Liet. mat. rink. LMD darbai, **48/49**, 2008, 354–359

Neuronų skaičiaus parinkimas vektorių kvantavimo metoduose

Olga KURASOVA, Alma MOLYTĖ (MII)
el. paštas: kurasova@ktl.mii.lt, alma.molyte@gmail.com

Reziumė. Darbe nagrinėjama neuronų skaičiaus parinkimo vektorių kvantavimo metoduose strategija. Analizuojami du neuroniniais tinklais pagrįsti metodai: saviorganizuojantys neuroniniai tinklai ir neuroninės dujos. Pasiūlytas būdas, pagal kurį parenkamas neuronų skaičius atsižvelgiant į analizuojamų duomenų specifiką.

Raktiniai žodžiai: vektorių kvantavimas, saviorganizuojantys neuroniniai tinklai, neuroninės dujos, kvantavimo paklaida.

1. Įvadas

Vektorių kvantavimas – tai procesas, kurio metu n-mačiai duomenų aibės vektoriai $X_1, X_2, \ldots, X_m$, čia m – vektorių skaičius, yra pakeičiami mažesniu kiekiu n-mačių vektorių $M_1, M_2, \ldots, M_N$, $N < m$. Dažniausiai vektorių kvantavimo metodai taikomi garsui ir vaizdui suspausti, tačiau jie tinka ir duomenims klasterizuoti bei klasifikuoti. Prie šių metodų grupės priskiriami saviorganizuojantys neuroniniai tinklai [4], vektorinio mokymo kvantavimas [4], neuroninių dujų metodas [5] ir kt. Kvantavimo rezultatui įvertinti skaičiuojama kvantavimo paklaida, ji yra mažiausia, kai $N = m$, tačiau kvantavimo metodų tikslas – sumažinti N. Yra problema nustatyti, kokia turi būti N reikšmė, kad rezultatas būtų priimtinas sprendžiamam uždaviniui. Šiame darbe nagrinėjami neuroniniais tinklais grindžiami vektorių kvantavimo metodai – saviorganizuojantis neuroninis tinklas (*angl. self-organizing map*) ir neuroninių dujų (*angl. Neural Gas*) metodas. Juose kvantuotų vektorių skaičius N vadinamas neuronų skaičiumi. Tyrimo tikslas – nustatyti neuronų skaičių N atsižvelgiant į analizuojamų duomenų specifiką.

2. Vektorių kvantavimo metodai

Tegul duomenų aibės matrica $X = \{X_1, X_2, \ldots, X_m\} = \{x_{ij},\ i = 1, \ldots, m,\ j = 1, \ldots, n\}$, jos eilutės yra vektoriai $X_i \in R^n$, t.y. $X_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, $i = 1, \ldots, m$, čia x_{ij} yra i-tojo vektoriaus j-toji komponentė, n – komponenčių (matmenų) skaičius, m – analizuojamų vektorių skaičius. Tiek neuroninių dujų (ND) metodu [5], tiek saviorganizuojančiu neuroniniu tinklu (SOM) [4] sukuriamas neuronų masyvas M. Neuronai – tai vektoriai, kurių matmenų skaičius lygus n. ND metode neuronų tinklas yra vienmatis $M = \{M_1, M_2, \ldots, M_N\}$, čia $M_k = \{m_{k1}, m_{k2}, \ldots, m_{kn}\}$, $k = 1, \ldots, N$, N – neuronų skaičius. SOM tinklas yra dvimatis $M = \{M_{ij}, i = 1, \ldots, r,$

$j = 1, \ldots, s\}$, čia $M_{ij} = (m_1^{ij}, m_2^{ij}, \ldots, m_n^{ij})$, r yra eilučių skaičius, s – stulpelių, neuronų skaičius $N = r \times s$. Metodų tikslas – pakeisti neuronų reikšmes taip, kad jie atspindėtų analizuojamos duomenų aibės vektorių X_i, $i = 1, \ldots, m$, savybes, t.y. mokymo pabaigoje neuronai tampa vektorių X_i kvantuotais vektoriais.

