



**Vilniaus Universiteto
Duomenų mokslo ir skaitmeninių technologijų institutas**



DAUGIAMAČIŲ PASLĖPTŲJŲ MARKOVO MODELIŲ TYRIMAS IR TAIKYMAS REALAUS LAIKO SISTEMŲ ANALIZEI

**INFORMATIKOS STUDIJŲ PROGRAMOS
DOKTORANTŪROS METINĖ ATASKAITA**

DOKTORANTŪROS LAIKOTARPIS: 2015 – 2019 m.

DOKTORANTĖ: JŪRATĖ VAIČIULYTĖ

DARBO VADOVAS: PROF. HABIL. DR. LEONIDAS SAKALAUSKAS

Tyrimų objektas

Rekursiniai algoritmai, skirti paslėptųjų Markovo modelių (PMM) daugiamačių parametrų vertinimui.

Darbo tikslas

Paslėptųjų Markovo modelių daugiamačių parametrų vertinimo rekursinių algoritmų tyrimas ir taikymas atsitiktinių sekų realaus laiko analizėje.

Darbo uždaviniai

1. Analitiškai apžvelgti paslėptųjų Markovo modelių parametrų vertinimo metodus, taikomus atsitiktinių sekų realaus laiko analizėje.
2. Sukurti rekursinius paslėptųjų Markovo modelių daugiamačių parametrų vertinimo algoritmus;
3. Sukurtus rekursinius paslėptųjų Markovo modelių daugiamačių parametrų vertinimo algoritmus ištirti statistinio modeliavimo būdu ir palyginti su kitais PMM parametrų vertinimo algoritmais.
4. Sukurtus rekursinius paslėptųjų Markovo modelių daugiamačių parametrų vertinimo metodus pritaikyti atsitiktinių sekų realaus laiko analizėje.

Tyrimų sritis ir problemos aktualumas (1)

Pastaraisiais metais ypač sparčiai populiarėja ir plečiasi dirbtinio intelekto sritis, tobuliami mašininio mokymo algoritmai dėl besiformuojančių skaičiavimo iššūkių.

Jie susiję su dideliu ir nuolat srautu gaunami duomenų rinkinių interpretavimu, mokymosi ir sprendimų priėmimu realiu laiku.

(pvz.: „AI at the Edge“).

Gartner Hype Cycle for Emerging Technologies, 2019



Tyrimų sritis ir problemos aktualumas (2)

Srautu gaunami duomenų rinkiniai gali būti modeliuojami kaip stochastiniai procesai. PMM parametrų rekursinio vertinimo būdai:

- stebėjimai turi būti apdorojami nuosekliai, nesaugant jų kompiuterio atmintyje,
- PMM parametrai gali kisti laike.

PMM parametrų vertinimas remiasi:

- rekursiniais didžiausio tikėtino metodoais,
- prognozavimo klaidų minimizavimo metodais,
- minimalios modelio divergencijos metodais,
- pseudo-tikėtino metodoais.
 - linkę konverguoti į jų tikslo funkcijų lokalius ekstremumus.

Mokslinis darbo naujumas

1. Sukurti ir eksperimentiškai ištirti daugiamačių stebėjimų rekursiniai klasifikavimo metodai, paremti paslėptaisiais Markovo modeliais.
2. Siūlomas daugiamačių Gauso paslėptųjų Markovo modelių būsenų perėjimo tikimybių rekursinis skaičiavimo metodas, paremtas „Čepmeno-Kolmogorovo“ lygtimi. Šis būsenų perėjimo tikimybių skaičiavimo metodas leidžia pasiekti didesnį klasifikavimo tikslumą nei naudojant įprastinę modifikuotą „Forward-Backward“ procedūrą žinomuose rekursiniuose algoritmuose.
3. Disertacijoje pateikiamas naujas rekursinis algoritmas paslėptųjų Markovo modelių parametrų vertinimui, kai stebėjimai yra pasiskirstę pagal Dirichlė skirstinį.
4. Sudaryti algoritmai buvo pritaikyti izoliuotų žodžių atpažinime, užimtumo nustatymo ir pulsarų nustatymo uždaviniuose. Eksperimentiniai rezultatai patvirtino pasiūlytų algoritmų efektyvumą - sumažėjęs algoritmų skaičiavimo laikas ir nežymiai pakitęs (iki 3%) klasifikavimo tikslumas.

Praktinė darbo reikšmė

1. rekursinis Gauso PMM parametrų vertinimo algoritmas buvo pritaikytas izoliuotų žodžių atpažinime, kai fiksuotas kiekis šnekos duomenų yra skirtas apmokymui, o tolimesni šnekos duomenys yra atpažįstami ir naudojami pakartotiniam parametrų vertinimui.
2. rekursinis Dirichlė PMM parametrų vertinimo algoritmas buvo pritaikytas užimtumo nustatymo uždavinyje, kai iš sensorių gautų duomenų reikia nustatyti, ar analizuojama patalpa yra užimta, ar ne.
3. rekursinis Dirichlė PMM parametrų vertinimo algoritmas buvo pritaikytas pulsarų kandidatų nustatymo uždavinyje.

Ginamieji teiginiai

1. Rekursinis PMM parametrų vertinimo algoritmas, kai išvesties tikimybinis skirstinys yra Gauso, yra iš esmės greitesnis ir klasifikavimo (stebėjimų atpažinimo) tikslumu prilygstantis tradiciniam (blokiniam) „Baum-Welch“ algoritmui.
2. „Chapman-Kolmogorov“ lygties taikymas modelio būsenų perėjimo tikimybių skaičiavime pagerina modelio parametrų konvergavimą lyginant su tradicine modifikuota „Forward-Backward“ procedūra.
3. Egzistuoja pakankamas duomenų rinkinys, skirtas pradinei algoritmo aproksimacijai, užtikrina algoritmo stabilumą ir neleidžia jam konverguoti į lokalius tikėtinumo funkcijos ekstremumus.
4. Rekursinis Gauso ir Dirichlė PMM parametrų vertinimo algoritmi gali būti taikomi kelių klasių klasifikavimo praktiniams uždaviniams, išreiškiamiems stochastiniu procesu ir modeliuojamiems paslėptaisiais Markovo modeliais, spręsti.

