INVESTICE
DO ROZVOJE
VZDĚLÁVÁNÍ

Restrikce měřící škály geodat pomocí genetických algoritmů

Jiří Dvorský, Vít Paszto
Katedra geoinformatiky, UP Olomouc
jiri.dvorsky@upol.cz,
vit.paszto@gmail.com
Lenka Skanderová
Katedra informatiky, VŠB – TU Ostrava
lenka.skanderova@vsb.cz

Úvod

- kvantitativní data, vzniklá měřením např. teplota vzduchu,
- problémy při dalším zpracování primárních dat:
  - primární data jsou příliš rozsáhlá,
  - přesnost překračující potřeby,
  - snižuje se naše schopnost porozumět měřenému jevu.
- je tedy nutné redukovat škálu dat,
- rozdělit data na intervaly (**restrikce škály**) a každý interval najít vhodného reprezentanta,
- tím ale dojde ke ztrátě informace,
- snahou je tuto ztrátu minimalizovat.

Úvod

Informační entropie:

- Claude E. Shannon (1948),
- míra informace,
- jednotkou je jeden bit,
- informační entropie je závislá na pravděpodobnosti jevu – čím méně pravděpodobný jev, tím více obsahuje informace a naopak,
- příklad – srovnej informační obsah jevu, že v menze je UHO/UHOŘ nebo půlkilový T-bone steak ...

Úvod

Cíl výzkumu

Využít genetický algoritmus k nalezení takové restrikci škály dat, která by minimalizovala informační ztrátu během transformace dat.

Restrikce škály v geovědách

- poměrně častá úloha,
- intervaly restrikce musí pokrývat celý rozsah naměřených hodnot,
- restrikce musí korespondovat s primárními daty,
- často používané metody:
  - rovnoměrné dělení – intervaly stejné délky,
  - dělení na základě kvantilů – intervaly jsou určeny na základě histogramu primárních dat,
  - **Jenksova metoda** – minimalizace střední kvadratické odchylky dat spadajících do určitého intervalu od jeho střední hodnoty a zároveň maximalizace téže odchylky od středních hodnot všech ostatních intervalů, nejpoužívanějších v geovědních oborech.
  - manuální dělení.

Možnosti matematiky v řešení problémů

- analytické řešení – rovnice, soustavy rovnic, vzorce,
- pro určité problémy:
  - není známo analytické řešení – voda tekoucí z kohoutku,
  - analytické řešení nelze spočítat v rozumném čase.
- obvykle ale umíme otestovat zda například „nějaká čísla“ jsou vůbec řešením našeho problému,
- umíme většinou změřit kvalitu tohoto řešení,
- problémem zůstává jak generovat „nějaká čísla“.

Možné další přístupy

Jednou z odpovědí jak systematicky generovat pokud možno stále lepší a lepší řešení jsou tzv. **genetické algoritmy**.

Genetický algoritmus (GA)

- patří do velké skupiny tzv. bio-inspirovaných výpočetních metod,
- představuje jakýsi obecný mechanismus, jak hledat kvalitní řešení problému,
- založen na darwinistických principech – přirozeném výběru, křížení a mutaci,
- GA spravuje populaci jedinců,
- jedinec představuje jedno z možných řešení problému, řešení je kódováno jako genom,
- každý jedinec je ohodnocen kvalitou jím reprezentovaného řešení, tzv. fitness,
- snažíme se systematicky generovat stále kvalitnější a kvalitnější jedince (řešení),
- jde minimalizaci fitness funkce.

Genetický algoritmus (GA)

- ze stávající generace v populaci je produkována generace nová,
- dvě základní operace:
  - křížení – ze dvou rodičovských jedinců vzniká nový jedinec, mix genů,
  - mutace – náhodná změna genomu s jistou malou pravděpodobností.
- nová populace sestavena z nejkvalitnějších jedinců,
- po daném počtu generací algoritmus ukončíme.

Před použitím GA nutno definovat

- reprezentace jednice,
- operace křížení a mutace,
- fitness funkce.

Odhad počtu možných restrikcí

Označení

- n různých unikátních vstupních hodnot,
- hledáme restrikci na M intervalů,
- počet možných restrikcí $R(n, N)$.

Počet všech možných restrikcí škály

n	N	$R(n, N)$
10	5	126
20	10	92 378
50	10	$\approx 2 \times 10^{9}$
100	10	$\approx 1.7 \times 10^{12}$
100	50	$\approx 5 \times 10^{28}$
200	100	$\approx 5 \times 10^{58}$

$$R(a, b) = \begin{cases} a - 1 & \text{pro } b = 2 \\ \sum_{i=1}^{a-b+1} R(a - i, b - 1) & \text{pro } b > 2 \end{cases}$$

Formální popis problému

Vstupní údaje

- vstupní hodnoty: $V = \{v_1, v_2, \ldots, v_n\}$, kde $v_i \in \mathbb{R}$,
- dále předpokládáme, že $v_i < v_{i+1}$ pro $i = 1, \ldots, n-1$,
- pro každou vstupní hodnotu v_i definujeme její četnost f_i,
- nutno rozlišit mezi unikátní vstupní hodnotou v_i a počtem jejích výskytů f_i ve vstupních datech.