Prieš tinklo mokymą generuojamos atsitiktinės pradinės neuronų komponenčių reikšmės intervale $(-0,5 \cdot 10^{-5}, 0,5 \cdot 10^{-5})$ (ND) arba $(0, 1)$ (SOM). Mokymo metu vienas po kito mokymo aibės X vektoriai pateikiami į tinklą nustatytą kiekį kartų. Kiekvienas vektorius į tinklą pateikiamas $\hat{e}$ kartų. Kadangi analizuojamų vektorių skaičius yra lygus m, tai mokymo iteracijų skaičius $t_{\max} = \hat{e} \times m$. Į tinklą pateikus vektorių X_l, $l \in \{1, \ldots, m\}$ suskaičiuojamas Euklido atstumas nuo jo iki visų neuronų.

ND metode neuronai $M_1, M_2, \ldots, M_N$ pakeičiami neuronais $W_1, W_2, \ldots, W_N$, čia $W_k \in \{M_1, M_2, \ldots, M_N\}$, $k = 1, \ldots, N$, taip, kad

$$\|W_1 + X_l\| \leqslant \|W_2 + X_l\| \leqslant, \ldots, \leqslant \|W_N + X_l\|.$$

Tada atstumas nuo X_l iki pirmo neurono W_1 yra mažiausias. Šis neuronas vadinamas neuronu nugalėtoju. Visų neuronų reikšmės keičiamos pagal formulę:

$$W_k(t+1) = W_k(t) + E(t) \cdot h_\lambda \cdot \big(X_l - W_k(t)\big), \tag{1}$$

čia t yra iteracijos numeris, $E(t) = E_i(E_f/E_i)^{(t/t_{\max})}$, $h_\lambda = \mathrm{e}^{-(k-1)/\lambda(t)}$, $\lambda(t) = \lambda_i(\lambda_f/\lambda_i)^{(t/t_{\max})}$, parametrų λ_i, λ_f, E_i, E_f reikšmės parenkamos prieš tinklo mokymą.

SOM mokyme į tinklą pateikus vektorių X_l, suskaičiuojamas Euklido atstumas nuo jo iki visų tinklo neuronų, randamas neuronas nugalėtojas $\hat{M}$, iki kurio atstumas nuo X_l yra mažiausias. Neuronų reikšmės keičiamos pagal formulę:

$$M_{ij}(t+1) = M_{ij}(t) + h_{ij}(t) \cdot \big(X_l - M_{ij}(t)\big), \tag{2}$$

čia t yra iteracijos numeris, $h_{ij}(t)$ taip vadinama kaimynystės funkcija, kurios reikšmė priklauso nuo vykdomos iteracijos numerio t ir perskaičiuojamo neurono vietos tinkle neurono nugalėtojo atžvilgiu. Procesui konverguoti būtina, kad $h_{ij}(t) \to 0$, kai $t \to \infty$.

Kai tinklas išmokytas, būtina įvertinti jo kokybę. Vektorių kvantavimo metoduose dažniausiai vertinama kvantavimo paklaida, apskaičiuojama pagal formulę

$$E_q = \frac{1}{m} \sum_{l=1}^{m} \|X_l - \hat{M}\|, \tag{3}$$

čia $\hat{M}$ yra vektoriaus X_l neuronas nugalėtojas, ND metode $\hat{M} = W_1$.

3. Analizuojami duomenys

Nagrinėti šie realaus pobūdžio duomenys [1]:

- *Fišerio irisų duomenys*. Yra išmatuota trijų veislių irisų: taurėlapių ilgis ir plotis, vainiklapių ilgis ir plotis. Sudaryti 4-mačiai vektoriai, $n = 4$, $m = 149$.

