Liet. matem. rink, **45***, spec. nr.*, 2005, 454–458

Tirpalų mišinių koncentracijų klasifikavimas naudojant apibendrintą pagrindinių komponenčių regresiją

Romas BARONAS (VU), Feliksas IVANAUSKAS (VU, MII),
Robertas PAULAUSKAS (VU), Pranas VAITKUS (VU)
el. paštas: romas.baronas@maf.vu.lt

Reziumė. Darbe biojutiklių atsakui į tirpalų mišinius analizuoti yra taikoma apibendrinta pagrindinių komponenčių regresija. Šiuo tiesiniu mišinių koncentracijų klasifikatoriumi gautieji rezultatai lyginami su rezultatais gautais taikant dirbtinius neuroninius tinklus.

Raktiniai žodžiai: apibendrinta pagrindinių komponenčių regresija, apibendrinta atvirkštinė matrica.

1. Įvadas

Biojutikliai – tai įrenginiai, kuriuos pagrindinai sudaro biologiškai aktyvi medžiaga, dažniausiai fermentas ir elektroninis signalo keitiklis [1]. Fermentui reaguojant su analizuojamu tirpalu, fizikiniai-cheminiai pasikeitimai yra paverčiami elektros signalu, kurio stiprumas priklauso nuo tirpalo koncentracijos. Biojutikliai plačiai taikomi įvairiose analitinėse sistemose.

Šio darbo tikslas – pritaikyti apindrintąją pagrindinių komponenčių regresiją (APKR) biojutiklio signalui klasifikuoti [4]. Biojutiklio atsako kreivių taškai naudojami kaip nepriklausomi kintamieji, koncentracijų reikšmės laikomos priklausomais kintamaisiais. Šis darbas tęsia ankstesnį darbą [3], kuriame biojutiklio atsakui klasifikuoti buvo taikomi dirbtiniai neuroniniai tinklai. Abiejų darbų rezultatai lyginami tarpusavyje.

2. Matematinio modelio parinkimas

Modeliuojamieji duomenys

Tegul $\vec{c} = (c_1, \ldots, c_L)$ yra L tirpalų koncentracijų vektorius, $\vec{z} = \vec{z}(\vec{c}) = (z_1(\vec{c}), \ldots, z_P(\vec{c}))$ yra biojutiklio signalas momentais $t_1, \ldots, t_P$. $\vec{z}$ yra biojutiklio atsakas į mišinį $\vec{c} = (c_1, \ldots, c_L)$. Pažymėkime, $C = \{\vec{c}\}$ – aibė visų tirpalų galimų koncentracijų vektorių ir $Z = \{\vec{z}(\vec{c})\}$ – stebėtų biojutiklio signalų aibė. Aibė Z suskaidoma į dvi dalis: apmokymo ir testinę imtis. Naudojant apmokymo aibės elementus, apskaičiuojami nepriklausomieji regresijos koeficientai B [4].

Apibendrinta pagrindinių komponenčių regresija

Apibendrinta pagrindinių komponenčių regresija (TPCR, [4]), kaip ir *dalinių mažiausiųjų kvadratų* (PLS) (angl. k. *partial least squares*) analizė, naudoja informaciją apie nepriklausomąjį kintamąjį ir kintamųjų paklaidos modelį.

Tarkime, tikrosios nepriklausomojo kintamojo reikšmės yra gaunamos iš nestebimų paslėptų (nematomų) kintamųjų, jie guli mažesnės apimties tiesiniame poerdvyje, apimančiame paslėptuosius kintamuosius. Sudarome ortonormalią stulpelio atžvilgiu matricą T_{NxK} ($K < P$ ir $T^T T = I$), kurios stulpeliai $\tilde{Z}$ ir $\tilde{C}$ yra poerdvio bazė:

$$\tilde{Z} = TG, \tag{2.1}$$

$$\tilde{C} = \tilde{Z}B = TGB = TF, \tag{2.2}$$

čia $G_{K\times P}$ ir $F_{K\times L}$ yra $\tilde{Z}$ ir $\tilde{C}$ pakrovimo matricos. T galima laikyti paslėpta struktūra abiems $\tilde{Z}$ ir $\tilde{C}$. Iš (2.1) ir (2.2) gauname (EIV) (angl. k. e*rror in variables*) paslėptų kintamųjų modelį:

$$C = TF + E_C,$$

$$Z = TG + E_Z.$$

Šiam uždaviniui formuluojamas tikslo kriterijus

$$\min_{\substack{T,G,F\\ T^T T-1}} \Big(\frac{\|Z - TG\|_F^2}{\sigma_Z^2} + \frac{\|C - TF\|_F^2}{\sigma_C^2}\Big)$$

arba

$$\min_T \big\|(I - TT^T)A\big\|_F^2,$$

kur $||M||_F$ žymi Frobenius matricos normą, t.y. $\|M\|_F = [tr(MM^T)]^{1/2}$, ir A yra $N \times (P + L)$ išplėstoji Z ir C matrica, tai yra $A = (Z, \lambda C)$ ir $\lambda^2 = \frac{\sigma_Z^2}{\sigma_C^2}$.