Darbo rezultatų aprobavimas (1)

Respublikinės konferencijos:

1. Lietuvos Matematikų Draugijos 59-oje konferencija. Pranešimas „*Rekurentinis paslėptųjų Markovo modelių parametrų vertinimo algoritmas ir jo taikymai*“. Lietuva, Kaunas, birželio 18-19 d., 2018.
2. „Kompiuterininkų dienų 2017“ konferencija. Pranešimas „*Rekurentinis paslėptųjų Markovo modelių parametro vertinimo algoritmas*“. Lietuva, Kaunas, rugsėjo 21–22 d., 2017.
3. Respublikinė mokslinė-praktinė konferencija „*Informacinių technologijų iššūkiai kūrybos ekonomikoje*“. Pranešimas „*Paslėptųjų Markovo modelių parametrų rekurentinis vertinimas*“. Lietuva, Šiauliai, kovo 17 d., 2017.
4. „*Data Analysis Methods for Software Systems*“ konferencija. Stendinis pranešimas „*Recurrent estimation of homogeneous Hidden Markov model parameters*“. Lietuva, Druskininkai, gruodžio 1-3 d., 2016.
5. Konferencija „*Operacijų tyrimas ir taikymai*“. Pranešimas „*Paslėptųjų Markovo modelių parametrų vertinimo algoritmo tyrimas*“. Lietuva, Kaunas, Balandžio 8 d., 2016.
6. „Kompiuterininkų dienų 2015“ konferencija. Pranešimas „*Automatinio šnekos atpažinimo metodų tyrimas ir taikymas balso įrašams stenografuoti*“. Lietuva, Panevėžys, rugsėjo 17–19 d., 2015.

Darbo rezultatų aprobavimas (2)

Tarptautinės konferencijos:

1. Tarptautinė konferencija *ICIC (3rd International Conference INNOVATIONS AND CREATIVITY)*. Pranešimas „*Recursive Dirichlet Hidden Markov Model Parameter Estimation Algorithm*“. Latvija, Liepoja, Birželio 6–8 d., 2019.
2. Tarptautinė konferencija *EURO 2018 (29th European Conference on Operational Research)*. Pranešimas „*Recurrent parameter estimation algorithm in hidden Markov models with application to multivariate data analysis and signal recognition*“. Ispanija, Valensija, Liepos 8-11 d., 2018.

Darbo rezultatų publikavimas (1)

Straipsniai recenzuojamuose Lietuvos ir užsienio leidiniuose:

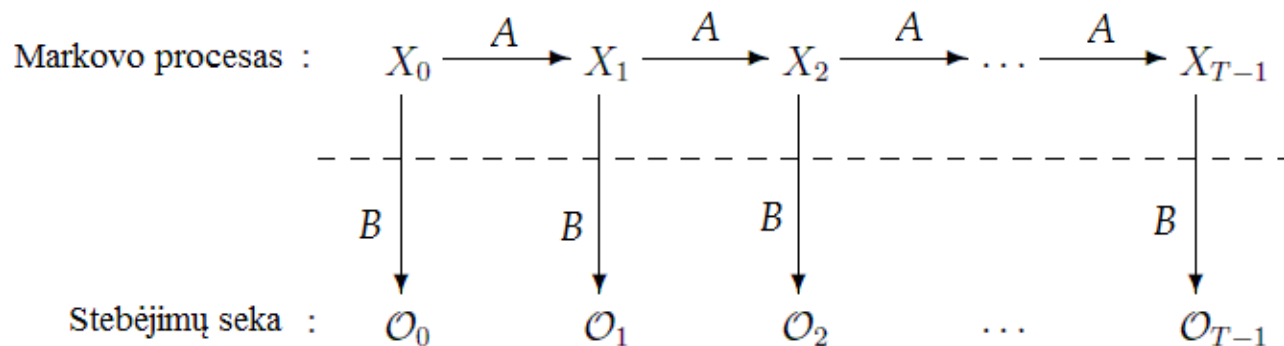
1. Vaičiulytė J., Sakalauskas L., 2019. Recursive parameter estimation algorithm of the Dirichlet hidden Markov model. *Journal of Statistical Computation and Simulation*. Taylor & Francis Production. <https://doi.org/10.1080/00949655.2019.1679144> (ISI WoS, JIF =0.767)
2. Vaičiulytė J., Sakalauskas L., 2019. Recursive estimation of multivariate hidden Markov model parameters. *Computational statistics*. Heidelberg, Springer. ISSN 0943-4062. eISSN 1613-9658. 2019, vol. 34, p. 1337-1353. <https://doi.org/10.1007/s00180-019-00877-z> (ISI WoS, JIF = 0.680)
3. Vaičiulytė J., Sakalauskas L., 2017. Rekurentinis paslėptųjų Markovo modelių parametro vertinimo algoritmas. *Computational Science and Techniques*. Vol 5, No 1 (2017), 529-540, eISSN: 2029-9966, <https://doi.org/10.15181/csat.v5i1.1561>. (Index Copernicus, CEEOL)
4. Vaičiulytė J., Felinskas G., 2016. Paslėptųjų Markovo modelių metodo tyrimas ir taikymas balso įrašams stenografuoti. *Jaunųjų Mokslininkų Darbai*, 1(45), 71-78. <https://doi.org/10.21277/jmd.v1i45.40> (Index Copernicus, CEEOL)
- **[publikacija recenzuojama]** Vaičiulytė J., Sakalauskas L., 2019. Isolated Word Recognition by Recursive HMM Parameter Estimation Algorithm. *Computing and Informatics*. (ISI WoS)

Darbo rezultatų publikavimas (2)

Santraukos konferencijų leidiniuose:

- Vaičiulytė, Jūratė. *Recurrent estimation of homogeneous Hidden Markov model parameters*. Data analysis methods for software systems : 8th international workshop on data analysis methods for software systems [abstracts book], Druskininkai, December 1-3, 2016. Vilnius : Vilniaus universiteto leidykla, 2016. ISBN 9789986680611. p. 62.

Paslēptieji Markovo modeļi



$B=\{b_j(o)\}$	Stebējimo tikimybės tankio funkcija būsenoje j
$A=\{a_{i,j}\}$ $a_{i,j}=P[x_{t+1}=X_j x_t=X_i]$	Perėjimo iš vienos būsenos į kitą tikimybių matrica Tikimybė iš būsenos i pereiti į būseną j
X_t	(Paslėpta) būsena laiko momentu t
$\pi_i = P(X_1=i)$	Pradinio buvimo būsenoje tikimybės (pirmoji paslėptojo Markovo modelio būsena $\pi_i = 1$)
N	Būsenų skaičius

Klasikinis EM algoritmas:

Stebėjimo tikimybės tankio funkcija - Gauso

E-žingsnis: logaritminės tikėtinumo funkcijos sąlyginio vidurkio apskaičiavimas

M-žingsnis: tikėtinumo funkcijos sąlyginio vidurkio maksimizavimas

$$\bar{\mu}_j = \frac{\sum_{j=1}^T \gamma(j) \cdot O}{\sum_{j=1}^T \gamma(j)} \quad \bar{\sigma}_j = \frac{\sum_{j=1}^T \gamma(j) \cdot (O - \mu_j)(O - \mu_j)^T}{\sum_{j=1}^T \gamma(j)} \quad \gamma_j = e^{-l(o, \mu_i, \sigma_i, \pi_i)}$$

$$l(o, \mu, \sigma, \pi) = \frac{(o - \mu)^T \sigma^{-1} (o - \mu)}{2} + \ln \left(\frac{\sqrt{|\sigma|}}{\pi} \right)$$

$$\bar{a}_{ij} = \frac{\text{expected number of transitions from state } S_i \text{ to the state } S_j}{\text{expected number of transitions from state } S_i}.$$