- *Automobilių*, pagamintų JAV, Europoje ir Japonijoje, duomenys. Automobilius charakterizuoja degalų sunaudojimas, cilindrų skaičius, variklio darbo tūris, arklio jėgų kiekis, svoris, greitis, pagaminimo metai. Sudaryti 7-mačiai vektoriai, $n = 7$, $m = 228$.
- *Kviečių duomenys*. Analizuotos penkios kviečių rūšys. Paimta kiekvieno grūdo skaitmeninė nuotrauka, išmatuota 12 parametrų: grūdo plotas, perimetras, spalvinės charakteristikos ir kt. Sudaryti 12-mačiai vektoriai, $n = 12$, $m = 400$.

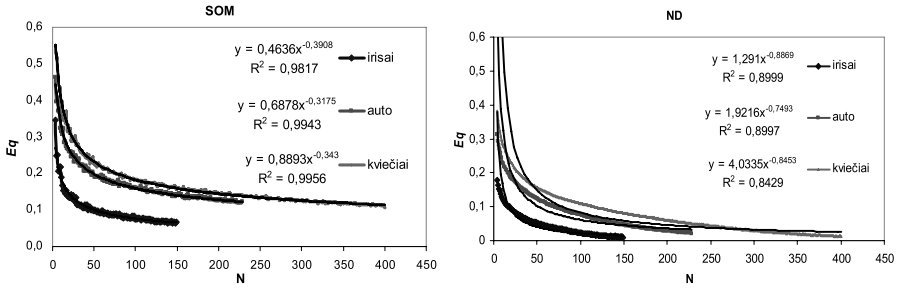
Taip pat tyrime naudoti dirbtinai sugeneruoti įvairaus didumo masyvai, sudaryti iš vektorių, kurių komponentės yra dydžiai, tolygiai pasiskirstę intervale (0, 1).

4. Eksperimentinio tyrimo rezultatai

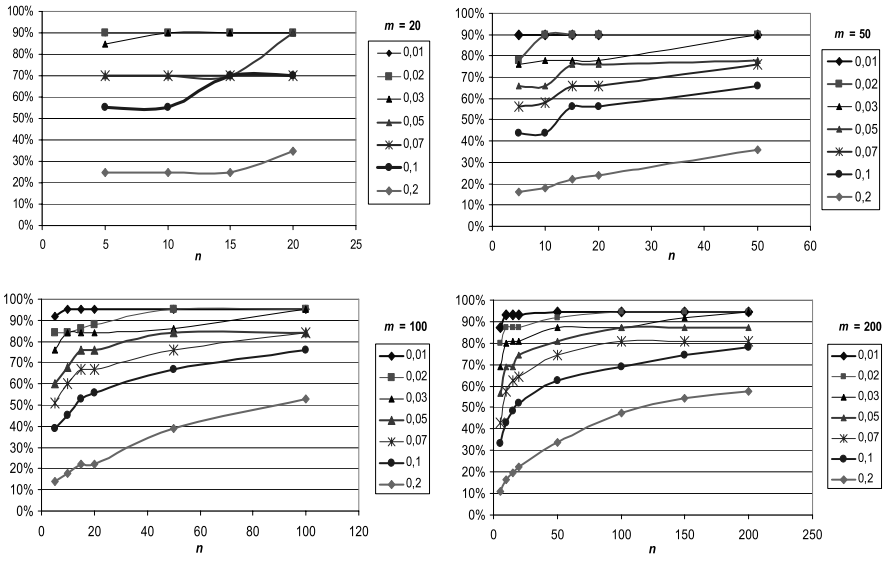
ND metodo rezultatai priklauso nuo mokymo parametrų $\lambda_i, \lambda_f, E_i, E_f$, mokymo žingsnių skaičiaus $\hat{e}$ bei neuronų skaičiaus N. Teorinių sprendinių konvergavimo įrodymų nėra, todėl norint rasti tinkamiausią sprendinį, parametrus reikia parinkti empiriškai. Darbe [6] nustatyta, kad geriausi rezultatai kvantavimo paklaidos prasme yra gaunami, kai $E_f = 0,1,\ \lambda_f = 0,01$, parametrų E_i, λ_i reikšmės imtos, kaip nustatyta [2], t.y. $E_i = 0,5,\ \lambda_i = N/2$. Taip pat darbe [6] ištirta, kad pakankamai stabilūs rezultatai yra gaunami, kai $\hat{e} = 200$. Didinti šį skaičių nėra prasmės, nes kvantavimo paklaida sumažėja nežymiai.

