УДК 005.01 DOI: 10.30837/0135-1710.2021.177.075

О.С. ЧАЛА

ЗГОРТКОВА ПОДВІЙНА РАДІАЛЬНО-БАЗИСНА НЕЙРОННА МЕРЕЖА НА ОСНОВІ АКТИВАЦІЙНИХ ЯДЕРНИХ ФУНКЦІЙ СПЕЦІАЛЬНОГО ТИПУ

Запропонована нейронна мережа, що має ядерні функції активації та призначена для вирішення задач розпізнавання та класифікації зображень в режимі онлайн. Модифікована імовірнісна нейронна мережа використовується в якості автоенкодера і отримує на вхідний шар зображення, представлені у оригінальній матричній формі. На виходах нейрокомпресора, кількість яких визначається кількістю класів у наборі даних, з'являється вектор сигналів попередньої класифікації. Потім утворений вектор-сигнал надходить на вхідний шар радіально-базисної нейронної мережі, яка утворює розділяючі гіперповерхні між класами довільної складної форми. Система характеризується не тільки високою точністю класифікації, а й високою швидкістю навчання, що дозволяє обробляти потоки даних, які послідовно подаються в режимі онлайн.

1. Вступ

Розпізнавання образів-зображень на сьогодні є однією з ключових задач інтелектуального аналізу даних, що вирішується у онлайн-режимі в рамках практичних завдань ідентифікації власників електронних пристроїв, знаходження правопорушників, розпізнавання об'єктів дорожньої інфраструктури безпілотними машинами, постановки попереднього діагнозу у медицині. Ця задача є досить складною у комп'ютерній реалізації через різноманітність форм кожного образу, тому для вирішення використовуються підходи, засновані на апараті обчислювального інтелекту [1-3].

На сьогодні найефективнішим апаратом для вирішення задачі розпізнавання образів-зображень є згорткові нейронні мережі (CNN), які є одним з найбільш передових різновидів глибоких нейронних мереж (DNN) [4-6]. Такі мережі забезпечують високу точність розпізнавання, проте їх використання стикається з цілим рядом проблем. По-перше, для навчання CNN потрібні великі обсяги навчальних виборок, які не завжди доступні при вирішенні практичних завдань. Використання передавального навчання далеко не завжди дозволяє вирішити цю проблему. По-друге, згорткові нейронні мережі містять, як правило, дуже велику кількість параметрів синаптичних ваг, та для свого навчання вимагають досить багато часу. Тому робота CNN в онлайн-режимі є практично неможливою. Згорткова нейронна мережа складається з двох секцій: нейрокомпресора, утвореного послідовністю шарів згортки та субдискретизації, та багатошарового персептрона (MLP), який, власне, і вирішує задачу апроксимації-розпізнавання. Вхідне зображення поступає на вхід нейрокомпресора, що перетворює вхідне зображення у вигляді матриці у вектор відносно невисокої розмірності, який надходить на входи багатошарового персептрона. Більшість витрат часу у згорткових нейронних мережах припадає на навчання багатошарових персептронів.

Для того, щоб різко підвищити швидкодію процесу розпізнавання, можна використати переваги ймовірнісних нейронних мереж (PNN), основою яких є ідеї байєсівського висновку ядерної регресії, вікон Парзена та оцінки Надарая-Ватсона [7-9]. Налаштування параметрів таких мереж реалізується на основі лінивого навчання [10] за принципом «Нейрони в точках даних», тобто проходить дуже швидко, з використанням «миттєвих моделей» [11]. Ймовірнісні нейронні мережі, незважаючи на суттєві переваги у швидкості, а саме, здатність навчатися та вирішувати задачі розпізнавання майже миттєво, програють за точністю згортковим нейронним мережам.

Основною вимогою, що висуваються до MLP, є універсальні апроксимуючі властивості. Такі властивості мають не тільки багатошарові персептрони, але й мережі радіально-базисних функцій (radial-basis function networks, RBFNs) [12-16], які є «близькими родичами» ймовірнісних нейронних мереж, що реалізують ядерну апроксимацію [9] і не поступаються за точністю MLP. В [16] була запропонована і досліджена CNN, в якій замість MLP

була використана RBFN. Така модифікована мережа є досить ефективною при вирішенні задачі розпізнавання образів, проте все ще має певні недоліки. По-перше, мережі радіально-базисних функцій страждають від «прокляття розмірності» коли кількість параметрів синаптичних ваг експоненційно зростає із зростанням розмірності вхідного вектора, що веде до формування вектора невисокої розмірності на виході останнього шару субдискретизації автоенкодера та відповідної втрати точності. По-друге, в такій мережі основний час витрачається на операції згортки та субдискретизації, а не власне на завдання апроксимації.