„Forward-Backward“ procedūra

$$N(o, \mu_s, \sigma_s) = \frac{1}{\sqrt{(2\pi)^n |\sigma_s|}} e^{-\frac{1}{2}(o - \mu_s)^T \sigma_s^{-1} (o - \mu_s)}$$

Sudarytas rekursinis EM algoritmas PMM parametrų vertinimui:

Stebėjimo tikimybės tankio funkcija - Gauso

Būsenų perėjimo tikimybės apskaičiuojamos pritaikant „Chapman-Kolmogorov“ lygtis :

$$\pi_t = A \cdot \pi_{t-1}$$

Vidurkių vektorius, kovariacijų matrica ir buvimo būsenoje tikimybės apskaičiuojamos pagal formules:

$$\mu_t^i = \mu_{t-1}^i + \frac{(o_t - \mu_{t-1}^i)}{t} \cdot \frac{\omega_t^i}{\beta_t},$$
$$\sigma_t^i = \left(\frac{\beta_{t-1}^i \cdot (t-1)}{\beta_t^i \cdot t} \right) \cdot \left(\sigma_{t-1}^i + \frac{(o_t - \mu_{t-1}^i)(o_t - \mu_{t-1}^i)^T}{t} \cdot \frac{\omega_t^i}{\beta_t} \right),$$

$$\beta_t^i = \beta_{t-1}^i + \frac{1}{t} (\omega_t^i - \beta_{t-1}^i).$$

čia koeficientai:

$$\beta_t^i = \frac{1}{t} \sum_{i=1}^t \omega_t^i \quad \omega_t^i = \frac{e^{-l(o_t, \hat{\mu}_i, \hat{\sigma}_i, \pi_i)}}{\sum_{i=1}^N e^{-l(o_t, \hat{\mu}_i, \hat{\sigma}_i, \pi_i)}}$$

Logaritminė tikėtinumo f-ja:

$$l(o, \mu, \sigma, \pi) = \frac{(o - \mu)^T \sigma^{-1} (o - \mu)}{2} + \ln \left(\frac{\sqrt{|\sigma|}}{\pi} \right)$$

Rekursinis EM algoritmas

Klasikinis EM algoritmas reikalauja kiekvienoje iteracijoje apdoroti visą turimą stebėjimų imtį, todėl apdorojant visą imtį kiekvienoje iteracijoje algoritmo sudėtingumas tampa kvadratinis, t.y. $O(T^2)$.

- Stebint procesą reikalingų vertinimų operacijų skaičius labai auga ir vertinimas realiu laiku tampa nebeįmanomas.

Rekursinio EM algoritmo atveju - PMM modelio parametrai yra atnaujinami su kiekvienu gautu nauju stebėjimu, neįsimenant ankstesnės mokymo aibės.

- Algoritmo sudėtingumas yra $O(T)$.

Eksperimentai

Duomenų rinkiniai klasterizavimui buvo atsitiktinai generuojami iš dviejų ($N = 2$) daugiamačių Gauso skirstinių, turinčių šias charakteristikas:

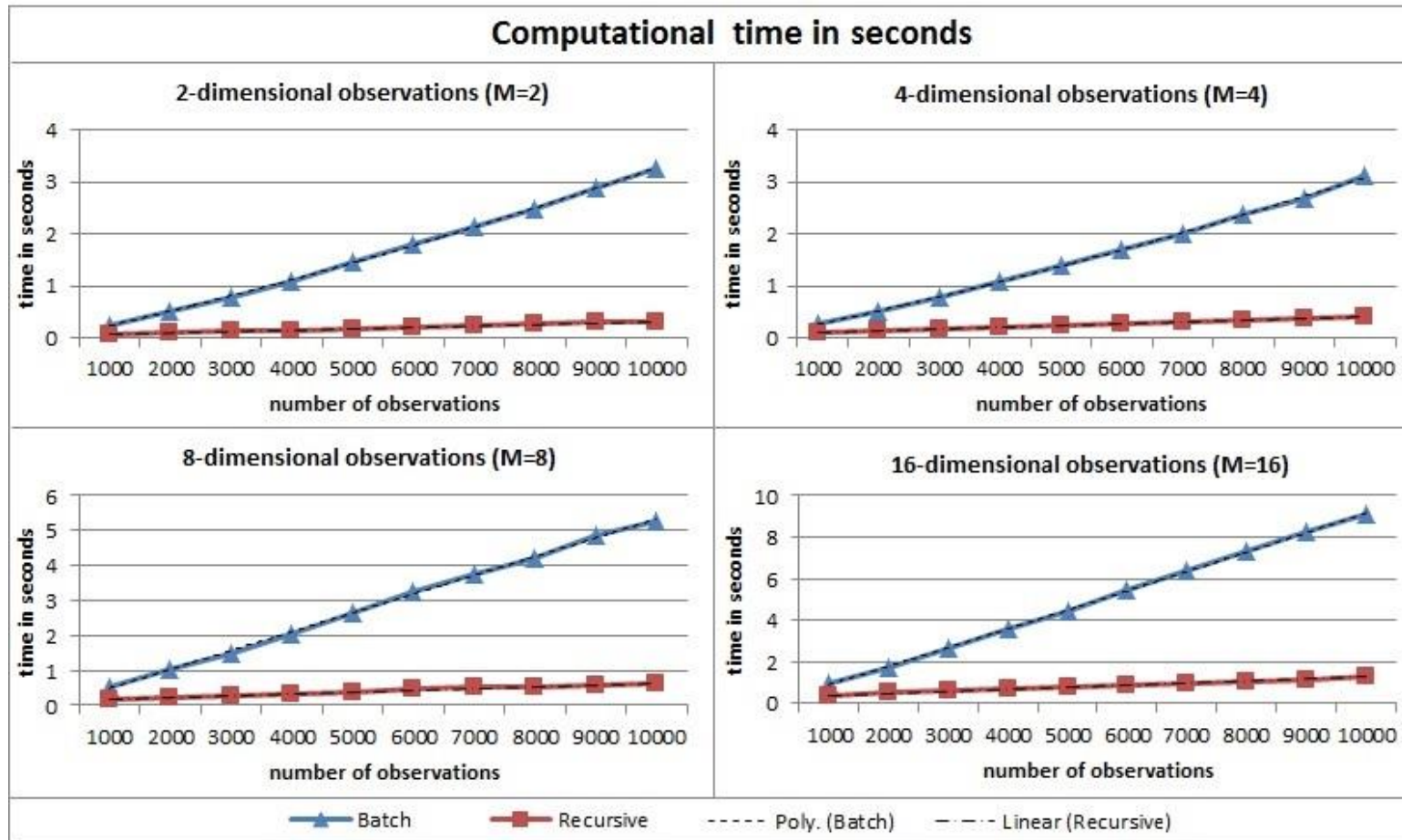
- Klasterių tikrieji centrai: $[500, 500, \dots]$ ir $[600, 600, \dots]$.
- Dimensijų skaičius: 2, 4, 8, 16.
- Standartinis nuokrypis: 20, 30, 40, 50.

