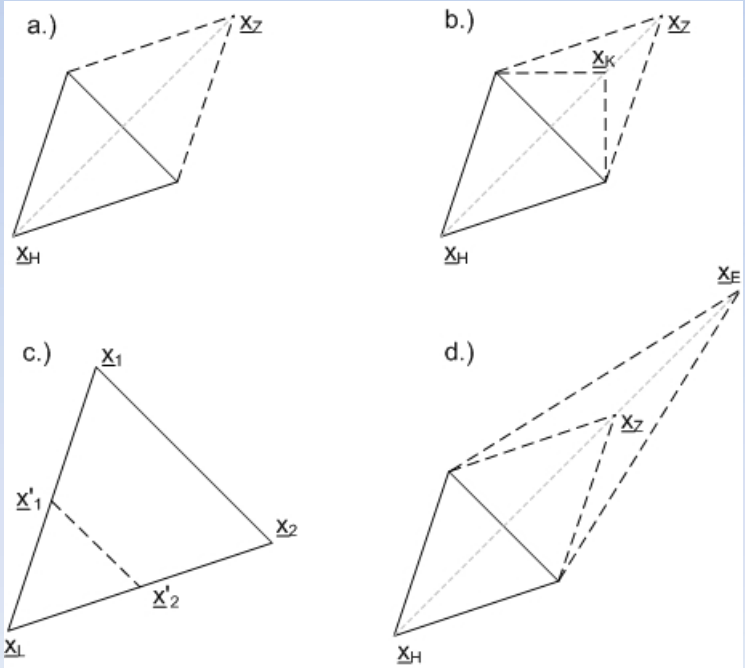


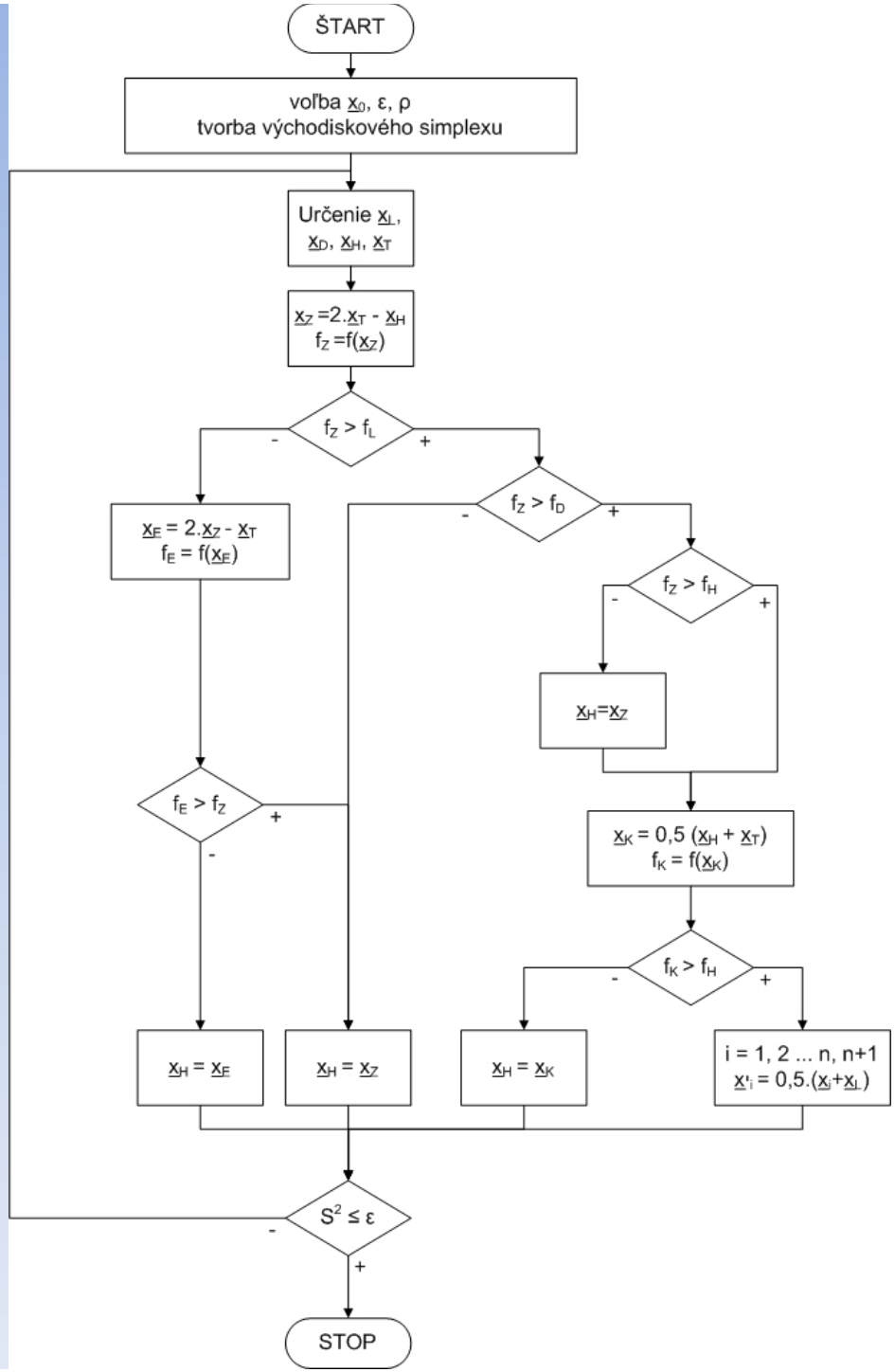
Úlohy vhodné pre optimalizáciu

	Pôvodná úloha	**Transformovaná úloha na úlohu optimalizácie** $f(\underline{x}) \rightarrow \min$
Maximalizácia funkcie	$g(\underline{x}) \rightarrow \max$ $\underline{x}_{max} = ?$	$f(\underline{x}) = -g(\underline{x})$
Hľadanie riešenia rovnice	$h(\underline{x}) = t(\underline{x})$ $h(\underline{x}) - t(\underline{x}) = 0$ $\underline{x}_0 = ?$	$f(\underline{x}) = abs(h(\underline{x}) - t(\underline{x}))$
Viac-kriteriálna optimalizácia	$f_1(\underline{x}) \rightarrow \min$ $f_2(\underline{x}) \rightarrow \min$ $f_3(\underline{x}) \rightarrow \min$ $\underline{x}_{min} = ?$	$f(x) = w_1.f_1(\underline{x})+w_2.f_2(\underline{x})+w_3.f_3(\underline{x})$ $w_1, w_2, w_3 > 0,$ kde w_i reprezentuje váhy jednotlivých kritérií