Подолати недоліки розглянутих мереж CNN та RBFN можна, використовуючи замість стандартного автоенкодера класичну ймовірнісну нейронну мережу. Остання повинна бути модифікована таким чином, щоб вхідний сигнал, який надходить до мережі, формувався не у вигляді вектора, а в формі зображення-матриці [17]. Варто зауважити, що класична ймовірнісна нейронна мережа також потерпає від «прокляття розмірності», тобто вона нарощує свою архітектуру у першому прихованому шарі патернів з надходженням нових спостережень з навчальної вибірки. Для того, щоб подолати цей недолік, можна замість стандартної RBFN застосувати, так звану мережу гіпербазисних функцій (HBFN) [18]. Ця мережа використовує замість рецепторних полів гіперсфер активаційних функцій RBFN, що засновано на гіпереліпсоїдах, які мають довільну орієнтацію осей. Введення додаткових контурів навчання параметрів цих гіпереліпсоїдів [19] дозволяє поліпшити апроксимаційні властивості мережі, зменшити кількість параметрів (подолання «прокляття розмірності») і захиститися від виникнення «дірок» у просторі ознак.

2. Архітектура подвійної радіально-базисної нейронної мережі з матричними входами

Запропонована нейронна мережа призначена для вирішення задачі розпізнавання зображень, і, подібно до традиційних глибоких згорткових нейронних мереж, також складається з двох секцій: автоенкодера у вигляді матричної ймовірнісної нейронної мережі і апроксиматора у вигляді радіально-базисної нейронної мережі з налаштовуваними рецепторними полями. При цьому архітектури обох секцій мережі досить близькі і засновані на використанні багатовимірних ядерних активаційних функцій [9]. Архітектура запропонованої мережі наведена на рис.1.

