Liet. matem. rink, **46***, spec. nr.*, 2006, 338–344

Biojutiklių atsako į mišinius statistinė analizė ir prognozavimas

Romas BARONAS (VU), Sigitas BŪDA (VU), Feliksas IVANAUSKAS (VU, MII), Pranas VAITKUS (VU)
el. paštas: s.buda@it.lt

Reziumė. Darbe nagrinėjami elektrocheminiai biojutikliai, jų reakcija į skirtingų cheminių komponentų mišinius. Pagrindinis darbo tikslas – išanalizavus skirtingų tipų biojutiklius ir jų galimybes atskirti mišiniuose esančius komponentus, sukurti matematinį modelį, kurio pagalba pagal biojutiklio atsako duomenis būtų galima prognozuoti mišinį sudarančių komponentų koncentracijų reikšmes. Prognozavimo matematinis modelis kuriamas neuroninių tinklų pagalba. Papildomai atliekama pagrindinių komponenčių ir daugiamatė dispersinė analizė. Šios analizės metu parenkama kokio tipo biojutiklis, kokiame matavimo režime geriausiai atskiria mišinio komponentus ir tinka komponentų koncentracijų prognozavimui. Darbo pabaigoje sukuriamas sudėtinis neuroninių tinklų medis, skirtas mišinių komponentų koncentracijoms prognozuoti.

Raktiniai žodžiai: biojutiklis, matematinis modelis, neuroniniai tinklai.

Tyrimo objektas ir darbo tikslas

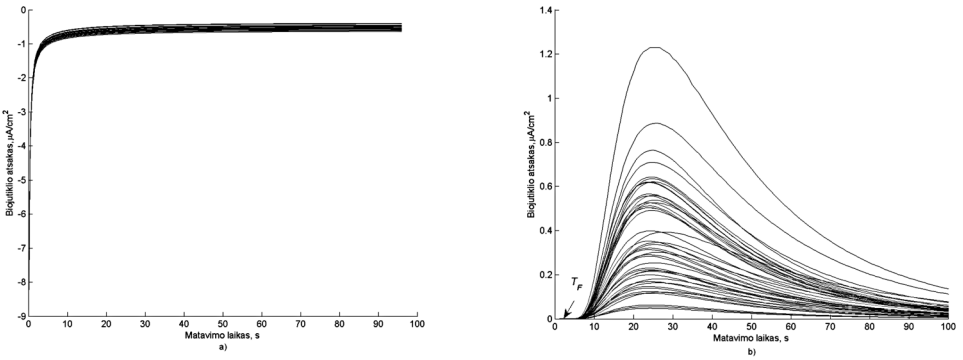
Tyrimo objektas – elektrocheminiai biojutikliai [1–3], tiriama kaip į skirtingų koncentracijų komponentų mišinius reaguoja dviejų tipų potenciometriniai (Po) ir amperometriniai (Am) biojutikliai. Kiekvieno tipo biojutiklių atsakas į mišinį buvo matuojamas dviem skirtingais režimais: vonios (BAT), kai biojutiklis viso atsako matavimo metu yra laikomas analizuojamame tirpale ir apipurškimo (FIA), kai biojutiklis trumpam yra panardinamas į tirpalą [4].

Pagrindinis šio darbo tikslas sukurti matematinį modelį, kuris pagal biojutiklio atsako į mišinį duomenis galėtų pakankamai tiksliai prognozuoti mišinį sudarančių komponentų koncentracijas. Ankstesniuose darbuose [6] visų komponentų koncentracijoms prognozuoti buvo naudojamas vienas bendras kelių sluoksnių neuroninis tinklas. Šiame darbe siūloma kiekvieno komponento koncentracijai prognozuoti kurti atskirus neuroninius tinklus.

Duomenys

Eksperimentų metu naudoti dirbtinai generuoti duomenys. Biojutiklio atsako į mišinius duomenys buvo gauti pasinaudojant matematiniu biojutiklio modeliu [5–6].