Pradinis parametrų vertinimas buvo atliktas naudojant fiksuoto dydžio ($t = 500$) mokymo stebėjimų rinkinį.

Eksperimento rezultatai gauti apskaičiavus vidurkius iš atliktų 100 eksperimento pakartojimų.

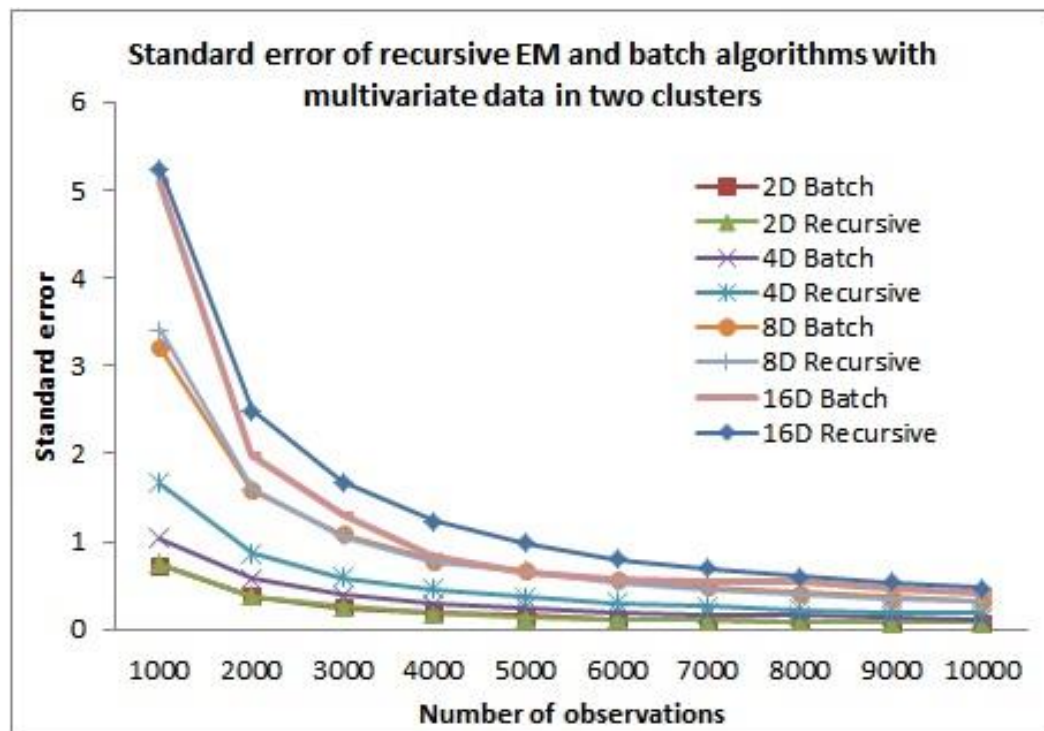
Eksperimentų rezultatai:

algoritmo skaičiavimo laikas



Eksperimentų rezultatai:

algoritmo konvergavimo savybė (1)