Tiek ND, tiek SOM metode mažiausia kvantavimo paklaida $E_q(3)$ gaunama, kai neuronų skaičius $N = m$, m – analizuojamų vektorių skaičius. Tačiau nuo tam tikros $N = N'$ reikšmės paklaida skiriasi nežymiai, palyginus su mažiausia. Kvantavimo paklaidų kitimo grafikai trims skirtingų matmenų duomenų aibėms pateikti 1 pav. Mažiausia kvantavimo paklaida gaunama analizuojant irisų duomenis ($m = 149$, $n = 4$), didesnė – automobilių ($m = 228$, $n = 7$), didžiausia – kviečių ($m = 400$, $n = 12$). Be to, pastebėta, kad paklaidos kitimo grafikas yra gana tiksliai aproksimuojamas laipsnine funkcija $y = ax^b$ su neigiamu laipsniu $b < 0$, aproksimacijos parametras $R^2 > 0,98$ (SOM) arba $R^2 > 0,84$ (ND).

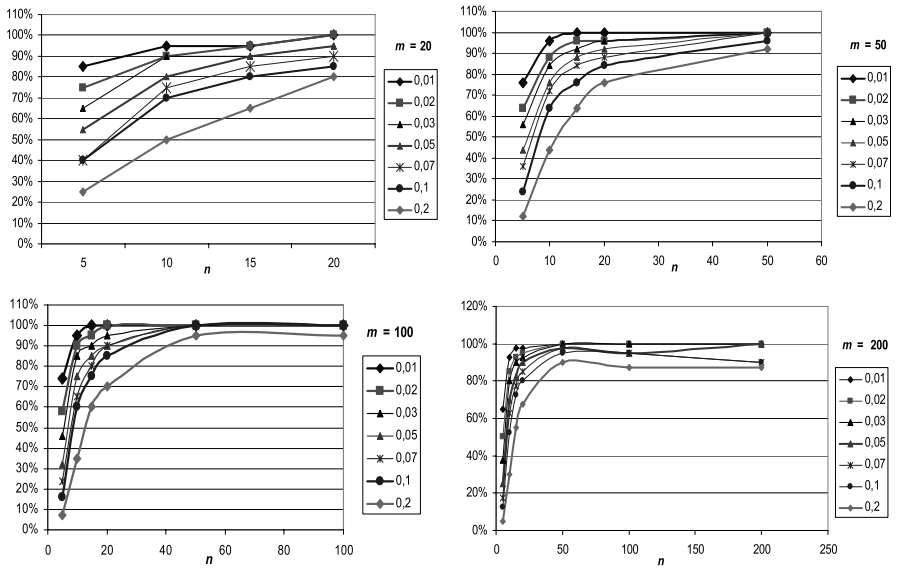
Neuronų skaičiaus priklausomybės nuo duomenų vektorių matmenų skaičiaus n tyrime naudoti dirbtinai sugeneruoti duomenys. Buvo sugeneruota 100 aibių, kurios sudarytos iš m vektorių, vektorių matmenų skaičius yra n. Išmokomi SOM ir ND tinklai.



1 pav. Kvantavimo paklaidos.



2 pav. Neuronų skaičius SOM metode.



3 pav. Neuronų skaičius ND metode.

Apskaičiuojamas kvantavimo paklaidų vidurkis. Eksperimentai atlikti su įvairiomis m ir n reikšmėmis, neuronų skaičius $N = 4, \ldots, m$. Fiksuotos neuronų skaičiaus $N' < N$ reikšmės, kai esant šiam neuronų skaičiui, kvantavimo paklaidos reikšmė nuo mažiau-

sios (kai $N = m$) skiriasi ne daugiau kaip $\varepsilon = 0,01;\ 0,02;\ 0,03;\ 0,05;\ 0,07;\ 0,1$ ar 0,2. Apskaičiuota neuronų skaičiaus N' reikšmių procentinė dalis nuo visų neuronų skaičiaus $N = m$.

