ROBUST'2002, 163 – 177

METODA SEGMENTACE V „CHANGE–POINT“ PROBLÉMU

MARTIN JANŽURA, JAN NIELSEN

Abstract. The method of segmentation is used to solve the change-point problem.

Резюмэ. Метод сегментации применен к решению проблемы детекции разладки.

1. Úvod

Obecně je úloha detekce změn formulována tak, že cílem je zjistit, zda v nějaké časové posloupnosti došlo ke změnám v pravděpodobnostním rozdělení, kterým se řídí pozorované veličiny. Máme tedy náhodné veličiny

$$X_1, \ldots, X_N,$$

kde – v případě jejich nezávislosti – předpokládáme, že veličina X_i má rozdělení p_i pro $i = 1, \ldots, N$. (U závislých veličin bychom museli uvažovat rozdělení podmíněná minulostí, tedy $\mathcal{L}(X_i|X_{i-1}, \ldots, X_1) = p_i(x_i|x_{i-1}, \ldots, x_1)$ pro $i = 1, \ldots, N$). Statistická úloha detekce změny je potom testování hypotézy

$$H_0 : p_1 = \cdots = p_N$$

proti různě obecným alternativám (viz např. Csörgö a Horváth (1997) nebo Antoch, Hušková a Jarušková (1998)).

Je přitom zřejmé, že zcela obecným a úplným řešením by bylo nalezení přímo posloupnosti

$$p_1, \ldots, p_N.$$

Takto obecnou úlohu však pochopitelně nelze řešit. Budeme tedy uvažovat některá zjednodušení. Předně nebudeme předpokládat zcela obecná rozdělení, ale vymezíme pevně danou konečnou množinu přípustných distribucí

$$\mathcal{P} = \{p^{(a)}\}_{a \in \mathcal{A}},$$

kde $\mathcal{A}$ je konečná množina „návěští“ („štítků“, „jmen“). Posloupnost distribucí $p_1, \ldots, p_N$, kde $p_i \in \mathcal{P}$ pro každé $i = 1, \ldots, N$, nám takto přirozeně indukuje posloupnost návěští

$$a^N = (a_1, \ldots, a_N) \in \mathcal{A}^N$$

kde vždy $p_i = p^{(a_i)}$ pro $i = 1, \ldots, N$.

Pro jednoduchost budeme předpokládat $p^{(a)} > 0$ pro všechna $a \in \mathcal{A}$. Můžeme nyní psát

$$\mathcal{L}(X_1, \ldots, X_N|a_1, \ldots, a_N) = \prod_{i=1}^{N} p^{(a_i)}(x_i)$$

2000 *Mathematics Subject Classification.* 62M99 62P12.

Klíčová slova. Segmentace, detekce struktuálních změn statistických modelů, statistický rozhodovací problém, metoda MAP, simulované žíhání, Metropolisnlv-Hastingsv algoritmus.

Tato práce je podporována grantem GA ČR 201/00/1149.

pro nezávislá pozorování (případně $\prod_{i=1}^{N} p^{(a_i)}(x_i|x_{i-1},\dots,x_1)$ pro závislá). Toto rozdělení pak budeme považovat za podmíněné rozdělení vektoru veličin $X_1,\dots,X_N$ za podmínky posloupnosti návěští $a_1,\dots,a_N$.

Navíc budeme předpokládat, že průběh změn v posloupnosti není libovolný, ale je ovlivněn nějakou vnitřní strukturou, a proto všechny posloupnosti nejsou nutně stejně možné, ale existuje nám známé apriorní rozdělení

$$\{P(a^{(N)})\}_{a^N\in\mathcal{A}^{(N)}},$$

kde opět pro jednoduchost $P>0$.

Tímto se přímo nabízí řešit úlohu nalezení posloupnosti $a^N=(a_1,\dots,a_N)$ na základě dat $x^N=(x_1,\dots,x_N)$, vzniklých pozorováním posloupnosti $X_1,\dots,X_N$, pomocí bayesovského přístupu. Vzhledem k tomu, že v typických situacích, které budeme mít na mysli, nebudou změny příliš časté a tudíž se posloupnosti návěští budou podle očekávání skládat z homogenních úseků (segmentů), metoda zvolená v tomto článku se podobá metodám segmentace (či klasifikace) známé např. v oblasti zpracování obrazu (viz Winkler (1995) pro přehled).

V následující části navrhneme řešení v rámci obecného bayesovského rozhodovacího problému.

2. Bayesovské řešení statistického rozhodovacího problému

Statistický rozhodovací problém je dán čtveřicí (viz např. Vajda (1987))

$$\{\mathcal{Y},\mathcal{Q},\mathcal{D},\mathcal{Z}\},$$

kde $\mathcal{Y}$ je stavový prostor, $\mathcal{Q}=\{q^\theta\}_{\theta\in\Theta}$ je parametrická rodina pravděpodobnostních rozdělení, $\mathcal{D}$ je množina možných „rozhodnutí“ a

$$\mathcal{Z}=\Theta\times\mathcal{D}\to[0,\infty)$$

je ztrátová funkce, kde $\mathcal{Z}(\theta,d)$ vyjadřuje ztrátu, která nastane při rozhodnutí d, jestliže skutečná hodnota parametru je $\theta\in\Theta$.

Pro volbu „dobrého“ rozhodnutí pak máme k dispozici pozorování $y\in\mathcal{Y}$, přičemž tuto volbu popisuje rozhodovací funkce

$$\delta:\mathcal{Y}\to\mathcal{D}.$$

Kvalitu rozhodovací funkce posuzujeme podle ztráty

$$\mathcal{Z}(\theta,\delta)=\int\mathcal{Z}(\theta,\delta(y))\,\mathrm{d}q^\theta(y)$$

spojené s touto rozhodovací funkcí. (Necháme v těchto úvahách stranou otázku měřitelnoti jednotlivých zobrazení. Můžeme např. pro jednoduchost považovat všechny množiny za konečné.)

Cílem úlohy je pak nalézt takovou rozhodovací funkci, která ztrátu v nějakém smyslu minimalizuje.