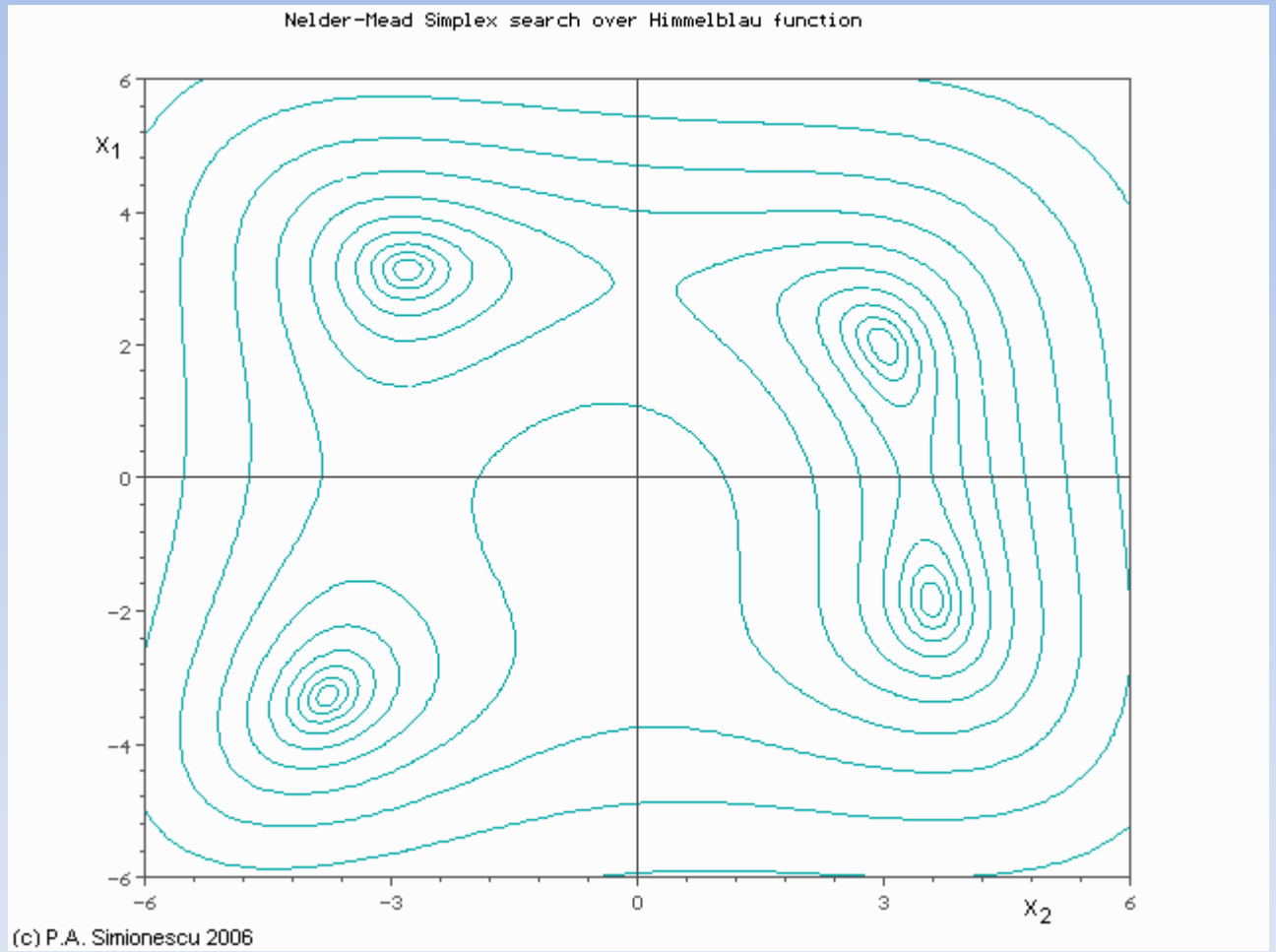
Metóda Neldera a Meada

$\underline{x}_L$ - "najlepší" vrchol simplexu
$\underline{x}_H$ - "najhorší" vrchol simplexu
$\underline{x}_D$ - "druhý najhorší" vrchol simplexu
$\underline{x}_Z$ - nový testovaný vrhol simplexu
$\underline{x}_E$ - vrchol získaný expanziou