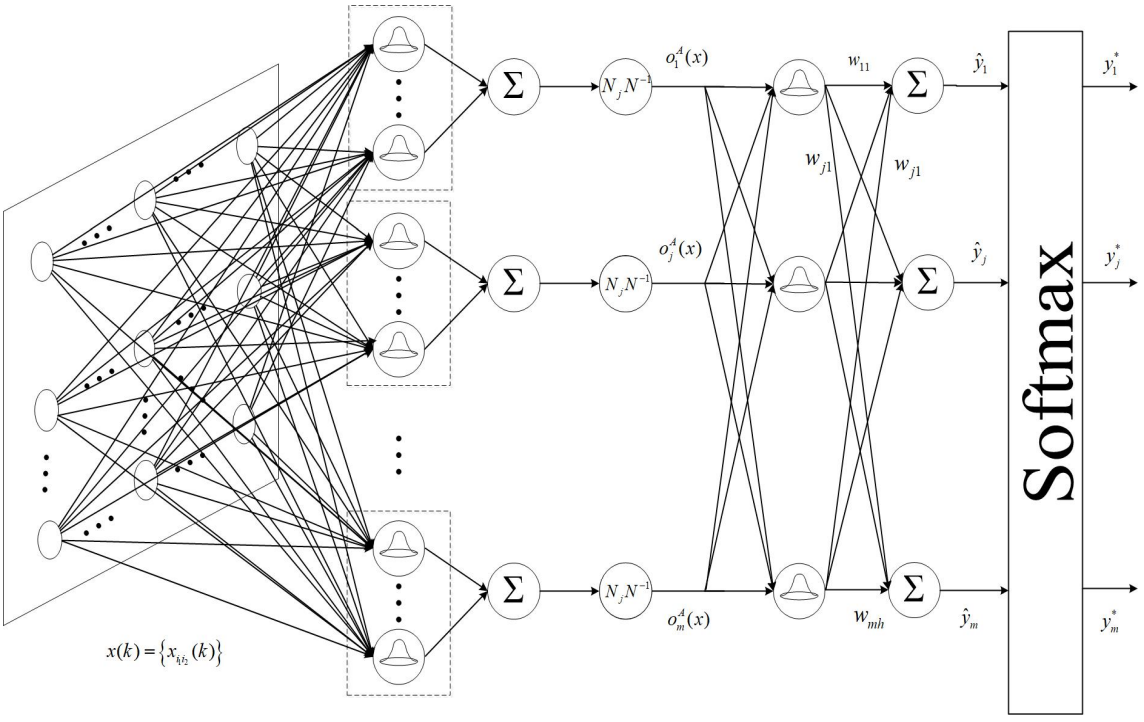

Рис. 1. Архітектура подвійної радіально-базисної нейронної мережі

Автоенкодер побудований на основі матричної ймовірнісної нейронної мережі [17] і містить три шари обробки інформації: шар образів, другий прихований шар, утворений m суматорами (тут *m* - кількість класів у вибірці) і третій (вихідний) шар корекції розподілу ймовірностей.

Навчальна вибірка представляє собою масив з N образів, кожен з яких являє собою ($n_1 \times n_2$) матрицю-зображення $x(k)=\{x_{i_1 i_2}(k)\}$, де $k=1,2,...,N$ - номер образу в навчальній вибірці. Також передбачається, що N_1 образів у початковому масиві даних відноситься до першого класу Cl_1, N_2 - до другого Cl_2 й нарешті, N_m - *m*-го класу Cl_m, тобто $\sum_{j=1}^{m} N_j = N.$

Кількість ядерних активаційних функцій (R-нейрони) у класичній PNN визначається обсягом вибірки N. Активаційні функції шару образів автоенкодера позначаються як

$$\varphi^A(x,\, c_{\tau_j}, \sigma^2_{\tau_j}), \tag{1}$$

де j - номер спостереження; τ_j варіюється в інтервалі ; $c_{\tau_j} \in R^{n_1 \times n_2}$ - матриця-центр активаційної функції, яка визначається в процесі навчання; $\sigma^2_{\tau_j}$ - параметр сферичного рецепторного поля відповідної дзвонуватої активаційної функції.

В якості активаційної функції була обрана ядерна функція В. Єпанечнікова [20], модифікована для матричного випадку:

$$\varphi^A(x,\, c_{\tau_j}, \sigma^2_{\tau_j}) = 1 - \left\| x - c_{\tau_j} \right\|^2 \sigma^{-2}_{\tau_j}, \tag{2}$$

де $\left\| x - c_{\tau_j} \right\|^2 = Tr(x - c_{\tau_j})(x - c_{\tau_j})^T$ - символ сліду матриці (матрична метрика Фробеніуса); $\sigma^2_{\tau_j}$ - радіус рецепторного поля активаційної функції.

Другий прихований шар нейрокомпресора утворений m суматорами, кожен з яких відноситься до конкретного класу Cl_j. На виходах цих суматорів розраховуються парзенівські оцінки щільності розподілу з урахуванням виходів попереднього шару $o^{[1]}_{\tau_j}(x)$ для спостереження x:

$$p_j(x) = o^{[2]}_j(x) = \sum_{\tau_j = N_1 + N_2 + ... N_{j-1} + 1}^{N_1 + N_2 + ... N_j} o^{[1]}_{\tau_j}(x). \tag{3}$$

У результаті роботи третього шару ці оцінки уточнюються з урахуванням значень емпіричних апріорних ймовірностей $N_j N^{-1}$ і на виходах автоенкодера з'являються ймовірності відношення спостереження *x*, що класифікується до *j*-го класу:

$$o^A_j(x) = o^{[2]}_j(x) N_j N^{-1}. \tag{4}$$

Таким чином, на виході нейрокомпресора з'явиться *m*-вимірний сигнал $o^A(x) = (o^A_1(x),...,o^A_j(x),...,o^A_m(x))^T$, який є попередньою оцінкою результатів класифікації.

Нейронна мережа-апроксиматор побудована на основі гіпербазисної нейронної мережі (HBNN), що є модифікацією популярної мережі радіально-базисних функцій з рецепторними гіпереліпсоїдними полями, що мають довільну орієнтацію осей. При цьому передбачається, що в процесі навчання HBFN параметри цих гіпереліпсоїдів можуть налаштовуватися одночасно з синаптичними вагами.