Pokud pro nějakou rozhodovací funkci δ_s je

$$\mathcal{Z}(\theta,\delta_s)=\min_\delta\mathcal{Z}(\theta,\delta)\quad\text{pro každé }\theta\in\Theta,$$

je tato δ_s stejnoměrně nejlepším řešením. Řešení minimaxní je dáno podmínkou

$$\max_{\theta\in\Theta}\mathcal{Z}(\theta,\delta_m)=\min_\delta\max_{\theta\in\Theta}\mathcal{Z}(\theta,\delta).$$

My zde budeme předpokládat, že máme na prostoru Θ nějakou apriorní distribuci $Q(\theta)$. Potom je bayesovským řešením takové rozhodovací funkce δ_b, která minimalizuje očekávanou ztrátu

$$\mathcal{Z}(\delta) = \int \mathcal{Z}(\theta, \delta)\, \mathrm{d}Q(\theta),$$

tedy

$$\mathcal{Z}(\delta_b) = \min_{\delta} \mathcal{Z}(\delta).$$

Označme nyní

$$\mathcal{Z}(d|y) = \int \mathcal{Z}(\theta, d)\, \mathrm{d}Q(\theta|y)$$

(kde přirozeně $Q(\theta|y) = \frac{Q(\theta;y)}{Q(y)} = \frac{q^{\theta}(y)\, Q(\theta)}{\int q^{\tau}(y)\, \mathrm{d}Q(\tau)}$).

Tvrzení: Bayesovské řešení δ_b rozhodovacího procesu je dáno předpisem

$$\mathcal{Z}(\delta_b(y)|y) = \min_{d \in \mathcal{D}} \mathcal{Z}(d|y).$$

Důkaz: Nechť δ je nějaká jiná rozhodovací funkce. Potom

$$\mathcal{Z}(\delta_b) = \int \mathcal{Z}(\delta_b(y)|y)\, \mathrm{d}Q(y) \leq \int \mathcal{Z}(\delta(y)|y)\, \mathrm{d}Q(y) = \mathcal{Z}(\delta).$$

Uvažujme nyní speciální případ, kdy máme „rozhodnout“ o parametru, tedy jej odhadnout. Potom $\Theta = \mathcal{D}$. Navíc se omezíme na speciální ztrátovou funkci

$$\mathcal{Z}(\theta, d) = \begin{cases} 0 & \text{pokud } \theta = d \\ 1 & \text{jinak.} \end{cases}$$

Potom máme

$$\mathcal{Z}(d|y) = 1 - Q(d|y)$$

a proto je (podle tvrzení nahoře) bayesovské řešení $\delta_b = \hat{\theta}$ dáno

$$Q(\hat{\theta}(y)|y) = \max_{\theta \in \Theta} Q(\theta|y).$$

Je to tedy odhad maximalizující aposteriorní pravděpodobnost (MAP), který je bayesovským řešením, neboť minimalizuje očekávanou ztrátu, danou zde pravděpodobností chyby

$$\mathcal{Z}(\hat{\theta}) = Q(\theta \neq \hat{\theta}(y)).$$

3. Řešení původní úlohy metodou MAP

Budeme řešit úlohu zformulovanou v části 1 metodou MAP, tzn. pro daná data $x^N = (x_1, \ldots, x_N)$ budeme hledat takovou posloupnost $\hat{a}^N$, která maximalizuje aposteriorní pravděpodobnost, tedy

$$\begin{aligned} \hat{a}^N \in \operatorname{argmax}_{a^N \in \mathcal{A}^N} P(a^N|x^N) &= \operatorname{argmax}_{a^N \in \mathcal{A}^N} P(a^N; x^N) = \\ &= \operatorname{argmax}_{a^N \in \mathcal{A}^N} \log P(a^N; x^N) \end{aligned}$$

kde

$$P(a^N; x^N) = \prod_{i=1}^{N} p^{(a_i)}(x_i) \cdot P(a^N).$$

Pro tuto úlohu diskrétní optimalizace však pro netriviální rozdělení $P(a^N)$ a netriviální dimenzi dat N není k dispozici účinný deterministický algoritmus. Pro numerické řešení použijeme tedy algoritmus stochastický, založený na znáhodněném prohledávání množiny $\mathcal{A}^N$ a známý pod jménem „simulované žíhání“ (viz např.

Winkler (1995)) nebo Janžura (1990)). Ten je založený na následujících úvahách. Nechť máme nalézt maximum funkce $F : \mathcal{A}^N \to R$.

Označme $Q^\tau(a^N) = \frac{e^{\tau F(a^N)}}{c(F,\tau)}$ pro reálné τ, kde $c(F,\tau) = \sum_{b^N \in \mathcal{A}^N} e^{\tau F(b^N)}$ je normalizační konstanta. Zřejmě nyní platí

$$Q^\tau \longrightarrow Q^\infty \quad \text{pro } \tau \to \infty,$$

kde

$$Q^\infty(a^N) = \begin{cases} \frac{1}{|M|} & \text{pro } a^N \in M = \operatorname{argmax} F \\ 0 & \text{jinak.} \end{cases}$$

Symbolem $|M|$ označíme kardinalitu množiny. Můžeme nyní vyvozovat, že nalézt $a^N \in M$ znamená to samé jako simulovat s rozdělením Q^∞, což je „skoro jako“ simulovat s rozdělením Q^τ pro dostatečně veliké $\tau > 0$. To však stále ještě neumíme, neboť nejsme schopni vyčíslit normalizační konstantu $c(F,\tau)$ pro příslušnou velikost množiny $\mathcal{A}^N$. Musíme tedy pokročit dále. Z teorie Markovových řetězců víme, že pokud je

$$R(\tau) = (R(\tau)_{a^N b^N})_{a^N, b^N \in \mathcal{A}^N}$$

nerozložitelná stochastická matice, která splňuje

$$Q^\tau \, R(\tau) = Q^\tau$$

(kde Q^τ chápeme jako řádkový vektor), platí

$$[R(\tau)]^n_{a^N_{(0)}\bullet} \longrightarrow Q^\tau(\bullet) \quad \text{pro } n \to \infty$$

při libovolném počátečním $a^N_{(0)} \in \mathcal{A}^N$.

Odtud můžeme opět vyvodit, že simulovat s rozdělením Q^τ je „skoro totéž“ jako simulovat s rozdělením $[R(\tau)]^n_{a^N_{(0)}\bullet}$ při nějakém dostatečně velkém n. A podstata simulovaného žíhání je ve spojení obou těchto limitních vlastností.