Tyrimo metu nagrinėjami $S_1, \ldots, S_K$ komponentų mišiniai. Tyrimui buvo pasirinktas keturių ($K = 4$) komponentų mišinys. Generuojant duomenis kiekvieno komponento koncentracija mišinyje atitiko vieną iš aštuonių ($M = 8$) galimų skirtingų koncentracijų reikšmių:



1 pav. Biojutiklio atsako į skirtingų koncentracijų mišinius kreivės: a) Po biojutiklis, b) Am biojutiklis. Matuota FIA režime, kontakto su mišiniu trukmė $T_F = 2$ sekundės.

$$S_k \in \left\{ S_{k,m} : S_{k,m} = \alpha_m \text{nmol/cm}^3,\ m = 1, ..., M \right\}, \quad k = 1, ..., K,$$

$$\alpha_1 = 1,\ \alpha_2 = 2,\ \alpha_3 = 4,\ \alpha_4 = 8,\ \alpha_5 = 12,\ \alpha_6 = 16,\ \alpha_7 = 32,\ \alpha_8 = 64. \quad (1)$$

Tokiu būdu buvo gauta $M^K = 8^4 = 4096$ skirtingų mišinių. Matematinio modeliavimo pagalba sugeneruoti duomenys yra biojutiklio atsako į šiuos mišinius kreivės taškai.

Uždavinys

Tegul $\vec{c} = (c_1, \ldots, c_K)$ yra mišinio komponentų koncentracijų reikšmių vektorius, $\vec{z}(\vec{c}) = (z_1(\vec{c}), \ldots, z_L(\vec{c})) \in \Re^L$ – biojutiklio atsako į mišinį reikšmės laiko matavimo momentais $t_1, \ldots, t_L$. Pagrindinis tikslas yra sukurti matematinį modelį F, kurio pagalba pagal biojutiklio atsako taškus būtų galima prognozuoti mišinio komponentų koncentracijas, t.y.:

$$F(\vec{z}) = \vec{c}. \quad (2)$$

Matematinio modelio kūrimui buvo pasirinkti dirbtiniai neuroniniai tinklai – daugiasluoksniai perceptronai (Multi-Layer Perceptron, sutr. MLP) [7–8]. Naudojant MLP, uždavinio sprendimas susiveda į tai, kad reikia parinkti neuronų aktyvacijos funkcijas, neuronų skaičių paslėptame sluoksnyje, rasti optimalias MLP svorių reikšmes taip, kad padavus MLP įėjimo sluoksniui biojutiklio atsako reikšmių vektorių $\vec{z}$, tinklo generuojamas išėjimas y_k kuo tiksliau prognozuotų komponento S_k koncentraciją c_k. Optimalios MLP svorių reikšmės yra randamos tinklo apmokymo metu. Jei biojutiklio atsako duomenų vektoriaus ilgis yra L, tai kiekvieno MLP įėjimo sluoksnio neurono svorių skaičius bus lygus $L + 1$. Per didelis svorių skaičius apsunkina MLP apmokymą, tuo pačiu pablogina prognozės kokybę.

Pagrindinių komponenčių analizė

Siekiant sumažinti biojutiklio atsako matavimo reikšmių vektoriaus ilgį (tuo pačiu ir svorių skaičių MLP įėjimo sluoksnyje), duomenys buvo transformuojami atliekant pa-

grindinių komponenčių analizę (Principal Component Analysis, sutr. PCA) [9–10]. PCA vienas iš metodų, kurio pagalba galima sumažinti duomenų erdvės matavimų skaičių minimaliai prarandant informaciją. Šio tyrimo metu tolimesnei analizei buvo atrinktos pagrindinės komponentės, kurių bendra dispersija viršija 95% originalių duomenų dispersijos.

PCA metu išskirtoms pagrindinėms komponentėms papildomai buvo atliekama daugiamatė dispersinė analizė (Multivariate Analysis of Variance, sutr. MANOVA) [11]. Šios analizės tikslas yra įvertinti kiek skiriasi biojutiklio atsakas į skirtingų komponentų koncentracijas. Tam tikslui PCA komponenčių duomenys buvo suskirstyti į keturias nesikertančias grupes. Kiekvieną G_k grupę sudaro mišiniai, kuriuose vieno iš komponentų S_k koncentracija c_k yra didesnė, likusių komponentų koncentracijos yra sąlyginai nedidelės, tokiu būdu stengiamasi kiekvienoje grupėje išskirti vieną iš komponentų, maksimaliai eliminuojant likusių komponentų įtaką:

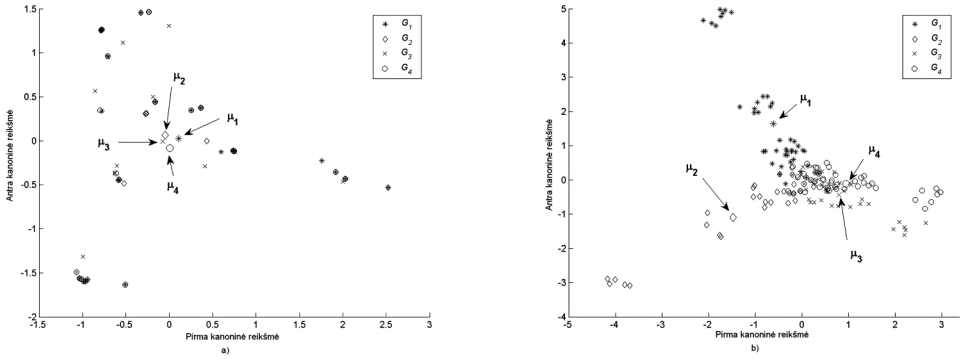
$$G_k \in \{\vec{c} = (c_1, \ldots, c_K) : \; c_k \geq \alpha_1 \cap \mathrm{c_n} \leq \alpha_2, \; j = 1, ..., K, \; n \neq k\}, \quad k = 1, ..., K,$$
$$\alpha_1 = 8\mathrm{nmol/cm}^3, \; \alpha_2 = 4\mathrm{nmol/cm}^3. \tag{3}$$

MANOVA analizės metu buvo matuojamas Mahalanobis'o atstumas D_M tarp grupių $G_1, \ldots, G_K$ vidurkių $\mu_1, \ldots, \mu_K$ ir tikrinama nulinė hipotezė H_0, kad artimiausių pagal atstumą grupių vidurkiai yra lygūs, t.y. praktiškai tikrinama ar biojutiklis vienodai reaguoja į skirtingus komponentus. Gauti MANOVA analizės rezultatai yra pateikiami 1 lentelėje.

Analizuojant gautus MANOVA rezultatus galima daryti prielaidą, kad pagal Po biojutiklio atsaką apskritai gali būti sudėtinga atskirti, kurio iš komponentų koncentracija tirpale yra dominuojanti: atstumas tarp gretimų grupių vidurkių artimas nuliui, visais atvejais neatmestinos hipotezės apie artimiausių grupių vidurkių lygybę. Tikėtina, kad pagal Po biojutiklio atsaką konkrečios komponento koncentracijos reikšmės prognozavimas apskritai yra sunkiai įmanomas. Am biojutiklio

1 lentelė. PCA ir MANOVA analizės metu gauti rezultatai: T_F – biojutiklio kontakto su mišiniu FIA režime trukmė sekundėmis, L – biojutiklio atsako matavimų skaičius, L_{PCA} – PCA analizės metu iš biojutiklio atsako išskirtų pagrindinių komponenčių skaičius, D_M – Mahalanobis'o atstumas iki artimiausios grupės, H_0 – hipotezės, apie artimiausių grupių vidurkių lygybę (reikšmingumo lygmuo $P = 0.05$), tikrinimo rezultatas: 0 – hipotezė neatmestina, 1 – hipotezė atmestina

Matavimo režimas	Tipas	T_F	L	L_{PCA}	G_1		G_2		G_3		G_4	
					D_M	H_0	D_M	H_0	D_M	H_0	D_M	H_0
BAT	Am	–	351	4	6.103	1	6.265	1	5.716	1	5.716	1
	Po	–	480	1	0.001	0	0.001	0	0.007	0	0.001	0
FIA	Am	2	121	4	6.692	1	6.692	1	7.510	1	7.778	1
		5	121	4	5.451	1	5.859	1	5.451	1	5.748	1
		10	121	4	5.931	1	5.931	1	6.306	1	6.023	1
	Po	2	120	4	0.010	0	0.015	0	0.023	0	0.009	0
		5	120	3	0.001	0	0.005	0	0.018	0	0.001	0
		10	120	2	0.006	0	0.016	0	0.006	0	0.009	0