Neuronų skaičiaus N' priklausomybė nuo analizuojamų vektorių matmenų skaičiaus n pavaizduota 2 ir 3 pav. Didesnė procentinė išraiška reiškia, kad tinklą reikia sudaryti iš daugiau neuronų, kad gautume norimo tikslumo paklaidą. Iš 2 ir 3 paveikslų matome, kad didėjant n, procentinė neuronų dalis taip pat didėja, esant fiksuotam paklaidų skirtumui ε. Tai reiškia, kad esant didesniam n, reikia imti daugiau neuronų, kad gautume pakankamai gerus rezultatus kvantavimo paklaidos prasme. Palyginus SOM tinklo ir ND metodo rezultatus pastebėta, kad SOM metode galima atsisakyti žymiai didesnės dalies neuronų neprarandant daug tikslumo. Pavyzdžiui, kai $m = 50$, $n = 20$, $\varepsilon = 0,2$, SOM metode užtenka tik šiek tiek daugiau nei 20% neuronų, o ND metode tam pačiam tikslumui pasiekti reikia beveik 80% neuronų.

5. Išvados

Šiame straipsnyje nagrinėta neuronų skaičiaus parinkimo strategija dvejuose vektorių kvantavimo metoduose – saviorganizuojančiame neuroniniame tinkle bei neuroninių dujų metode. Tirta, iš kiek neuronų reikia sudaryti tinklą, kad analizuojama vektorių aibė būtų sumažinta taip, kad kvantavimo paklaida nuo mažiausios, kuri gaunama, kai neuronų skaičius sutampa su analizuojamų vektorių skaičiumi, skirtųsi mažu dydžiu. Nustatyta, kad esant didesniam vektorių matmenų skaičiui, tinklą reikia sudaryti iš daugiau neuronų, kad kvantuoti vektoriai kuo tiksliau atspindėtų analizuojamų vektorių savybes. SOM tinklą užtenka sudaryti iš žymiai mažesnio skaičiaus neuronų. Norint pasiekti tą patį tikslumą ND metode reikia naudoti daugiau neuronų.

Literatūra

1. A. Asuncion, D.J. Newman, *UCI Machine Learning Repository*, University of California, School of Information and Computer Science, Irvine, CA (2007), `http://www.ics.uci.edu/~mlearn/MLRepository.html`
2. E. Alhoniemi, J. Himberg, J. Parhankangas, J. Vesanto, *SOM Toolbox for Matlab 5*, Helsinky University of Technology, Report A57 (2000), `http://www.cis.hut.fi/projects/somtoolbox/`
3. G. Dzemyda, O. Kurasova, J. Žilinskas, *Daugiamačių duomenų vizualizavimo metodai*, Mokslo aidai, Vilnius (2008).
4. T. Kohonen, *Self-Organizing Maps*, 3nd ed., Springer series in information sciences, Springer-Verlag, Berlin (2001).
5. T.M. Martinetz, K.J. Schulten, A neural-gas network learns topologies, in: T. Kohonen, K. Mäkisara, O. Simula and J. Kangas (Eds.), *Artificial Neural Networks*, Amsterdam, North-Holland (1991), pp. 397–402.
6. A. Molytė, O. Kurasova, Vektorių kvantavimo metodo *Neural-Gas* tyrimas, iš *11-osios Lietuvos jaunųjų mokslininkų konferencijos „Mokslas – Lietuvos ateitis“ 2008 metų teminės konferencijos „Informatika“ straipsnių rinkinys*, Vilnius, VGTU, Technika (2008), pp. 198–205.

SUMMARY

O. Kurasova, A. Molytė. Selection of number of neurons for vector quantization methods

In this paper, a strategy of the selection of the neurons number for vector quantization methods has been investigated. Two methods based on neural networks have been analysed: self-organizing map and neural gas. There is suggested a way under which the number of neurons is selected taken into account the particularity of the analysed data set.

Keywords: vector quantization, self-organizing map, neural gas, quantization error.