Věta: Nechť $\tau_n \le \Delta(F) \log n$, potom

$$[R(\tau_1) \cdot R(\tau_2) \cdot \ldots \cdot R(\tau_n)]_{a^n_{(0)}\bullet} \longrightarrow Q^\infty(\bullet) \quad \text{pro } n \to \infty.$$

Důkaz: Např. Winkler (1995), Věta 5.2.1.

Algoritmus nyní můžeme popsat takto:

(1) zvolíme $a^N_{(0)}$

(2) pro $j = 1, \ldots, n_{\text{STOP}}$ vybereme $a^N_{(j)}$ s rozdělením $R(\tau_j)_{a^N_{(j-1)}\bullet}$

(3) ponecháme $a^N_{(n_{\text{STOP}})}$.

Takto získané $a^N_{n_{\text{STOP}}}$ pro dostatečně velké n_{STOP} považujeme za vybrané „skoro jako“ s rozdělením Q^∞ a tudíž za řešení úlohy.

Jak ale zvolit stochastickou matici $R(\tau)$? Jeden z návodů poskytuje metoda Metropolise–Hastingse:

Zvolme nerozložitelnou stochastickou matici $\tilde{R}$ a definujme

$$\begin{aligned} R(\tau)_{a^N b^N} &= \tilde{R}_{a^N b^N} \cdot \min\left(1, \frac{Q^\tau(b^N)\,\tilde{R}(\tau)_{b^N a^N}}{Q^\tau(a^N)\,\tilde{R}(\tau)_{a^N b^N}}\right) \quad \text{pro } b^N \neq a^N \\ R(\tau)_{a^N a^N} &= 1 - \sum_{b^N \neq a^N} R(\tau)_{a^N b^N}. \end{aligned}$$

Jelikož $0 \le R(\tau)_{a^N b^N} \le \tilde{R}_{a^N b^N}$, máme také $\sum_{b^N \neq a^N} R(\tau)_{a^N b^N} \le 1$ a proto je matice $R(\tau)$ stochastická.

Abychom zaručili nerozložitelnost, budeme navíc předpokládat, že

$$\tilde{R}_{a^N b^N} = 0 \quad \text{právě když} \quad \tilde{R}_{b^N a^N} = 0.$$

Potom z $R(\tau)_{a^N b^N} = 0$ plyne $\tilde{R}_{a^N b^N} = 0$ a nerozložitelnost $R(\tau)$ plyne z nerozložitelnosti $\tilde{R}$. Tento dodatečný předpoklad je snadno splněn např. pro symetrické matice $\tilde{R}$.

Nakonec ověříme invarianci. Můžeme psát

$$\sum_{a^N} Q^\tau(a^N)\, R(\tau)_{a^N b^N} = \sum_{a^N \neq b^N} \min\left(Q^\tau(a^N)\, \tilde{R}_{a^N b^N},\, Q^\tau(b^N)\, \tilde{R}_{b^N a^N}\right) +$$

$$+\, Q^\tau(b^N) \left(1 - \sum_{a^N \neq b^N} R(\tau)_{b^N a^N}\right) = Q^\tau(b^N)$$

neboť také

$$Q^\tau(b^N)\, R(\tau)_{b^N a^N} = \min\left(Q^\tau(b^N)\, \tilde{R}_{b^N a^N},\, Q^\tau(a^N)\, \tilde{R}_{a^N b^N}\right).$$

Tím jsme ukázali, že tato matice má potřebné vlastnosti a přitom na rozdělení Q^τ závisí pouze prostřednictvím podílů

$$\frac{Q^\tau(b^N)}{Q^\tau(a^N)} = e^{\tau(F(b^N) - F(a^N))},$$

které již neobsahují problémovou normalizační konstantu $c(F, \tau)$ a proto mohou být vyčísleny.

Stochastickou matici $\tilde{R}$ volíme s ohledem na jednoduchost simulace, zpravidla „nezávisle“ a „rovnoměrně“, tj.

$$\tilde{R}_{a^N b^N} = \frac{1}{|\mathcal{A}^N|} \quad \text{pro všechna } a^N,\, b^N \in \mathcal{A}^N.$$

Nebo také můžeme v každém kroku měnit náhodně pouze jednu souřadnici, potom máme

$\tilde{R}_{a^N b^N} = \frac{1}{N \cdot |\mathcal{A}|}$ pokud existuje $t \in \{1, \ldots, N\}$ tak, že $a_s = b_s$ pro každé $s \neq t$,

$\tilde{R}_{a^N b^N} = 0$ jinak.

(Zde nejprve náhodně vybereme souřadnici t s pravděpodobností $\frac{1}{N}$ a novou hodnotu b_t vybereme s pravděpodobností $\frac{1}{|\mathcal{A}|}$.)

Máme-li tedy zvolenu matici $\tilde{R}$, probíhá j-tý krok simulace ve dvou fázích:

1.: návrh: simulujeme $\tilde{a}^N$ s rozdělením $\tilde{R}_{a^N_{(j-1)}}\bullet$

2.: přijetí: pokud

$$\Delta_j = \frac{Q^{\tau_j}(\tilde{a}^N)\, \tilde{R}_{\tilde{a}^N a^N_{(j-1)}}}{Q^{\tau_j}(a^N_{(j-1)})\, \tilde{R}_{a^N_{(j-1)} \tilde{a}^N}} \geq 1,$$

dosadíme přímo $a^N_j := \tilde{a}^N$.

Pokud $\Delta_j < 1$, učiníme další nezávislou simulaci a dosadíme $a^N_{(j)} := \tilde{a}^N$ s pravděpodobností Δ_j nebo ponecháme $a^N_{(j)} := a^N_{(j-1)}$ s pravděpodobností $1 - \Delta_j$.

Poznámka: Často se postupuje také tím způsobem, že konfiguraci $a^N \in \mathcal{A}^N$ pozměňujeme postupně „po složkách“ (po jedné nebo několika málo) v nějakém deterministickém pořadí. To znamená, že pro každé $j = 1, \dots, N$ máme $S_j \subset \{1, \dots, N\}$ tak, že

$$\tilde{R}^{(j)}_{a^N b^N} = 0 \quad \text{pokud } b_s \neq a_s \text{ pro nějaké } s \notin S_j.$$

Tedy opět např.

$$\begin{aligned} \tilde{R}^{(j)}_{a^N b^N} &= \tfrac{1}{|\mathcal{A}^{S_j}|} \quad \text{pokud } a_s = b_s \text{ pro každé } s \notin S_j. \\ &= 0 \qquad \text{jinak.} \end{aligned}$$

Tím ovšem každá jednotlivá „návrhová“ matice $\tilde{R}^{(j)}$ (a tudíž i od ní odvozená $R(\tau)^{(j)}$) není nerozložitelná a musíme tedy brát

$$R(\tau) = R(\tau)^{(1)} \cdots R(\tau)^{(N)}.$$

Pokud přitom platí $\bigcup_{j=1}^N S_j = \{1, \dots, N\}$, je už tato součinová matice nerozložitelná. Tento postup má tu výhodu, že je každý jeho dílčí krok velmi rychle a snadno realizovatelný.