2 pav. MANOVA analizės metu išskirtų pirmųjų dviejų kanoninių reikšmių grafikai: a) Po biojutiklis, b) Am biojutiklis. Matuota FIA režime, kontakto su mišiniu trukmė $T_F = 2$ sekundės. Rodyklėmis pažymėti skirtingų grupių vidurkių taškai.

atveju hipotezės apie grupių vidurkių lygybę atmestinos – Am biojutiklio atsakas į skirtingus komponentus skiriasi.

Akivaizdus skirtumas tarp skirtingų tipų biojutiklių atsako į mišinius matyti 2 pav., kuriame pateikiama MANOVA metu iš PCA komponenčių išskirtų pirmųjų dviejų kanoninių reikšmių grafikai (atsakas matuotas FIA režime, kontakto trukmė $T_F = 2$ sekundės). Kairėje paveikslėlio pusėje (2 pav., a) matyti, kad Po biojutiklio atsakas į skirtingus komponentus praktiškai nesiskiria – skirtingų grupių G_k kanoninių reikšmių taškai persidengia, visų grupių vidurkiai $\mu_k$artimi nuliui. Pagal Am biojutiklio atsaką (2 pav., b) skirtingų komponentų grupės atsiskiria, grupių vidurkiai nutolę vienas nuo kito.

Atsižvelgiant į PCA ir MANOVA analizės metu gautus rezultatus ir į didžiausius atstumus tarp skirtingų mišinių grupių, komponentų koncentracijoms prognozuoti buvo pasirinktas Am tipo biojutiklis, atsakas matuojamas FIA režime, biojutiklio kontakto su mišiniu trukmė $T_F = 2$ sekundės.

Tirpalo komponentų koncentracijų prognozavimas pagal biojutiklio atsaką

Visų prognozavimo eksperimentų metu uždavinio sprendimui naudotas dvisluoksnis perceptronas (MLP) su sigmoidinė aktyvacijos funkcija $\varphi(u) = 1/(1 + \exp(-u))$. MLP neuronų skaičius paslėptame sluoksnyje buvo parenkamas eksperimentinių bandymų metu.

Mišinio komponentų koncentracijų ir biojutiklio atsako duomenų aibė buvo padalinta į dvi dalis: tinklo apmokymo ir testavimo aibes. Apmokymo aibei iš turimų 4096 skirtingų mišinių atsitiktiniu būdu atrinkta $N_1 = 2000$ mišinių duomenys, likę $N_2 = 2096$ buvo priskirti testinei aibei. MLP įėjimo sluoksniui buvo paduodama iš biojutiklio atsako duomenų išskirtos pagrindinių komponenčių reikšmės. MLP buvo apmokomas naudojant Levenberg–Marquart atgalinės eigos (Back Propagation, sutr. BP) algoritmą [7,8]. Apmokymo metu MLP mokomas fiksuotą iteracijų skaičių.

Apmokyto MLP prognozės tikslumas buvo vertinamas pasinaudojant testavimo aibės duomenimis. Visų pirma įvertinama tinklo išėjimo reikšmės y_k ir prognozuo-

jamos komponento koncentracijos c_k vidutinė absoliutinė paklaida:

$$E_k = \frac{1}{N_2}\sum_{i=1}^{N_2} |y_{k,i} - c_{k,i}|, \quad \text{čia } N_2 \text{ – testavimo aibės tūris.} \tag{4}$$

Kitas pasirinktas tikslumo matas Q_k – procentinis skaičius stebėjimų, kuriems absoliutinė paklaida viršijo $\Delta c = 0.5\,\text{nmol/cm}^3$:

$$Q_k = \frac{1}{N_2}\sum_{i=1}^{N_2} Ind\left(y_{k,i} \in \left[c_{k,i} - \Delta c; c_{k,i} + \Delta c\right]\right). \tag{5}$$