Restrikce škály

Rozdělení n vstupních hodnot z množiny V do posloupnosti N nepřekrývajících se intervalů $V_1, \ldots, V_N$.

Definice intervalů restrikce škály

Definice intervalů

- vektor dolních mezí intervalů $\vec{l} \in \mathbb{N}^N$,
- vektor horních mezí intervalů $\vec{u} \in \mathbb{N}^N$

Interval V_i je pak definován:

$$V_i = \left[v_{\vec{l}_i}, v_{\vec{u}_i} \right]$$

Podmínky

$$\begin{aligned} 1 \leq \vec{l}_i &\leq n \\ 1 \leq \vec{u}_i &\leq n \\ \vec{l}_1 &= 1 \\ \vec{u}_N &= n \\ \vec{l}_{j+1} &= \vec{u}_j + 1 \\ \vec{l}_i &\leq \vec{u}_i \\ V &= \bigcup_{i=1}^{N} V_i \end{aligned}$$

Definice jedince a populace

- každá posloupnost intervalů $V_1, V_2, \ldots, V_N$ definuje jednu možnou restrikci škály na původní množině V,
- a stejně tak definuje jednoho jedince v populaci genetického algoritmu,
- populace jedinců se tak skládá z M jedinců $I_1, I_2, \ldots, I_M$,
- každý jedinec I_j je dán svou posloupností intervalů:

$$I_j = (V_1^j, \ldots, V_N^j)$$

Informační entropie vstupních dat

Pravděpodobnost hodnoty v_i dána jako:

$$p_i = \frac{f_i}{F}$$

kde

$$F = \sum_{i=1}^{n} f_i$$

Střední hodnota entropie H_0 vstupní množiny V je dána jako

$$H_0 = -\sum_{i=1}^{n} p_i \log_2 p_i$$

Informační entropie jedince

- entropie jedince $I_j = (V_1^j, \ldots, V_N^j)$ založena na entropii intervalů V_i^j,
- entropie těchto intervalů je založena na pravděpodobnosti výskytů všech hodnot z daného intervalu:

$$f_{V_i^j} = \sum_{k=\vec{I}_i}^{\vec{u}_i} f_k$$

$$p_{V_i^j} = \frac{\sum_{k=\vec{I}_i}^{\vec{u}_i} f_k}{F}$$

- střední entropie jedince $I_j = (V_1^j, \ldots, V_N^j)$ je pak dána jako

$$H_j = -\sum_{i=1}^{N} p_{V_i^j} \log_2 p_{V_i^j}$$

- Fitness funkci $f(I_j)$ jedince I_j definujeme jako:

Experimentální data

- průměrná roční teplota vzduchu v České republice za období 1961 až 2001,
- zaokrouhleno na celé stupně Celsia,
- ESRI grid o velikosti buňky 100 metrů,
- výsledkem je grid o 4865 sloupcích a 2780 řádcích,
- střední hodnota entropie $H_0 = 2,135$ bitů na jednu buňku gridu.

v_i [°C]	f_i [počet pixelů]
1	1968
2	12680
3	50204
4	160566
5	468285
6	1564203
7	3128512
8	2135303
9	382978
10	950

Experimenty s genetickým algoritmem

- restrikce škály dat od 2 do 9 intervalů,
- tomu odpovídá vělikost jedince,
- počet jedinců v populaci $M = 2000$,
- počet generací $G = 1500$
- pro menší počet generací a velikosti populace nedosaženo dobrých výsledků.