На входи HBFN з навчальної вибірки надходить послідовність $o^A(x(k)) = o^A(k) = (o^A_1(x),...,o^A_j(x),...,o^A_m(x))^T, k = 1,2,...,N$, яка передається на *h* гіпербазисних R-нейронів, на виходах яких формуються сигнали

$$\varphi^H_l(o^A(k),\, c_l, \Sigma^{-1}_l) = \; 1 - (c\sum\nolimits^{-1}(o^A(k) - c_l) = 1 - \left\| o^A(k) - c_l \right\|^2_{\Sigma^{-1}_l}, l = 1,2,...,h, h >> m, \tag{5}$$

де $c_l \in R^m$ - векторний центр активаційної функції $\varphi^H_l(\circ)$, Σ^{-1}_l - коваріаційна матриця, яка визначає форму, розмір та орієнтацію осей рецепторного поля відповідної активаційної функції.

Вихідні сигнали R-нейронів подаються на вихідний шар HBNN, утворений елементарними персептронами Розенблата з активаційними функціями типу softmax. Таким чином, на виходах HBNN в цілому формуються сигнали

$$\hat{y}_j(k) = \mathrm{w}_{j0} + \sum_{l=1}^{h} \mathrm{w}_{jl} \varphi_l^H \left(\left\| o^A(k) - c_l \right\|_{\Sigma_l^{-1}}^2 \right) = \sum_{l=0}^{h} \mathrm{w}_{jl} \varphi_l^H \left(\left\| o^A(k) - c_l \right\|_{\Sigma_l^{-1}}^2 \right), \varphi_0^H(\circ) \equiv 1, \tag{6}$$

$$y_j^*(k) = \mathrm{softmax}\, \hat{y}_j(k) = \exp \hat{y}_j(k) \left(\sum_{p=1}^{m} \exp \hat{y}_p(k) \right)^{-1}. \tag{7}$$

Якщо сигнал на виході PNN визначає рівень ймовірності того, що образ, який класифікується, відноситься до конкретного класу, то сигнал на виході введеної HBFN задає рівні нечіткої належності цього спостереження до того ж класу.

3. Навчання подвійної радіально-базисної нейронної мережі з матричними входами

Навчання подвійної радіально-базисної нейронної мережі відбувається окремо для автоенкодера і апроксиматора, базуючись на різних принципах. Налаштування нейрокомпресора реалізується за допомогою лінивого навчання за принципом «Нейрони в точках даних» і практично миттєво. Тобто центр ядерної активаційної функції встановлюється в точці з координатами вхідного образу з певною класифікацією, забезпечуючи високу точність класифікації і високу швидкість навчання [11].

Налаштування гіпербазисної нейронної мережі реалізується на основі контрольованого навчання з «гарячим» кодуванням навчального сигналу, тобто елементи зовнішнього навчального сигналу $y_j(k)$ можуть приймати тільки два значення: 1, якщо $x(k)$ належить до конкретного класу, і 0 в іншому випадку.

В якості критерію навчання HBFN використовується стандартна кросентропія:

$$E = -\sum_{k=1}^{N} \sum_{j=1}^{m} y_j(k) \ln y_j^*(k) = -\sum_{k=1}^{N} \sum_{j=1}^{m} y_j(k) \ln \exp \hat{y}_j(k) \left(\sum_{p=1}^{m} \exp \hat{y}_p(k) \right)^{-1}. \tag{8}$$

Далі для скорочення запису введемо додаткові позначення: $w_j = \left(w_{j0}, w_{j1}, ..., w_{jl}, ..., w_{jh} \right)^T$,

$\varphi^H(o^A(k), c_l, \Sigma_l^{-1}) = \left(1, \varphi_1^H(o^A(k), c_1, \Sigma_l^{-1}), ..., \varphi_l^H(o^A(k), c_l, \Sigma_l^{-1}), ..., \varphi_h^H(o^A(k), c_h, \Sigma_l^{-1})\right)^T$.