4. Implementace a příklady

V naší úloze jsme použili Metropolisův–Hastingsův algoritmus v modifikaci popsané v poznámce v předchozí části, kdy jsme brali $S_j = \{j-1, j, j+1\}$ pro $j = 2, \dots, N-1$. Abychom eliminovali směrový efekt při simulaci, měnili jsme periodicky směr času, brali jsme tedy vlastně

$$R(\tau) = R(\tau)^{(1)} \cdots R(\tau)^{(N)} \cdot R(\tau)^{(N)} \cdots R(\tau)^{(1)}.$$

Velmi podstatným parametrem úlohy je volba funkce τ_n (nazývaná v terminologii simulovaného žíhání jako „ochlazovací plán“.) Vzhledem ke své pomalé rychlosti konvergence je logaritmus všeobecně považován pro reálně dosažitelné n_{STOP} za příliš pomalý (viz např. Winkler (1995), Kapitola 6). Jde zde však třeba postupovat velmi opatrně, neboť volba příliš rychlé konvergence zavádí metodu okamžitě do nějakého lokálního extrému blízko počáteční konfigurace. Po určitých numerických experimentech se v našem případě osvědčilo brát $\tau_n = (1 + 10^{-\ell})^n$, kde můžeme zvolit $\ell \in \{3, 4, 5\}$. Potom zpravidla postačuje $n_{\text{STOP}} = 10^5$.

Metodu jsme aplikovali na simulovaná i reálná data. Vzhledem k tomu, že reálná data se týkala ročních průměrných teplot v Praze, pracovali jsme s podobným oborem hodnot i pro simulovaná data. Zvolili jsme

$$\mathcal{A} = \{8, 9, 10, 11\}$$

a

$$P(a^N) \propto \exp\left\{-\alpha \sum_{i=2}^{N} |a_i - a_{i-1}|\right\},$$

což znamená, že apriorní rozdělení je markovské a pro $\alpha > 0$ jsou preferovány blízké (nejvíce přímo totožné) hodnoty v sousedících bodech. To odpovídá méně četným a méně významným změnám. Dále pro každé $i = 1, \dots, N$ bylo

$$p_i^{(a)}(x_i) = \frac{1}{\sqrt{2\pi}\sigma}\, e^{-\frac{1}{2\sigma^2}(x_i - a)^2},$$

takže jsme měli data s nezávislým gaussovským šumem. Sdružené rozdělení $\mathcal{L}(X_1, \dots$ $\dots, X_N)$ je tedy dáno tzv. skrytým markovským modelem (viz např. MacDonald

a Zucchini (1997) pro podrobnou studii). Simulovaná data vznikla tak, že jsme nejprve simulovali posloupnost a^N a potom ji „zašuměli“.

Pro daná data $\hat{x_1}, \ldots, \hat{x}_N$ vede metoda MAP na úlohu

$$\min_{a^N} \left[\sum_{i=1}^{N} (\hat{x}_i - a_i)^2 + 2\gamma \sum_{i=2}^{N} |a_i - a_{i-1}| \right]$$

kde $\gamma = \sigma^2 \cdot \alpha$ je nyní jediným parametrem úlohy, vyjadřujícím relaci mezi „důvěryhodností“ dat a „důvěryhodností“ informace obsažené v sousedních hodnotách. (Např. pokud je malý rozptyl v datech a malé závislosti mezi sousedy, dáváme větší váhu informaci v datech.)

Za počáteční konfiguraci jsme volili vždy postupně dvě krajní možnosti, a to jednak konfiguraci získanou „z dat“ (tj. vyřešením úlohy při $\gamma = 0$), a potom konfiguraci konstantní (tu můžeme považovat za řešení při $\gamma = \infty$). Pokud se při takto významně odlišných počátečních hodnotách řešení nelišila, mohli jsme to považovat za známku spolehlivosti výsledku.

Samozřejmě, že obecně bychom neznali ani „správnou“ množinu $\mathcal{A}$ a parametr γ (když už bychom připustili daný typ rozdělení). Proto jsme experimenty prováděli při různých volbách těchto parametrů.

Reálnými daty byly průměrné roční teploty naměřené v pražském Klementinu od roku 1775 (autoři děkují J. Antochovi za laskavé zapůjčení). Aplikovali jsme na ně stejný model jako na data simulovaná.

5. Výsledky

Výsledky uvádíme formou obrázků, kde slabá čára vždy značí průběh dat, silná výslednou konfiguraci a pro simulovaná data pak středně silná skoková čára znázorňuje průběh „správné“ konfigurace. Vzhledem k tomu, že jsme měli na mysli detekci „změn“, volili jsme tuto konfiguraci se skokovým průběhem.

Výsledky uvádíme pro troje různá data. Nejprve (Obr. 1 – Obr. 6) pro data simulovaná s rozptylem $\sigma^2 = 0,5$ (při $\alpha = 5$, $\mathcal{A} = \{8, 9, 10, 11\}$), další série (Obr. 7 – Obr. 12) obsahuje výsledky pro jiná data při větším rozptylu $\sigma^2 = 1$ (stejné α a $\mathcal{A}$) a poslední série (Obr. 13 – Obr. 18) se týká reálných dat z Klementina. Pro každá data uvádíme řešení pro všechny kombinace volby parametru $\gamma = 1$ nebo $\gamma = 5$ a množiny $\mathcal{A} = [8, 11]$ s přesností 0,1 („Krok $= 0,1$“), $\mathcal{A} = \{8; 8,5; 9; 9,5; 10; 10,5; 11\}$ („Krok $= 0,5$“) a $\mathcal{A} = \{8; 9; 10; 11\}$ (Krok $= 1$).