$\underline{x}_K$ - vrchol získaný kontrakciou
$\underline{x}_T$ - "ťažiskový" vrchol simplexu





Priebeh konvergencie algoritmu N-M



Hybridný optimalizačný algoritmus

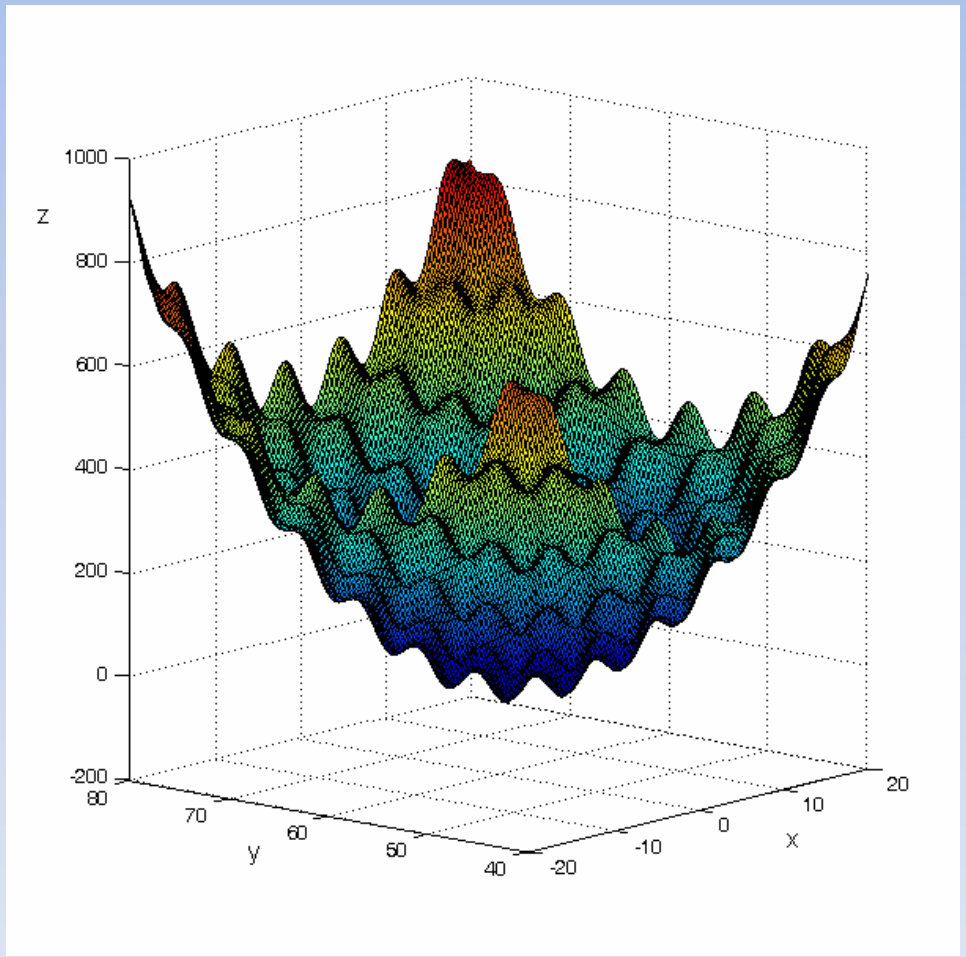
- motivácia
- výhody / nevýhody
- idea hybridného algoritmu (genetický algoritmus + metóda pružného simplexu)