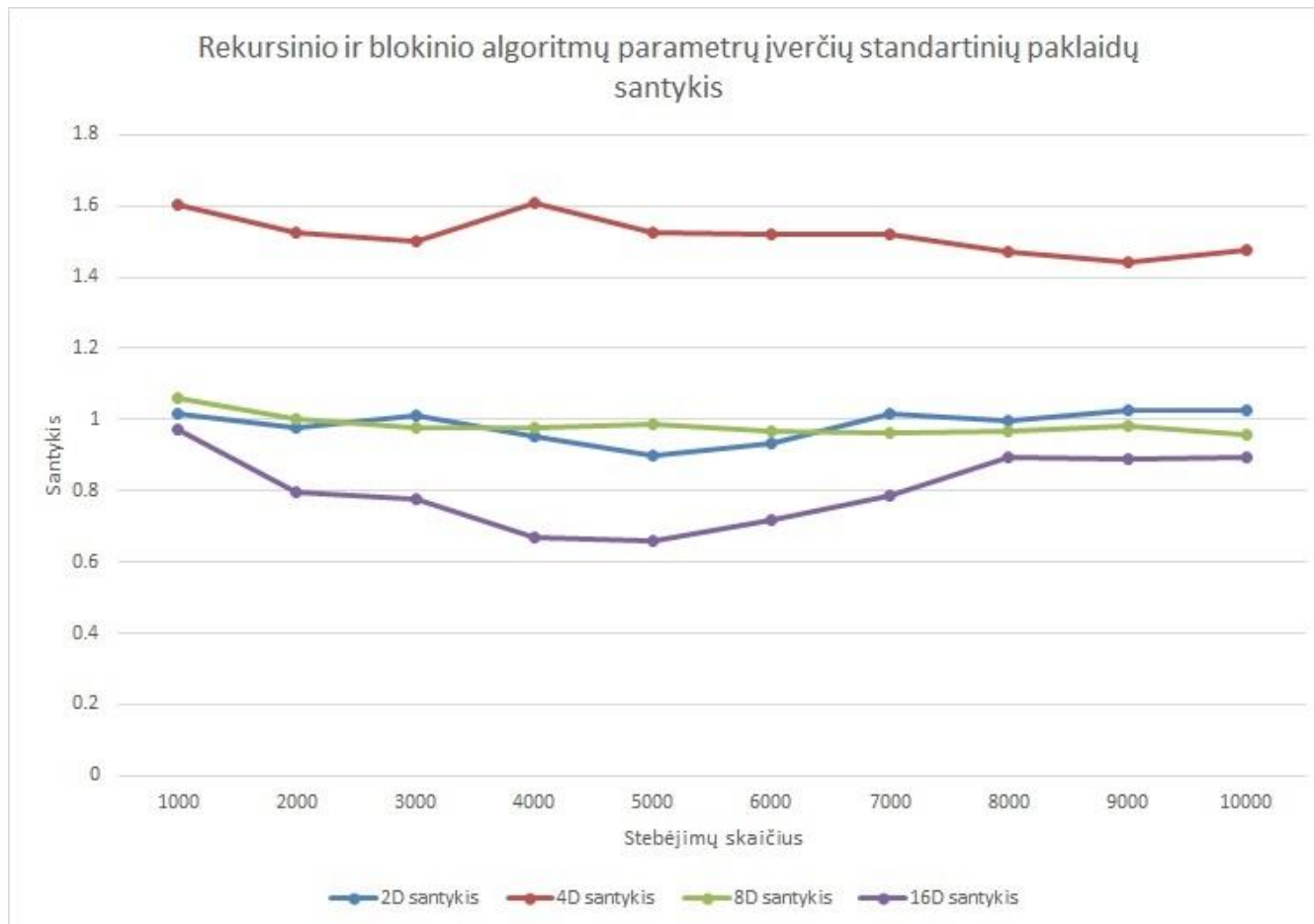


Blokinio ir rekursinio algoritmų būsenų perėjimo tikimybių matrica, gauta PMM parametrų vertinimo metu

		M = 2			
T=		Blokinis		Rekursinis	
1000		0.497	0.503	0.497	0.503
		0.472	0.528	0.472	0.528
10000		0.506	0.494	0.506	0.494
		0.498	0.502	0.498	0.502

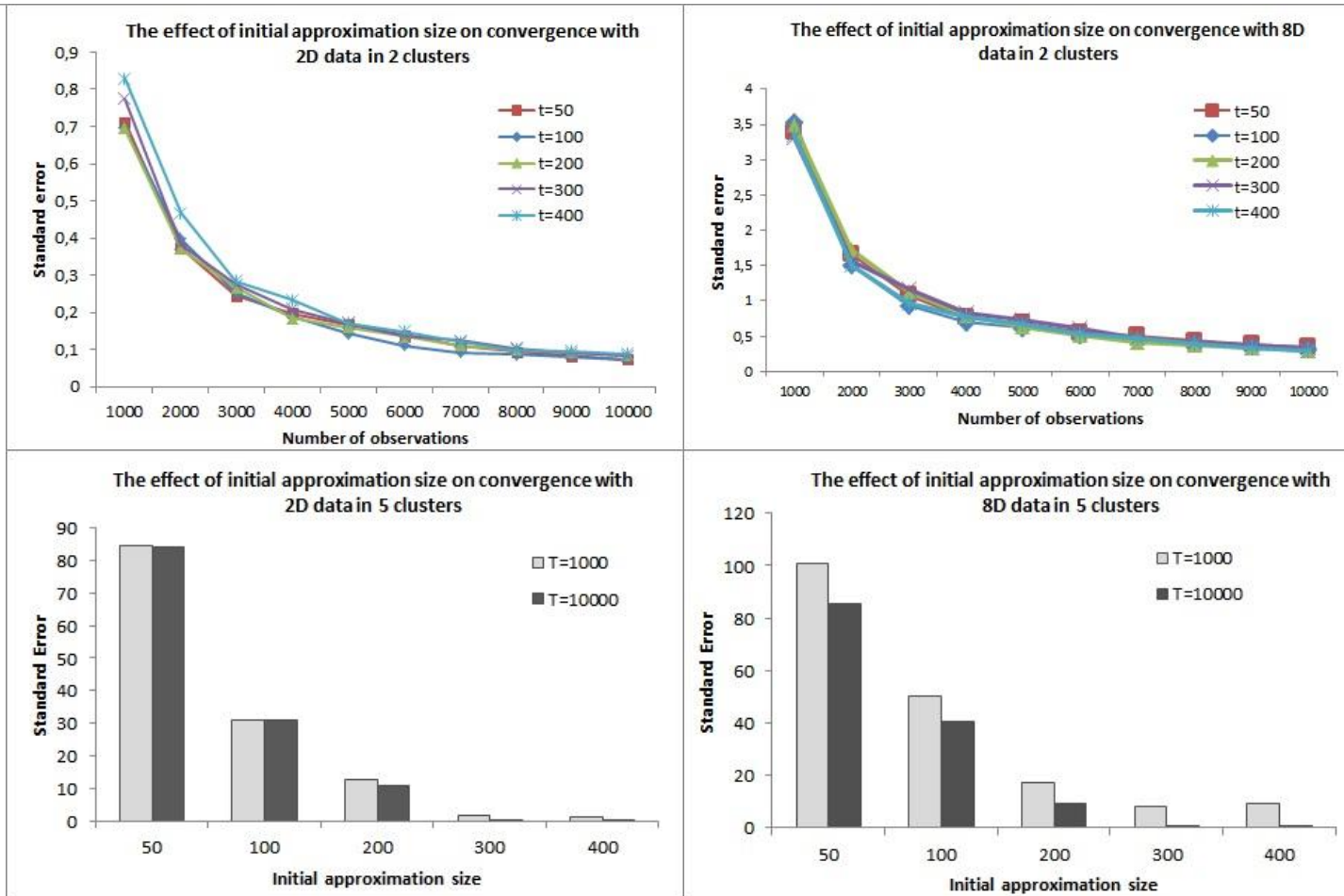
Eksperimentų rezultatai:

algoritmo konvergavimo savybė (2)



Eksperimentų rezultatai:

algoritmo pradinės aproksimacijos įtakos analizė



Eksperimentų rezultatai:

palyginimas su rekursiniu [Stenger01] algoritmu

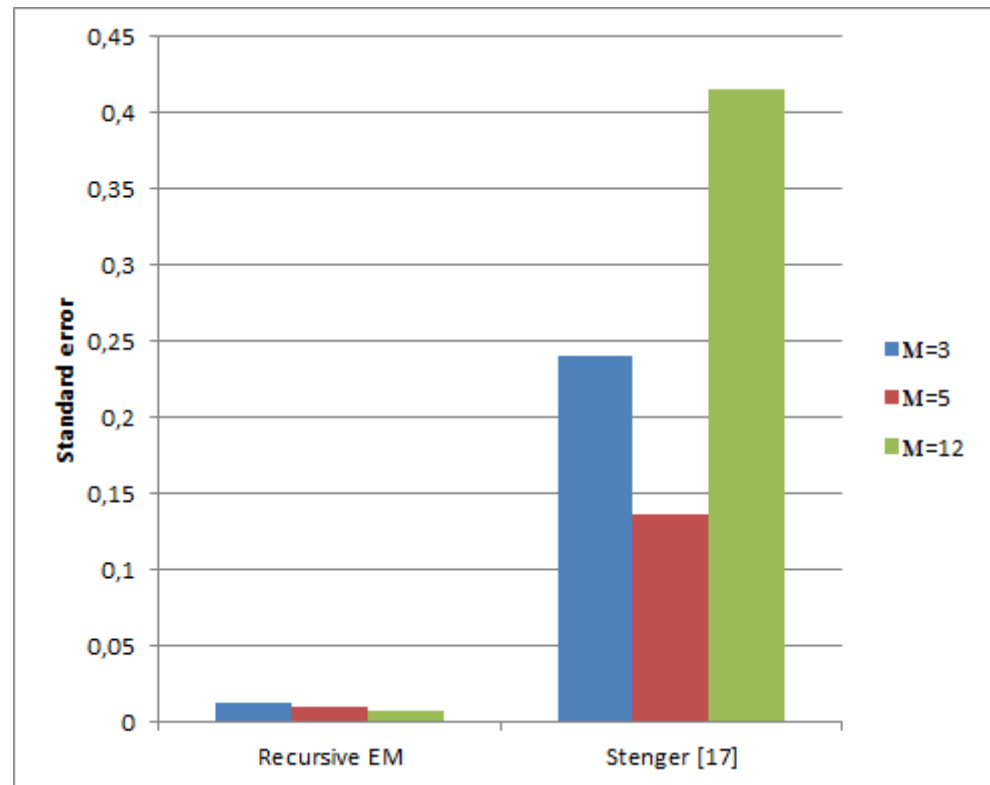
Generuojami stebėjimų vektoriai pasiskirstę pagal daugiamatį Gauso skirstinį.