6. Závěry

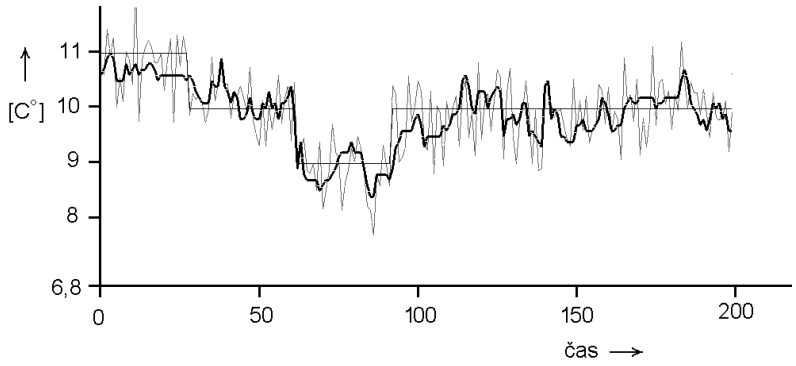
(1) Správná volba parametru γ je důležitá. Při malém $\gamma = 1$ řešení obvykle příliš sleduje data (s výjimkou na Obr. 5, kde při správně zvolené množině $\mathcal{A}$ je i takto malý parametr γ postačující, vedoucí ke stejnému řešení jako $\gamma = 5$ na Obr. 6). Proto pro dobré výsledky bylo třeba zpravidla použít $\gamma = 5$.

(2) Volba stavové množiny $\mathcal{A}$ (čili „Kroku“) také ovlivňuje řešení podstatně. Vzhledem k principu metody, která preferuje stejné nebo blízké hodnoty, dochází při větší stavové množině (tj. jemnější škále) k „vyhlazení“ skokové funkce tvořící správnou konfiguraci.

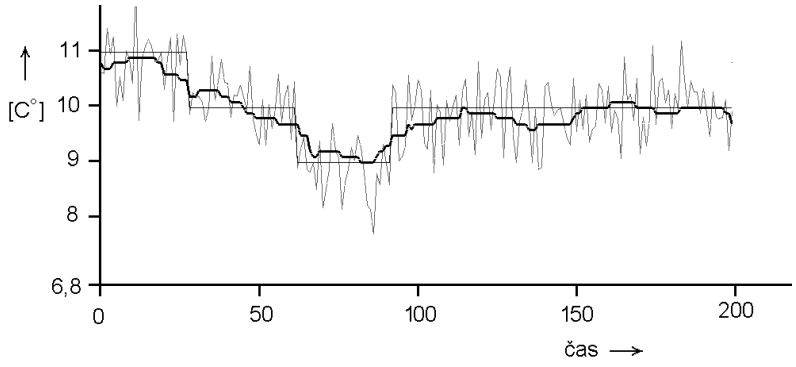
(3) Při správně zvolených parametrech funguje metoda dostatečně spolehlivě.

(4) Výsledky pro klementinská data nejsou v rozporu s očekáváním a výstupy z jiných zpracování a poskytují zajímavé náhledy a interpretační možnosti.

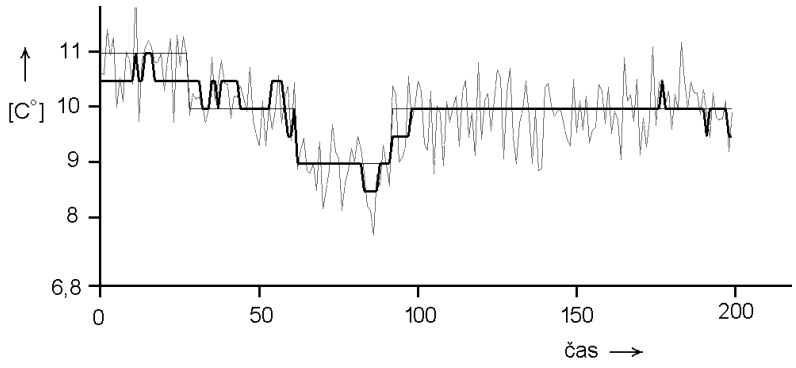
(5) Celkově je možné říci, že zejména pro reálná data poskytuje metoda i nástroj pro generování hypotéz, které mohou být dále zkoumány jinými postupy.



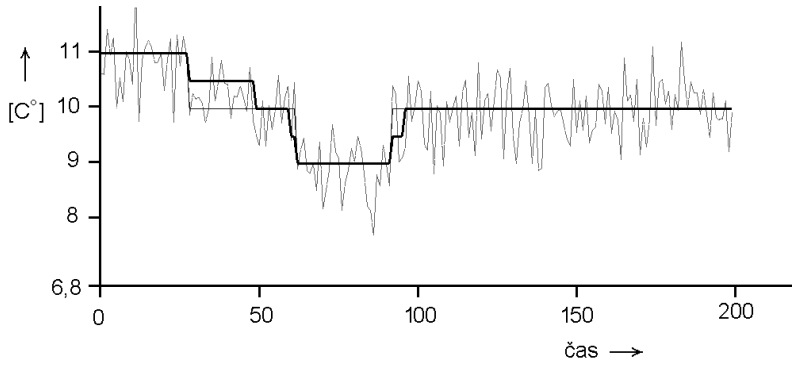
Obr. 1 $Rozptyl = 0,5,\ Gamma = 1,\ Krok = 0,1.$



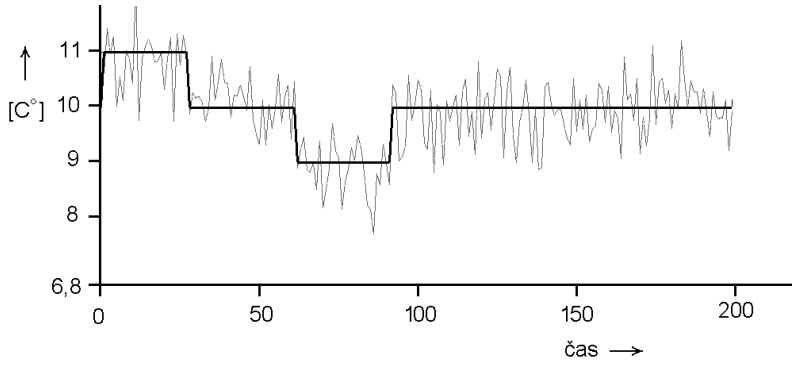
Obr. 2 $Rozptyl = 0,5,\ Gamma = 5,\ Krok = 0,1.$