Tegul ypatingosios reikšmės matricos A dekompozicija yra

$$A = U\Sigma V^T,$$

čia $U = (u_1, \ldots, u_N) \in R^{N\times N}$ yra kairysis ypatingasis vektorius su $U^T U = I_N$, ir $V = (v_1, \ldots, v_{(P+L)}) \in R^{(P+L)\times(P+L)}$ – dešinysis ypatingasis vektorius, kuriam $V^T V = I_{(P+L)}$, Σ yra įstrižainė matrica su ypatingosiomis reikšmėmis įstrižainėje ir kitais elementais lygiais 0. Tegul T yra pirmieji K matricos U stulpeliai

$$T = (u_1, \ldots, u_K).$$

Tada $\tilde{Z}$ ir $\tilde{C}$ įverčius galima gauti iš

$$\hat{Z} = TT^T Z,$$

$$\hat{C} = TT^T C.$$

TPCR įvertinti regresijos koeficientai yra gaunami iš

$$\hat{B} = (T^T Z)^+ T^T C,$$

kur viršutinis indeksas „+“ žymi apibendrintą atvirkštinę matricą.

Regresijos koeficientai įvertinami klasikiniu būdu

$$\hat{B} = (T^T T)^{-1} T^T C.$$

Metodo tikslumui įvertinti taikome

$$Q_k = \frac{1}{L} \sum_{i=1}^{L} Ind(\hat{C}_{i,k} \in \Delta y) \cdot Ind(C_{i,k} = y) \cdot 100\%, \tag{2.3}$$

kur indikatoriaus funkcija $Ind(\hat{C}_{i,k} \in \Delta y)$ yra lygi vienam, kai y_{ik} priklauso koncentracijų intervalui $(y - \delta_{1,y}, y + \delta_{2,y})$, kitu atveju 0, L – stebėjimų skaičius testinėje arba apmokymo aibėje.

3. Skaičiavimų rezultatai

Sumodeliuotų duomenų analizė

Naudojant modelį pateiktą [2], buvo modeliuojamas biojutiklių atsakas į keturių ($L = 4$) tirpalų mišinį, vonios (BA) ir apipurškimo (FIA) režimais. Vonios režimu biojutiklio veikimas modeliuotas dviems membranos sluoksniams: $d = 0,02$ ir $d = 0,04$ cm. Apipurškimo režime storis buvo $d = 0,02$ cm. Naudotos aštuonios kiekvieno tirpalo koncentracijos: 1, 2, 4, 8, 12, 16, 32, 64 nmol/cm^3. Iš viso buvo gauta 4096 skirtingų tirpalų koncentracijų vektorių. Biojutiklių signalo modeliavimo laikas: BA atveju kai ($d = 0,02$) 301 sekundžių ($N = 301$), BA atveju ($d = 0,04$) 501 sekundžių ($N = 501$) ir FIA atveju kai ($d = 0,02$) 151 sekundžių ($N = 151$), FIA atveju ($d = 0,04$) 301 sekundžių ($N = 301$).

Sudarant testinę aibę atsitiktinai be grąžinimo išrinkti 2000 tirpalų koncentracijų vektorių, likę 2096 vektoriai sudarė apmokymo aibę. Bandymo būdu buvo parinkti pagrindinių komponenčių skaičius bei meta parametras λ. Šiuo atveju $\lambda = 5$ ir pagrindinių komponenčių skaičius 300 atveju kai BA ($d = 0,02$), 500 kai BA ($d = 0,04$), 82 kai FIA ($d = 0,02$) ir 84 kai FIA ($d = 0,04$). Nustačius šiuos parametrus ir pasirinkus tikslumo intervalus Δ (pateikti 1 lentelėje) apskaičiuoti klasifikacijos tikslumai pagal (2.3), pateikti 2 lentelėje.