Nejlepší restrikce škály nalezené pomocí GA

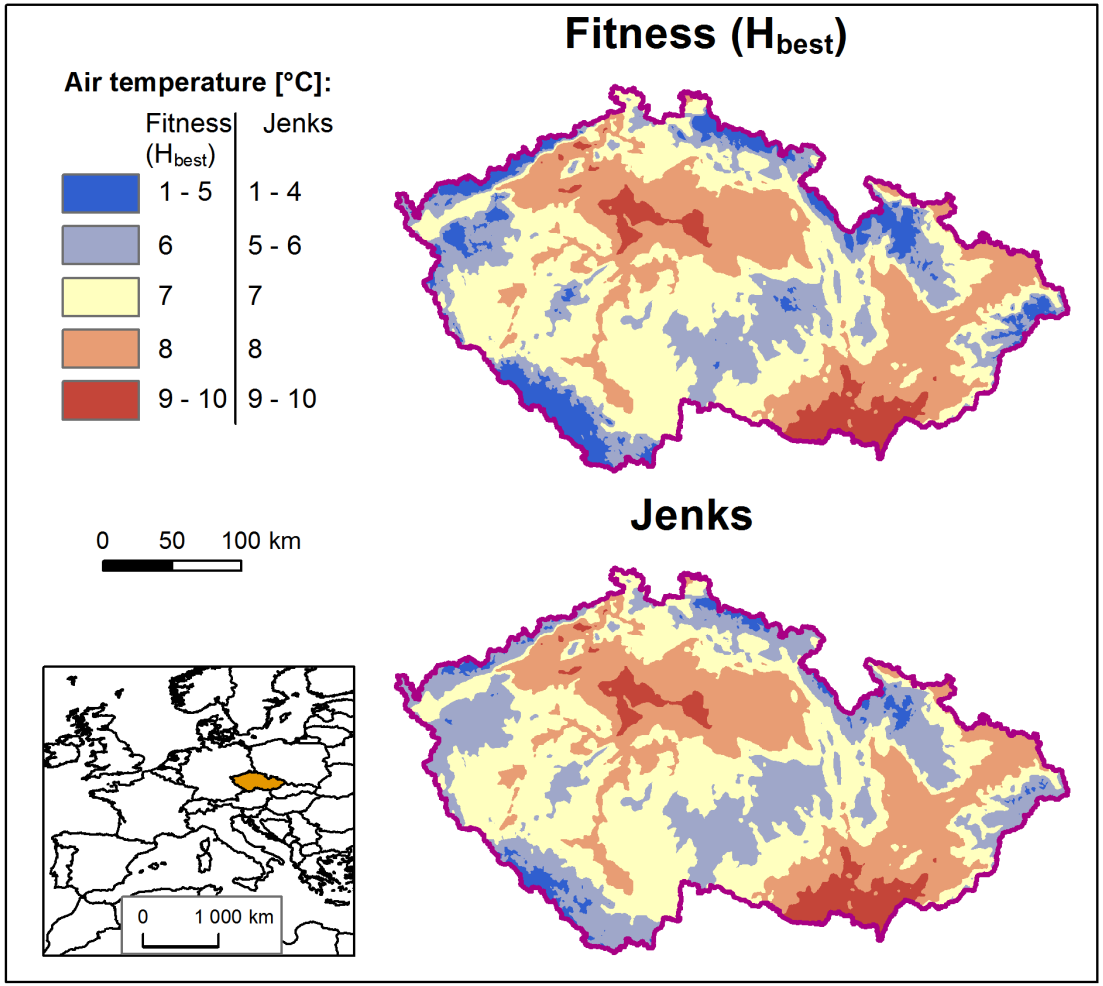
N	I_{Best}	$f(I_{Best})$ [bity]
2	$[1,2,3,4,5,6,7]$ $[8,9,10]$	1,232
3	$[1,2,3,4,5,6]$ $[7]$ $[8,9,10]$	0,564
4	$[1,2,3,4,5]$ $[6]$ $[7]$ $[8,9,10]$	0,309
5	$[1,2,3,4,5]$ $[6]$ $[7]$ $[8]$ $[9,10]$	0,113
6	$[1,2,3,4]$ $[5]$ $[6]$ $[7]$ $[8]$ $[9,10]$	0,033
7	$[1,2,3]$ $[4]$ $[5]$ $[6]$ $[7]$ $[8]$ $[9,10]$	0,009
8	$[1,2]$ $[3]$ $[4]$ $[5]$ $[6]$ $[7]$ $[8]$ $[9,10]$	0,002
9	$[1]$ $[2]$ $[3]$ $[4]$ $[5]$ $[6]$ $[7]$ $[8]$ $[9,10]$	0,001

- I_{Best} označuje nejlepšího jedince v populaci
- $f(I_{Best})$ hodnotu jeho fitness funkce

Výsledky experimentů

- data lze dělit nejvýše do 8 intervalů,
- maximální počet možných restrikcí škály je dosaženo pro 5 intervalů – celkem 126 možností,
- genetický algoritmus a Jenksova metoda srovnána pro 5 intervalů,
- výsledky se **zásadním** způsobem neliší,
- restrikce genetickým algoritmem zachovává formálně více informace,
- genetický algoritmus zdůraznil nejčastější položky a potlačil extrémy – „vypíchnul střed peletonu“,
- Jenksova metoda rozděluje škálu dat rovnoměrněji.

Nejlepší řešení pomocí GA a Jenksovy metody



Závěr

- využití genetických algoritmů pro restrikci škály geodat,
- návrh reprezentace jedince, operátoru křízení a mutace,
- návrh a formalizace fitness funkce založené na minimalizaci rozdílu entropie původních dat a dat restringovaných,
- experimentální výsledky a srovnání s klasickými přístupy.

Děkuji za pozornost

Otázky?