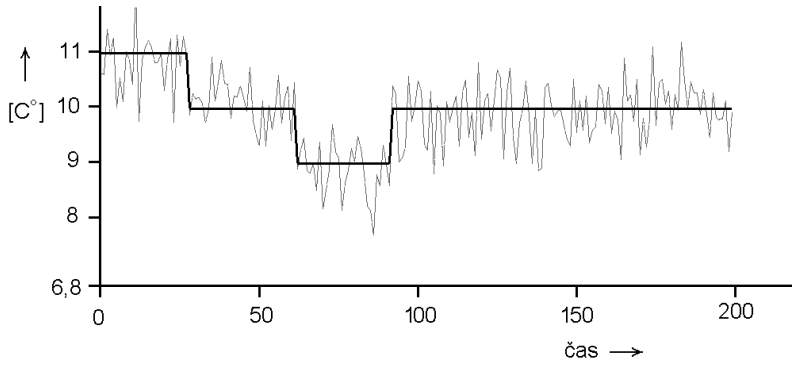
Obr. 3 $Rozptyl = 0,5,\ Gamma = 1,\ Krok = 0,5.$



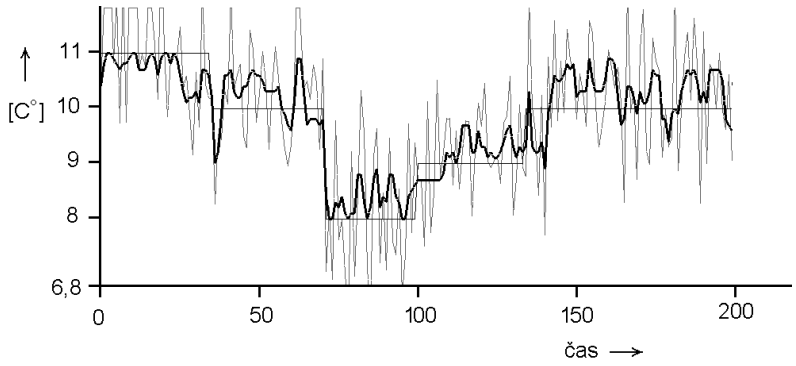
Obr. 4 *Rozptyl* $= 0,5$, *Gamma* $= 5$, *Krok* $= 0,5$.



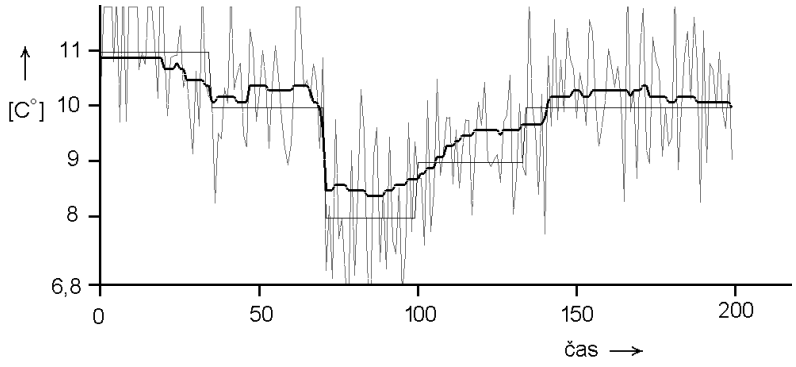
Obr. 5 *Rozptyl* $= 0,5$, *Gamma* $= 1$, *Krok* $= 1,0$.



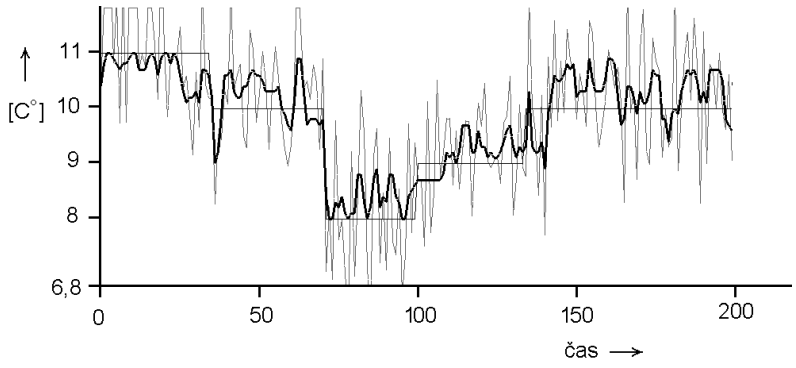
Obr. 6 *Rozptyl* $= 0,5$, *Gamma* $= 5$, *Krok* $= 1,0$.



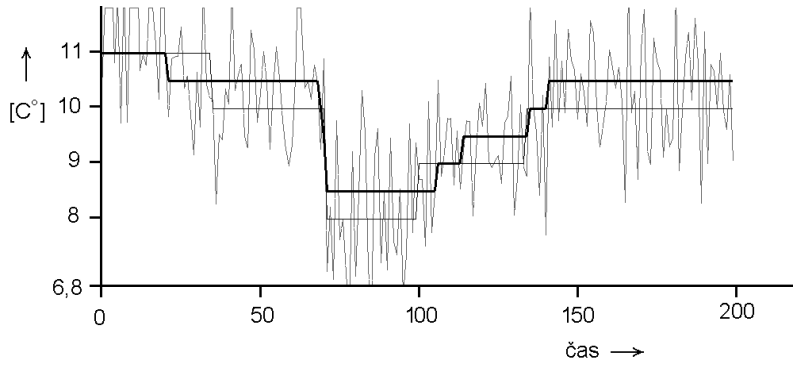
Obr. 7 *Rozptyl* $= 1,0$, *Gamma* $= 1$, *Krok* $= 0,1$.



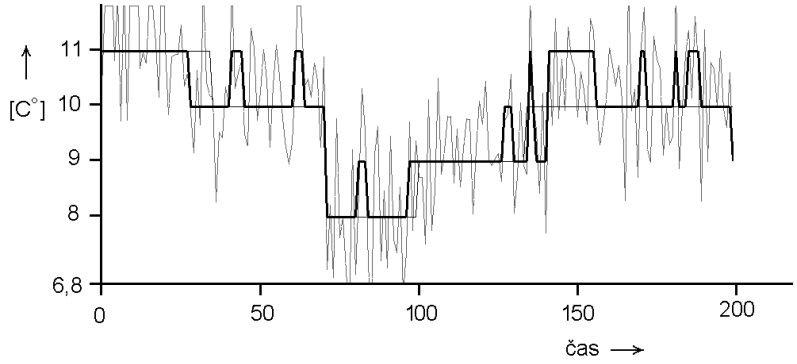
Obr. 8 *Rozptyl* $= 1,0$, *Gamma* $= 5$, *Krok* $= 0,1$.