Lyginant rezultatus gautus dirbtinių neuroninių tinklų pagalba (pateikti 3 lentelėje) ir gautus panaudojus apibendrintą pagrindinių komponenčių regresiją (2 lentelė) matome, kad vonios režimu klasifikavimo rezultatai gauti APKR yra gerokai prastesni tais atvejais kai membranos storis $d = 0,02$ ir $L = 3$ ir $L = 4$ skirtumai yra apie 8 ir 39 procentinių punktų atitinkamai, tačiau kai membranos storis $d = 0,04$ šie skirtumai sumažėja iki 2 ir 30 procentinių punktų. Apipurškimo režimo (FIA) rezultatatai yra vienodai geri.

1 lentelė. Tikslumo intervalai Δ prognozuojamoms koncentracijoms

y (nmol/cm^3)	1	2	4	8	12	16	32	64
$\Delta_1 y$ (nmol/cm^3)	<1,5	[1,5; 3)	[3; 6)	[6; 10)	[10; 14)	[14; 24)	[24; 48)	$\geqslant 48$
$\Delta_2 y$ (nmol/cm^3)	[0; 1,5)	[1,5; 2,9)	[3,1; 5)	[7; 9)	[11; 13)	[15; 17)	[31; 33)	[63; 65)

2 lentelė. Klasifikavimo tikslumas vonios ir apipurškimo režimais, naudojant intervalus Δ_1 ir Δ_2 bei APKR

L	BA, Δ_1, $d = 0,02$		FIA, Δ_2, $d = 0,02$		BA, Δ_1, $d = 0,04$		FIA, Δ_2, $d = 0,04$	
	Apm. aibė	Test. aibė	Apm. aibė	Test. aibė	Apm. aibė	Test. aibė	Apm. aibė	Test. aibė
1	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
2	100,00	99,20	100,00	100,00	100,00	100,00	100,00	100,00
3	92,46	91,40	100,00	100,00	99,71	98,00	100,00	100,00
4	53,10	48,35	100,00	99,85	78,87	68,70	100,00	100,00

3 lentelė. Klasifikavimo tikslumas vonios ir apipurškimo režimais, naudojant intervalus Δ_1 ir Δ_2 bei dirbtinius neuroninius tinklus

L	BA, Δ_1, $d = 0,02$		FIA, Δ_2, $d = 0,02$		BA, Δ_1, $d = 0,04$	
	Apm. aibė	Test. aibė	Apm. aibė	Test. aibė	Apm. aibė	Test. aibė
1	100,00	100,00	100,00	100,00	99,90	100,00
2	100,00	100,00	100,00	100,00	99,80	99,80
3	99,76	99,60	100,00	100,00	100,00	100,00
4	87,97	86,95	99,85	99,75	99,95	99,90

4. Išvados

Naudojant APKR kaip tiesinį klasifikatorių klasifikuojant biojutiklio signalą gaunami geri rezultatai ir išvengiama problemų su kuriomis susiduriame konstruojant neuroninį tinklą, t.y. architekūros parinkimas, apmokymo algoritmo subtilumai, tikslo funkcijos lokalieji ekstremumai. Duomenų masyvą sudarė atsako kreivės taškai, kurių yra BA atveju kai ($d = 0,02$) 301, BA atveju ($d = 0,04$) 501 ir FIA atveju kai ($d = 0,02$) 151, FIA atveju ($d = 0,04$) 301. Ateityje bus ieškomi esminiai atsako kreivių taškai, kurie turi didžiausią įtaką koncentracijų vertinimui, tokiu būdu bus stengiamasi sumažinti triukšmo įtaką realių duomenų vektoriams.

Literatūra

1. U. Wollenberger, F. Lisdat, F.W. Scheller, *Frontiers in Biosensorics 2, Practical Applications*, Birkhauser Verlag, Basel (1997).
2. R. Baronas, J. Christensen, F. Ivanauskas, J. Kulys, Computer simulation of amperometric biosensor response to mixtures of compounds, *Nonlinear Analysis*: *Modelling and Control*, **7**(2), 3–14 (2002).
3. R. Baronas, F. Ivanauskas, R. Maslovskis, P. Vaitkus, An analysis of mixtures using amperometric biosensors and artificial neural networks, *Journal of Mathematical Chemistry*, **36**(3), 281–297 (2004).
4. Y. Tan, L. Shi, W. Tong, Charles Wang, Multi-class cancer classification by total principal component regression (TPCR) using microarray gene expression data, *Nucleic Acids Res.*, **33**(1), 56–65 (2005).

SUMMARY

R. Baronas, F. Ivanauskas, R. Paulauskas, P. Vaitkus. The classification of concentration of mixture of analytes using total principal component regression

In this paper total principal component regression is used for biosensors response to mixtures of compounds classification. The results are compared with the results obtained using artificial neural networks.

Keywords: total principal component regression, pseudo inverse matrix.