Pirmųjų bandymų etapo (I) metu buvo bandoma sukurti vieną MLP, kuris pagal biojutiklio atsako pagrindinių komponenčių reikšmes prognozuotų visų keturių komponentų koncentracijas mišinyje. Toks MLP turi keturis neuronus išėjimo sluoksnyje, tinklą bandoma apmokyti taip, kad kiekvieno iš šių išėjimo sluoksnio neuronų atsakas prognozuotų konkretaus komponento koncentraciją. Gautas rezultatas yra pateikiamas 2 lentelės I etapo rezultatų skiltyje. Kaip matome gerai sekasi prognozuoti S_1 ir S_2 komponentų koncentracijas – prognozės atvejų skaičius, kai viršijama toleruotina $\Delta c = 0.5\,\text{nmol/cm}^3$ paklaidos riba, nesiekia net 1% ($Q_2 = 0.04\%$). Likusių dviejų S_3 ir S_4komponentų prognozės rezultatai žymiai blogesni – S_3 komponentui beveik 100% atvejų skirtumas tarp MLP generuojamos prognozės ir realios koncentracijos viršijo toleruotiną ribą ($Q_3 = 99.72\%$).

Antrojo (II) etapo metu buvo sukurti keturi atskiri neuroniniai tinklai, kiekvienas iš jų skirtas konkretaus komponento koncentracijai mišinyje prognozuoti. Iš gautų rezultatų matyti, kad žymiai pagerėjo S_3 ir S_4 komponentų koncentracijų prognozė. Skaičius atvejų, kai S_3 komponento prognozės paklaida viršijo toleruotiną Δc ribą, nuo 100% sumažėjo iki $Q_3 = 3.41\%$! Verta atkreipti dėmesį į tai, kad prognozuojant tik vieno komponento koncentraciją, MLP paslėptame sluoksnyje užtenka mažesnio kiekio neuronų.

Lyginant MLP prognozę su tikrąja tiriama komponento koncentracija buvo pastebėta, kad visų komponentų prognozės tikslumas mažėja didėjant S_3 ir S_4 komponentų koncentracijoms. Tikėtina, kad MLP įėjimui paduodant papildomą informaciją apie S_3 ir S_4 koncentracijas mišinyje, komponentų prognozės tikslumas turėtų pagerėti. Trečio (III) etapo metu buvo sukurti papildomi atskiri neuroniniai tinklai, kurie kiekvienam mišiniui prognozuoja ar S_3 ir S_4 komponentų koncentracijos mišinyje neviršija $8\,\text{nmol/cm}^3$. Gauti prognozės rezultatai buvo prijungti prie pradinės MLP apmokymo aibės – biojutiklio atsako pagrindinių komponenčių duomenų. Naujai suformuota apmokymo aibė buvo panaudota III etapo metu kuriamiems MLP apmokyti. Kaip matome iš 1 lentelės III skiltyje pateiktų rezultatų, papildoma informacija apie S_3 ir S_4 komponentų koncentracijas pagerino visų komponentų prognozės tikslumą – didžiausia vidutinė prognozės paklaida $E_3 = 0.067\,\text{nmol/cm}^3$ apie 8 kartų mažesnė už toleruotiną Δc ribą; didžiausia buvusi Q_3reikšmė sumažėjusi iki $Q_3 = 1.81\%$.

Siekiant dar pagerinti S_3 ir S_4 komponentų prognozės tikslumą, ketvirto (IV) etapo metu tinklo apmokymo aibė buvo praplėsta pridedant S_1 ir S_2 komponentų koncen-

2 lentelė. Prognozavimo eksperimentų rezultatai. Čia h – neuronų skaičius paslėptame MLP sluoksnyje, E_k – vidutinė absoliutinė prognozės paklaida nmol/cm^3, Q_k – procentinis skaičius stebėjimų, kuriems absoliutinė paklaida viršijo $\Delta c = 0.5\,\text{nmol/cm}^3$.