Būsenų skaičius: 5

Stebėjimų dimensijų sk.: 3, 5, 12

Kiekvienos aibės tūris: 800 stebėjimų

Eksperimentas pakartotas 100 kartų, apskaičiuota įverčių standartinė paklaida.



Dirichlė PMM modelio parametrų rekursinis vertinimas

Dirichlė PMM parametrai:

- PMM būsenų skaičius: N
- Būsenų perėjimų tikimybių matrica P
- Pradinio buvimo būsenoje tikimybių vektorius π
- $\alpha_1, \dots, \alpha_K$, čia $\alpha_i > 0$ ir $K > 2$
- $x_1, \dots, x_K$, čia $x_i \in (0,1)$ ir $\sum_{i=1}^K x_i = 1$
- Logaritminė tikėtinumo funkcija:

$$\log Dir(x, \alpha) = \ln \Gamma\left(\sum_{i=1}^K \alpha_i\right) - \sum_{i=1}^K \ln \Gamma(\alpha_i) + \left[(\alpha_K - 1) \ln\left(1 - \sum_{i=1}^{K-1} x_i\right)\right] + \sum_{i=1}^{K-1} [(\alpha_i - 1) \ln(x_i)]$$

Didžiausio tikėtinumo metodu išvestos rekursinės formulės Dirichlė PMM parametrų vertinimui

$$\theta_t^{\langle q \rangle} = P_t^{\langle q \rangle} \log \text{Dir}(x_t, \alpha^{\langle q \rangle})$$

$$\omega_t^{\langle q, s \rangle} = \omega_{t-1}^{\langle q, s \rangle} + \frac{1}{t} \left(\frac{\ln(x_t^{\langle s \rangle}) \theta_t^{\langle q \rangle}}{\sum_{j=1}^N \theta_t^{\langle j \rangle}} - \omega_{t-1}^{\langle q, s \rangle} \right), 1 \leq q \leq N$$

$$\omega_t^{\langle q, K \rangle} = \omega_{t-1}^{\langle q, K \rangle} + \frac{1}{t} \left(\frac{\ln(1 - \sum_{s=1}^{K-1} x_t^{\langle s \rangle}) \theta_t^{\langle q \rangle}}{\sum_{j=1}^N \theta_t^{\langle j \rangle}} - \omega_{t-1}^{\langle q, K \rangle} \right)$$

$$\gamma_t^{\langle q \rangle} = \gamma_{t-1}^{\langle q \rangle} + \frac{1}{t} \left(\frac{\theta_t^{\langle q \rangle}}{\sum_{j=1}^N \theta_t^{\langle j \rangle}} - \gamma_{t-1}^{\langle q \rangle} \right)$$

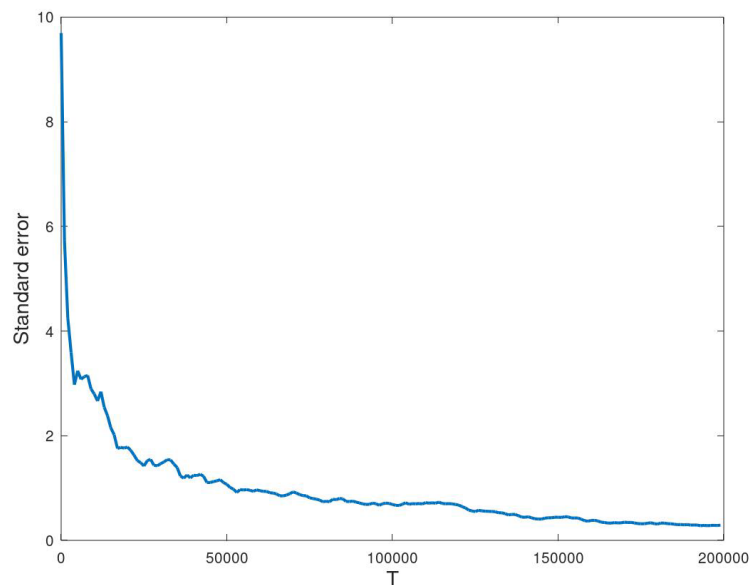
$$\alpha_t^{\langle q, s \rangle} = \Psi^{-1} \left[\text{Psi} \left(\sum_{i=1}^K \alpha_i^{\langle q \rangle} \right) + \frac{\omega_t^{\langle q, s \rangle}}{\gamma_t^{\langle q \rangle}} \right], 1 \leq s \leq K$$

- Ψ^{-1} – atvirkštinė digamma funkcija

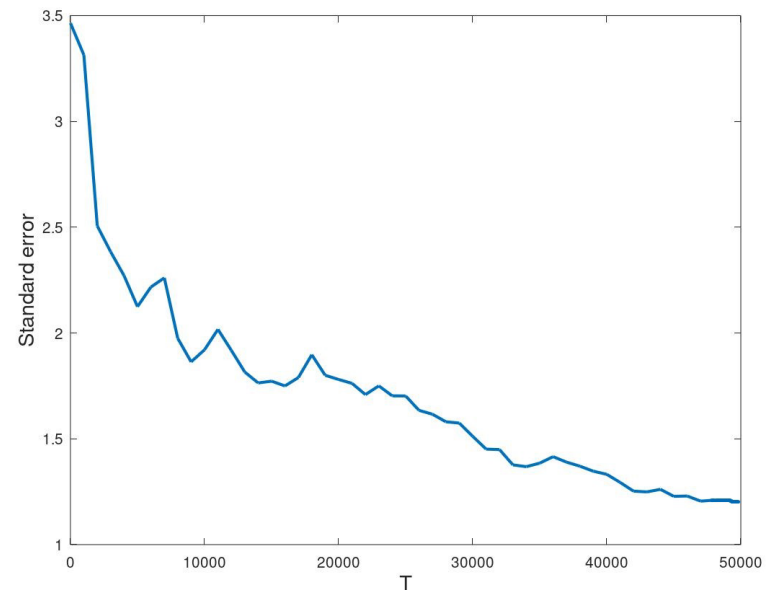
Eksperimentų rezultatai:

algoritmo konvergavimo savybė

Modelio parametrų įverčių standartinė paklaida (3-būsenų PMM su trimačiais stebėjimais)