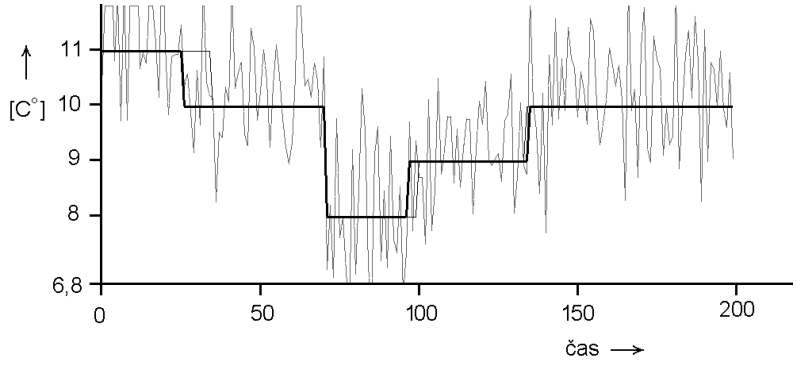
Obr. 9 *Rozptyl* $= 1,0$, *Gamma* $= 1$, *Krok* $= 0,5$.



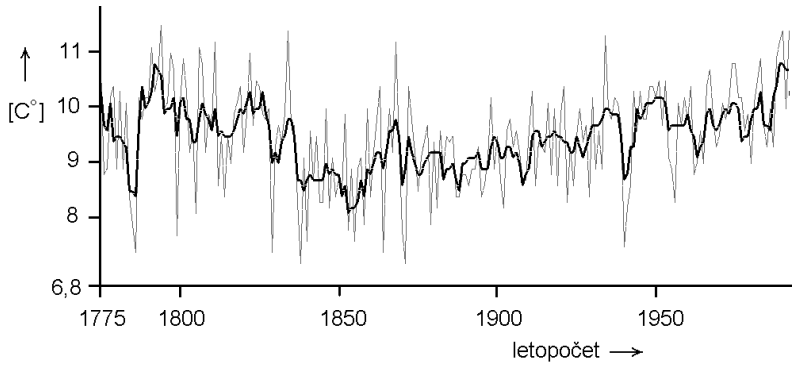
Obr. 10 *Rozptyl* $= 1,0$, *Gamma* $= 5$, *Krok* $= 0,5$.



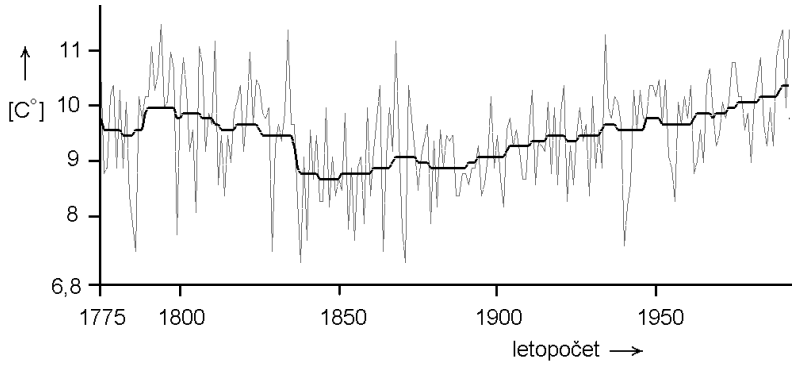
Obr. 11 *Rozptyl* $= 1,0$, *Gamma* $= 1$, *Krok* $= 1,0$.



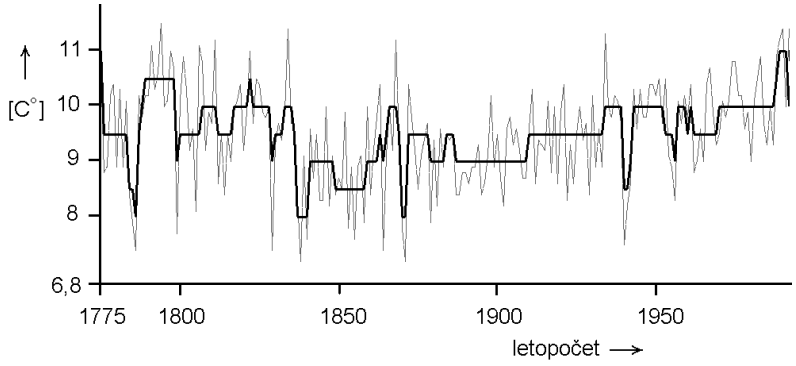
Obr. 12 *Rozptyl* $= 1,0$, *Gamma* $= 5$, *Krok* $= 1,0$.



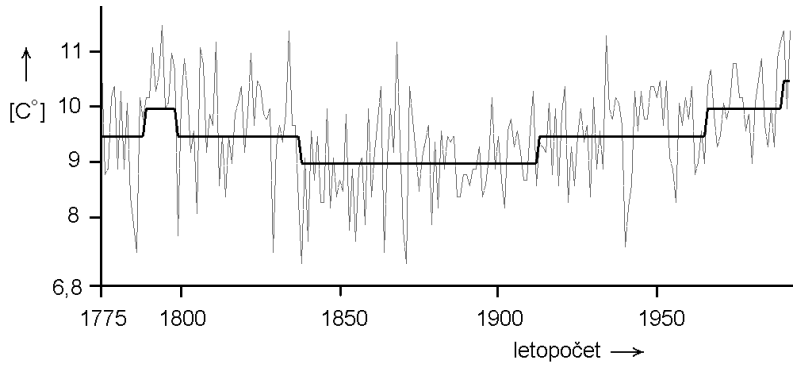
Obr. 13 *Klementinum, Gamma* $= 1$*, Krok* $= 0,1$.