S_k	I etapas			II etapas			III etapas			IV etapas		
	h	E_k	Q_k	h	E_k	Q_k	h	E_k	Q_k	h	E_k	Q_k
S_1	10	0.113	0.00	4	0.020	0.00	6	0.015	0.00	–	–	–
S_2	10	0.163	0.04	4	0.062	0.00	6	0.043	0.00	–	–	–
S_3	10	2.142	99.72	6	0.148	3.41	6	0.067	1.81	6	0.045	0.00
S_4	10	1.139	80.26	6	0.128	1.32	6	0.071	0.08	6	0.039	0.04

tracijos mišinyje prognozės rezultatus. Kaip matome, apmokius MLP, S_3 ir S_4 komponentų prognozės tikslumas praktiškai pasiekė S_1 ir S_2 komponentų prognozės lygį. Komponento S_3, kurio koncentraciją ankstesnių etapų metu sunkiausiai sekėsi prognozuoti, vidutinė prognozės paklaida sumažėjo iki $E_3 = 0.045\,\text{nmol/cm}^3$, skaičius atvejų, viršijančių Δc ribą $Q_3 = 0.00\%$.

Darbo išvados

Darbo metu buvo įvertinti dviejų skirtingų tipų – amperometriniai (Am) ir potenciometriniai (Po) biojutikliai. Pagrindinių komponenčių ir daugiamatės dispersinės analizės metu buvo nustatyta, kad Po biojutikliai prasčiau nei Am analogai atskiria mišinyje esančius komponentus, pagal Po biojutiklio atsaką komponentų koncentracijos reikšmių prognozavimas yra sunkiai įmanomas.

Matematinis komponentų koncentracijos prognozavimo modelis sukurtas remiantis Am biojutiklio atsaku. Geriausi prognozės rezultatai gauti sukūrus atskirus neuroninius tinklus kiekvieno komponento koncentracijai prognozuoti. Iš to galima daryti prielaidą, kad kiekvieno konkretaus komponento koncentracijai tirpale nustatyti turėtų būti kuriami atskiri biojutikliai.

Prognozuojant vieno komponento koncentraciją mišinyje labai svarbi yra informacija apie kitų komponentų koncentracijas. Žinant vienų komponentų kiekį mišinyje, likusių komponentų koncentracijos prognozė gaunasi tikslesnė.

Matematinis biojutiklio modeliavimas, jo atsako statistinė analizė gali padėti kuriant realius biojutiklius: parinkti biojutiklio tipą, matavimo režimą, įvertinti atsako matavimo trukmę, kontakto su mišiniu laiką, enzimo membranos storį ir panašiai.

Literatūra

1. L.C. Clarc and C. Loys, *Ann. N.Y. Acad. Sci.*, **102** (1962).
2. A.P.F. Turner, I. Karube and G.S. Wilson, *Biosensors: Fundamentals and Applications*, Oxford University Press, Oxford (1987).
3. F. Scheller and F. Schubert, *Biosensors*, Elsevier, Amsterdam (1992).
4. J. Ruzicka and E.H. Hansen, *Flow Injection Analysis*, John Wiley and Sons, New York (1988).
5. 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).
6. R. Baronas, F. Ivanauskas, R. Maslovskis, P. Vaitkus, An analysis of mixtures using amperometric biosensors and artificial neural networks, *Journal of Mathematical Chemistry*, **36** (2004).

7. Š. Raudys, *Statistical and Neural Classifiers: an Integrated Approach to Design*, Springer, London (2001).
8. S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed., Prentice Hall, New York (1999).
9. J.R. Llinas and J.M. Ruiz, in: *Computer Aids to Chemistry*, G. Vemin and M.Chanon (Eds.), John Wiley, New York (1986).
10. H. Martens and T. Nes, *Multivariate Calibration*, Wiley, Chichester (1989).
11. W.J. Krzanowski, *Principles of Multivariate Analysis*, Oxford University Press (1988).

SUMMARY

R. Baronas, S. Būda, F. Ivanauskas, P.Vaitkus. Biosensor response to multi-component mixtures statistical analysis and forecasting

This paper deals with an analysis of the electrochemical biosensors and their response to multi-component mixtures. The main task is to build a mathematical model for estimation the concentration of each mixture component from the biosensor response data. Two different types of biosensors: amperometric and potenciometric are analysed. Due to high dimensionality of biosensor output data the principal component analysis is applied. Additional multivariate analysis of variance is used to analyze the response sensitivity of each biosensor type. Finally a concentration estimation model based on ensemble of neural networks is presented.

Keywords: biosensor, modelling, neural networks.