Popis algoritmu

- 10 generácií x 50 jedincov
- Inicializácia jedincov $N(\mu, \sigma^2) = N(0, 100^2)$
- Operácie na množine jedincov :
  - kríženie
  - mutácia
  - priemerovanie
  - nový náhodný jedinec
  - metóda N-M

Testovanie algoritmu

Teoretická funkcia



$z = f(x, y) = 50.\sin(x).\cos(y-58) + x^2 + (y-58)^2$

Dáta z projektu ADMIRE

- Pilotná aplikácia *Flood Forecasting* Simulation Cascade
- Modely pre predikovanie hydrometeorologických veličín
- 8 vstupných veličín
- 8760 záznamov

$$f(\underline{p}) = \sum_{i=1}^{M} (\mathrm{model}(\underline{p}, \underline{d}_i) - y_i)^2$$

Nastavenie algoritmu

- Genetického algoritmu
  - počet generácií, počet jedincov v generácií
  - pravdepodobnosť mutácie
  - štartovacie rozdelenie jedincov $X \sim N(0, 100^2)$
  - tolerancia
- Algoritmu N-M
  - krok
  - epsilon
  - koeficienty expanzie, kontrakcie, redukcie

Prehľad najlepších jedincov v jednotlivých generáciách

generácia	0.	1.	2.	Č. exp.
Hodnota krit. f-cie najlepšieho jedinca v danej gener.	463.166	-47.2008	-47.6275	1
	77.6319	-47.6275	-47.6275	2

generácia	0.	1.	2.	Č. exp.
Hodnota krit. f-cie najlepšieho jedinca v danej gener.	6699E6	194098	18655	1
	6439E6	36312	18655	2

Číselné charakteristiky modelu (Admire)

- model(p, d) = p(0).d(0) + p(1).d(1) + p(2).d(2) + p(3).d(3) + p(4).d(4) + p(5).d(5) + p(6).d(6) + p(7)
  - Root mean squared error : 1.459308
  - Correllation coefficient : 0.9639
  - Mean absolute error : 1.1791
  - Relative absolute error : 23.8739 %

- model (p, d) = p(0).d(0) + p(1).d(1) + p(2).d(2) + p(3).d(3) + p(4).d(4) + p(5).d(5) + p(6).log(d(6)) + p(7)
  - Root mean squared error : 1.398013 (1.0386)
  - Correllation coefficient : 0.9716 (0.9821)
  - Mean absolute error : 1.1712 (0.7748)
  - Relative absolute error : 20.4017 % (15.6884 %)

Paralelizácia hybridného algoritmu

- Parametrická štúdia
  - každý výpočtový uzol prehľadáva inú oblasť priestoru parametrov
- Paralelné testovanie rôznych štruktúr modelu
- Paralelizácia na úrovni GA
  - paralelné realizovanie mutácií, krížení, konvergovanie alg. N-M

Zhrnutie

- Vo všetkých prípadoch experimentov bol nájdený globálny extrém.
- Vhodný na hľadanie štruktúry modelu.
- Plánované rozšírenia algoritmu
  - dynamická zmena štruktúry modelu
  - implementácia paralelizmu pre klaster

Referencie

- Evolučné výpočty a ich využitie v praxi, SEKAJ Ivan, Iris, 2005, ISBN 80-89018-87-4
- Optimalizácia, HUDZOVIČ Peter, STU Bratislava, 2004, ISBN 80-227-2072-0
- Projekt ADMIRE - Advanced Data Mining and Integration Research for Europe (ADMIRE), EU FP7 ICT project no. 215024. http://www.admire-project.eu (accessed Sept. 2010).
- Analyzing GeneXproTools Models Statistically, Relative Absolute Error, http://www.gepsoft.com/gxpt4kb/Chapter10/Section4/SS15.htm
- WEKA, University of Waikato, http://www.cs.waikato.ac.nz/ml/weka/
- WITTEN H., EIBE F.: Data mining: Practical Machine Learning Tools and Techniques, University of Waikato, Eslevier: 2005.

Ďakujem za pozornosť

otázky ...