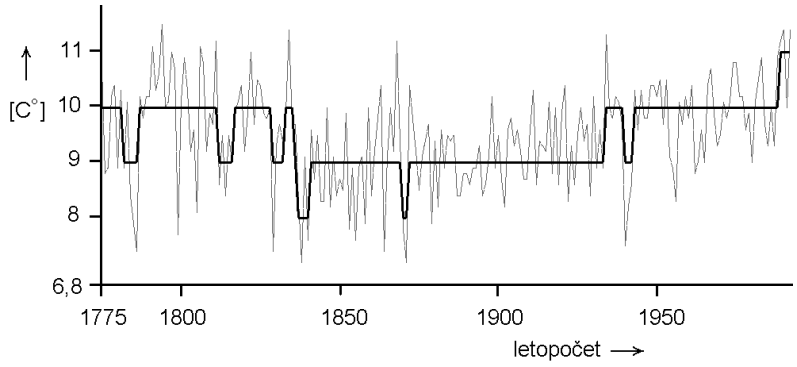
Obr. 14 *Klementinum, Gamma* $= 5$*, Krok* $= 0,1$.



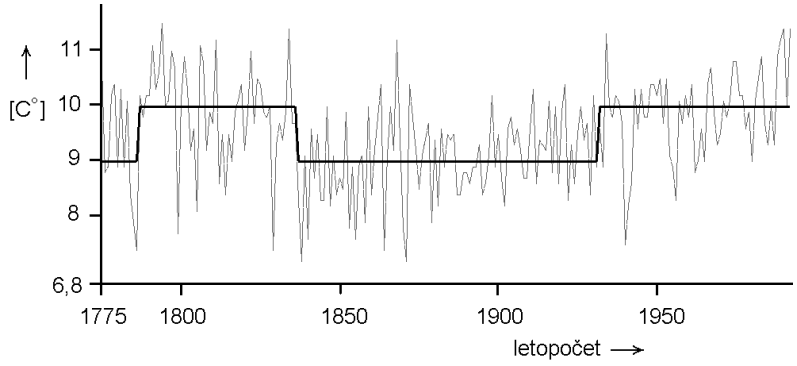
Obr. 15 *Klementinum, Gamma* $= 1$*, Krok* $= 0,5$.



Obr. 16 *Klementinum, Gamma* $= 5$*, Krok* $= 0,5$.



Obr. 17 *Klementinum, Gamma* $= 1$*, Krok* $= 1,0$.



Obr. 18 *Klementinum, Gamma* $= 5$*, Krok* $= 1,0$.

7. Další zobecnění

V předchozích částech jsme uvažovali pouze nezávislé veličiny lišící se střední hodnotou. Nyní si ukážeme, že metoda je dostatečně obecná a že záleží jen na naší schopnosti zformulovat řešený problém v jejím rámci.

i) Trend
Uvažujme systém, který se může nacházet ve svou stavech, a to ve stavu stagnace nebo růstu. Stagnaci v čase $i \in \{1, \dots, N\}$ vyjádříme tak, že $EX_i = EX_{i-1}$, a růst tak, že $EX_i = EX_{i-1} + c$, kde $c > 0$ je známá konstanta. Předpokládáme opět gaussovský šum a markovskou apriorní distribuci. Metoda MAP pak vede na

$$\min_{a^N} \left[\sum_{i=1}^{N} \left(\hat{x}_i - c \sum_{j=1}^{i} a_j \right)^2 + \gamma \sum_{i=2}^{N} |a_i - a_{i-1}| \right]$$

kde $a^N \in \{0,1\}^N$ ($a_i = 0$ znamená stagnaci a $a_i = 1$ růst).

ii) AR procesy
Nyní předpokládejme, že pozorované veličiny tvoří gaussovskou autoregresní posloupnost, jejíž parametry se mohou měnit. Máme dán seznam modelů

$$\left(\mu^{(a)}, \{c_k^{(a)}\}_{k=1}^{K} \right)_{a \in \mathcal{A}}$$

kde $\mu^{(a)}$ jsou střední hodnoty a $\left\{ c_k^{(a)} \right\}_{k=1}^{K}$ autoregresní koeficienty. Máme-li data $\hat{x}_{-k+1}, \dots, \hat{x}_N$, řešíme úlohu

$$\min_{a^N} \left[\sum_{i=1}^{N} \left(\hat{x}_i - \mu^{(a_i)} - \sum_{k=1}^{K} c_k^{(a_i)} \hat{x}_{i-k} \right)^2 + \gamma \cdot \sum_{i=2}^{N} \mathrm{DIST}(a_i, a_{i-1}) \right]$$

kde $a^N \in \mathcal{A}^N$ a $\mathrm{DIST}(\cdot,\cdot)$ je relace odvozená od apriorní distribuce.

V těchto úlohách jsme stále pracovali se „známou“ množinou $\mathcal{A}$. Jednalo se tedy, striktně vzato, o klasifikaci a ne o segmentaci.

Z předpokladu gaussovského šumu a markovské apriorní distribuce můžeme však odvodit i účelovou funkci typu

$$\sum_{i=2}^{N} \left[(x_i - x_{i-1})^2 \, \delta(a_i, a_{i-1}) + \gamma \cdot (1 - \delta(a_i, a_{i-1})) \right],$$

kde

$$\delta(a,b) = \begin{cases} 1 & \text{pro } a = b, \\ 0 & \text{jinak.} \end{cases}$$

Zde potom volíme pouze velikost množiny $\mathcal{A}$, na konkrétních hodnotách jejich elementů nezáleží. Pak se jedná o pravou „nesupervizovanou“ segmentaci.

Literatura

[1] Antoch, J., Hušková, M. a Jarušková D. (1998): Change point problem po deseti letech. Sborník Robust'98, JČMF Praha, 1–42.
[2] Csörgö, M. a Horváth, L. (1997): Limit Theorems in Change–Point Analysis. Wiley.
[3] Janžura, M. (1990): O jednom pravděpodobnostním algoritmu pro optimalizační úlohy. Sborník Robust'90, JČMF Praha, 84–88.

[4] MacDonald, I. L. a Zucchicni, W. (1997): Hidden Markov and Other Models for Discrete–valued Times Series. Chapman & Hall.
[5] Vajda, I. (1987): Theory of Information and Statistical Decision (in Slovak). Alfa, Bratislava.
[6] Winkler, G. (1995): Image Analysis, Random Fields and Dynamic Monte Carlo Methods. Springer.

Ústav teorie informace a automatizace AV ČR, Pod Vodárenskou věží 4, 182 08 Praha 8