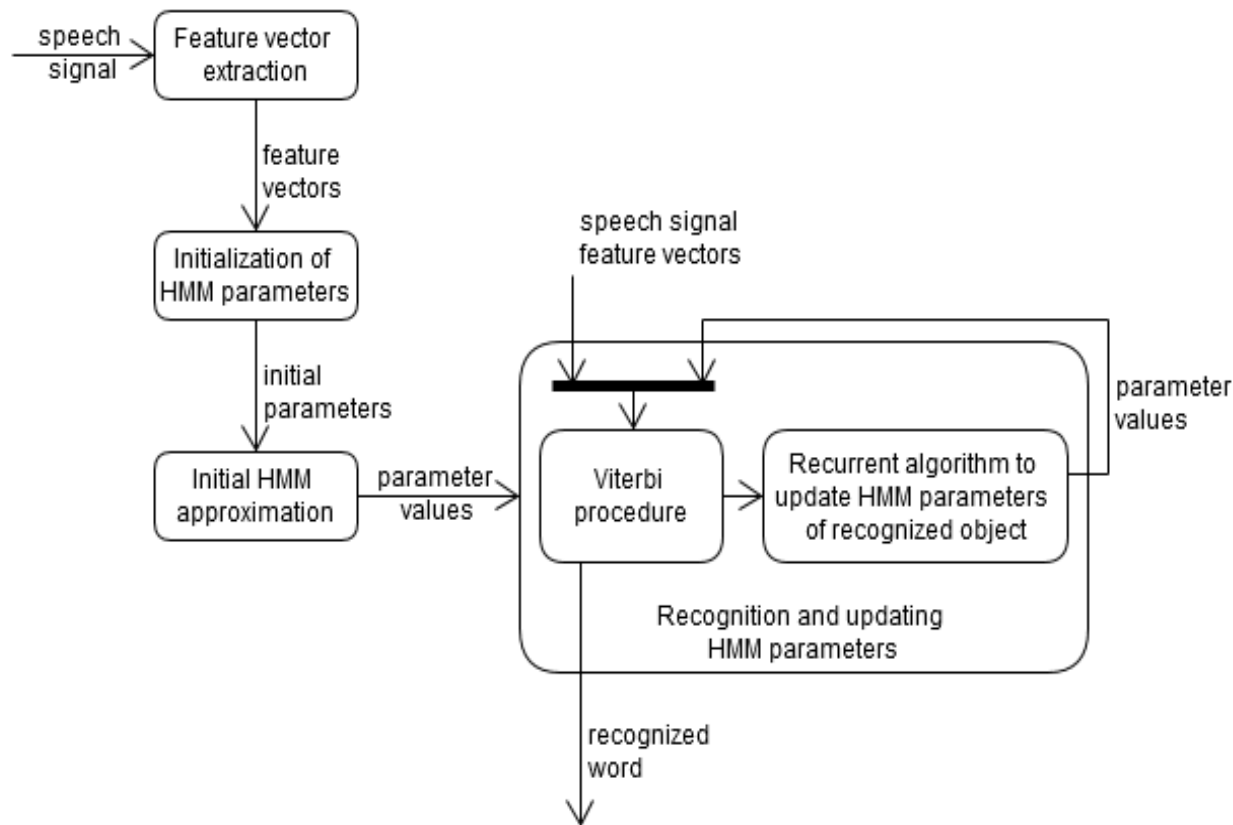
Modelio parametrų įverčių standartinė paklaida (5-būsenų PMM su trimačiais stebėjimais)



Eksperimentas pakartotas 100 kartų, apskaičiuota įverčių standartinė paklaida.

Taikymai

Gauso PMM parametų vertinimo rekursinis algoritmas pritaikytas izoliuotų žodžių atpažinimui



Izoliuotų žodžių atpažinimo eksperimentų rezultatai

Eksperimentai atlikti su „TIDIGITS“ izoliuotų žodžių duomenų aibe:

Mokymo imties dydis (žodžiais)	Atpažinimo tikslumas (%)
100	92.53
500	94.33
1000	95.87
1500	97.60
2000	97.27

Eksperimentai atlikti su „Spoken Arabic Digits“ garsynu, kurį sudaro mokymo (8143 stebėjimai) ir testavimo (2665 stebėjimai) poaibiai.

- Gautas izoliuotų žodžių atpažinimo tikslumas - 91,86%

Eksperimentai:

užimtumo nustatymas

Eksperimentams duomenų rinkinys buvo suskaidytas į dvi dalis: mokymo (8143 stebėjimai) ir testavimo (2665 stebėjimai).

7 požymių stebėjimų vektoriai.

Maišos matrica užimtumo nustatymui

		Priskirta klasė	
		Neužimta	Užimta
Tikra klasė	Neužimta	TP	FN
	Užimta	FP	TN

Algoritmų efektyvumo rodikliai

	Dirichlė	Gauso
Accuracy	0,977110694	0,870168856
Precision	0,940948693	0,830168776
Recall	1	0,809670782
F-score	0,96957606	0,819791667

Aukščiausius F-Score reikšmes gavę
ne rekursiniai algoritmai
(LR, SVM ir ANN) [2].

LR	SVM	ANN
95,6	95,55	95,32

Eksperimentai:

pulsarų nustatymas

HTRU2 duomenų rinkinyje yra 16259 netikri (triukšmo sukelti) kandidatai (ne pulsarai) ir 1639 realūs pulsarai.

Kiekvienas stebėjimų vektorius susideda iš 8 požymių.

Palyginti rekursiniai Dirichlė PMM, Gauso PMM ir ne rekursinis Fuzzy knn algoritmas

	Dirichlė	Gauso	Fuzzyknn
Accuracy	0,92	0,82	0.978
FPR	0,065	0,183	0.17
F-score	0,571	0,365	0.873
Gmean	0,886	0,862	0.961

Maišos matrica pulsarų nustatymui

		Priskirta klasė	
		Pulsaras	Ne pulsaras
Tikra klasė	Pulsaras	TP	FN
	Ne pulsaras	FP	TN

Išvados (1)

Pagrindiniai rezultatai:

- Remiantis didžiausio tikėtinumo metodu ir klasikiniu EM algoritmu, suformuoti rekursiniai PMM parametrų vertinimo algoritmai, kai daugiamačiai Gauso ir Dirichlė dėsniai yra modelio būsenų išvesties tikimybiniai skirstiniai.
- PMM parametrų vertinimo dalyje būsenų perėjimo tikimybės siūlomos skaičiuoti remiantis „Chapman-Kolmogorov“ lygtimi. Tokiu būdu pakeista tradicinė „Forward-Backward“ procedūra, įprastai taikoma „Baum-Welch“ algoritmuose, o rekursiniuose - modifikuojama praleidžiant „Backward“ dalį.
- Atliktas pasiūlytų rekursinių algoritmų eksperimentinis tyrimas su sintetiniais duomenimis Monte-Karlo metodu, izoliuotų žodžių, užimtumo nustatymo ir pulsarų nustatymo duomenų rinkiniais.

Išvados (2)

Rekursinis Gauso PMM parametrų vertinimo algoritmas yra greitesnis ir klasifikavimo tikslumu prilygstantis tradiciniam (blokiniam) „Baum-Welch“ algoritmui.

- Atlikus eksperimentus gautas duomenų apdorojimo greitis rekursiniu algoritmu yra pranašesnis negu blokinio algoritmu. Rekursinio algoritmo skaičiavimo laikas yra 3-9 kartus greitesnis nei blokinio algoritmo, kai apdorojami skirtingo dimensijų skaičiaus stebėjimai.
- Blokinio ir rekursinio Gauso PMM algoritmų būsenų perėjimo tikimybių matricos, gautos PMM parametrų vertinimo metu apdorojant iki 10000 stebėjimų, nesiskiria.
- Rekursiniu ir blokinio algoritmais gautų PMM parametrų įverčių standartinių paklaidų santykis rodo, kad rekursinis algoritmas, apdorojant skirtingų dimensijų stebėjimus, duoda minimaliai nuo blokinio algoritmo besiskiriančius paramaterų įverčius.
- Atlikus eksperimentinį tyrimą ir palyginus rekursinio PMM parametrų vertinimo algoritmo rezultatus su žinomu rekursiniu algoritmu, matome, jog būsenų perėjimo tikimybių skaičiavimas su „Chapman-Kolmogorov“ lygtimi pagerina parametrų konvergavimą į tikrąsias modelio parametrų reikšmes, lyginant su algoritmu, kuriame būsenų perėjimo tikimybių matrica yra skaičiuojama su modifikuota „Forward-Backward“ procedūra.

Išvados (3)

Pradinis duomenų rinkinys, skirtas pradinei algoritmo aproksimacijai, užtikrina algoritmo stabilumą ir neleidžia jam konverguoti į lokalius tikėtinumo funkcijos ekstremumus, tačiau turint pakankamą pradinės aproksimacijos duomenų rinkinį ir jį didinant atpažinimas iš esmės nesikeičia.

- Standartinė parametrų įverčių paklaida mažėja, didėjant bendram duomenų rinkinio dydžiui ir esant mažam klasterių skaičiui, net jei pradinės aproksimacijos duomenų rinkinio dydis yra mažas. Didėjant klasterių skaičiui, pradinės aproksimacijos duomenų rinkinys taip pat turi būti didinamas.
- Esant nedideliame klasterių (PMM būsenų) skaičiui, pradiniam mokymui naudojamų duomenų rinkinio dydis taip pat gali būti mažas, nepriklausomai nuo stebėjimo vektorių dimensijų skaičiui.
- Pritaikius rekursinį Gauso PMM parametrų vertinimo algoritmą izoliuotų žodžių atpažinimui pasiektas 97% atpažinimo tikslumas „TIDIGITS“ garsyno atveju, o „Spoken Arabic Digits“ garsyno atveju atpažinimo tikslumas siekė 91%.

Išvados (4)

Duomenų sklaida ir dimensijų skaičius įtakoja rekursinį modelio parametrų vertinimo efektyvumą - augant dimensijų skaičiui, didėja standartinė paklaida - skirtumas tarp tikrųjų parametrų reikšmių ir parametrų įverčių.

- Standartinė paklaida, kai duomenys yra dvimačiai ir sklaida - 20, yra apie 4-6 kartus mažesnė negu standartinė paklaida, gauta apdorojus 16-mačius duomenis, kurių sklaida yra 50.
- Didėjant duomenų sklaidai ir dimensijoms, lėtėja algoritmo konvergavimo laikas. Tokiu atveju algoritmo konvergavimui reikalingas duomenų kiekis padidėja.
- Rekursinio ir blokinio PMM parametrų vertinimo algoritmų standartinių paklaidų santykis svyruoja nuo 1.1 iki 1.5 karto, priklausomai nuo apdorojamų duomenų dimensijų ir sklaidos didėjimo.

Išvados (5)

Atlikus eksperimentinius tyrimus su rekursiniu Dirichlė PMM parametrų vertinimo algoritmu, suformuluotos tokios išvados bei jas patvirtinantys eksperimentiniai rezultatai:

- Atliktų eksperimentų, skirtų patikrinti rekursinio Dirichlė PMM parametrų vertinimo algoritmo efektyvumą lyginant su rekursiniu Gauso PMM parametrų vertinimo algoritmu, rezultatai parodė, kad plačiai paplitę Gauso PMM negali būti taikomi uždaviniuose, kuriuose stebėjimai yra pasiskirstę pagal Dirichlė dėsnį - tokiu atveju Dirichlė PMM yra efektyvesnis ir duoda didesnį stebėjimų klasifikavimo tikslumą.
- Atlikus eksperimentus nustatyta, kad atpažinimo (klasifikavimo) tikslumas didėjant mokymo duomenų kiekiui yra nuosekliai gerinamas, o standartinė parametrų įverčių paklaida mažėja.
- Pritaikius Dirichlė PMM praktiniuose uždaviniuose, gauti rezultatai parodė, kad užimtumo nustatymo duomenų rinkinio atveju pasiektas 97% atpažinimo tikslumas, o pulsarų nustatymo duomenų rinkinio atveju atpažinimo tikslumas siekė 93%.



Vilniaus Universiteto
Duomenų mokslo ir skaitmeninių technologijų institutas



PASLĖPTŲJŲ DAUGIAMAČIŲ MARKOVO MODELIŲ TYRIMAS IR TAIKYMAS REALAUS LAIKO SISTEMŲ ANALIZEI

DOKTORANTĖ: JŪRATĖ VAIČIULYTĖ

DARBO VADOVAS: PROF. HABIL. DR. LEONIDAS SAKALAUSKAS