Градієнтна процедура навчання [19] для налаштування синаптичних ваг, центрів і матриць рецепторних полів має вигляд:

$$w_{jl}(k+1) = w_{jl}(k) + \eta_w(k+1) \left(y_j(k+1) - w_j^T(k) \varphi^H \left(\left\| o^A(k+1) - c(k) \right\|_{\Sigma^{-1}(k)}^2 \right) \right) \varphi_l^H \left(\left\| o^A(k+1) - c_l(k) \right\|_{\Sigma_l^{-1}(k)}^2 \right), \tag{9}$$

$$c_l(k+1) = c_l(k) - \eta_c(k+1) \left(y_j(k+1) - w_j^T(k+1) \varphi^H \left(\left\| o^A(k+1) - c_l(k) \right\|_{\Sigma^{-1}(k)}^2 \right) \right) w_{jl}(k+1) \bullet$$
$$\bullet \left(\varphi^H \left(\left\| o^A(k+1) - c_l(k) \right\|_{\Sigma^{-1}(k)}^2 \right) \right) \Sigma_l^{-1}(k) \left(o^A(k+1) - c(k) \right), \tag{10}$$

$$\Sigma_l^{-1}(k+1) = \Sigma_l^{-1}(k) \eta_\Sigma(k+1) \left(y_j(k+1) - w_j^T(k+1) \varphi^H \left(\left\| o^A(k+1) - c_l(k) \right\|_{\Sigma^{-1}(k)}^2 \right) \right) \bullet$$
$$\bullet\, w_{jl}(k+1) \left(\varphi^H \left(\left\| o^A(k+1) - c_l(k) \right\|_{\Sigma^{-1}(k)}^2 \right) \right) \left(o^A(k+1) - c_l(k+1) \right) \left(o^A(k+1) - c_l(k+1) \right)^T, \tag{11}$$

де w_j й $\varphi^H\left(\left\|o^A(k)-c_l\right\|^2_{\Sigma_l^{-1}}\right)$ - $(h+1)\times 1$ вектори синаптичних ваг на j-му виході системи і сигналів на виходах R-нейронів HBNN відповідно; $\eta_w(k+1)$, $\eta_c(k+1)$, $\eta_\Sigma(k+1)$ - параметри кроку навчання для налаштування змінних.

Таким чином, на відміну від традиційних CNN, де налаштовуються тільки синаптичні ваги нейронів, в запропонованій системі одночасно уточнюються параметри активаційних функцій, що забезпечують їй додаткову гнучкість і швидкодію.

4. Висновки

Запропоновано нейромережеву систему, яка була розроблена для вирішення задачі класифікації образів-зображень за мінімальний час. На відміну від традиційних згорткових нейронних мереж, запропонована система використовує ядерні активаційні функції. Вона складається з двох частин - нейрокомпресора та системи розпізнавання. В якості нейрокомпресора використовується модифікована ймовірнісна нейронна мережа з матричними входами, а система розпізнавання базується на гіпербазисній нейронній мережі.

Особливість розробленої системи полягає в комбінованому навчанні двох незалежних блоків. Перший блок, що представлений автоекодером, налаштовується за допомогою лінивого навчання, базуючись на концепції «Нейрони в точках даних». Система розпізнавання функціонує за парадигмою навчання з учителем та дозволяє налаштовувати не тільки синаптичні ваги, але й характеристики функцій активації - центри та рецепторні поля. Завдяки такому підходу до навчання, система в цілому характеризується високою швидкістю та високою точністю. Ця система дозволяє з високою точністю вирішувати задачу розпізнавання образів-зображень за умов коротких та довгих вибірок та вирішувати задачу класифікації потоку даних. При цьому запропонована система не потерпає від «прокляття розмірності», тобто дозволяє класифікувати спостереження, що надходять послідовно у онлайн-режимі та мають форму матриць.

Список літератури: 1. *Mumford C.* Computational Intelligence//Springer Berlin Heidelberg. 2009. P. 732. DOI: 10.1007/978-3-642-01799-5. 2. *Kacprzyk J., Pedrycz W.* Springer Handbook of Computational Intelligence// Berlin Heidelberg: Springer, Verlag. 2015. P. 1634. DOI: 10.1007/978-3-662-43505-2. 3. *Kruse R., Borgelt C., Klawonn F., Moewes C., Steinbrecher M., Held P.* Computational Intelligence: A Methodological Introduction// Berlin: Springer-Verlag. 2013. P.492. DOI: 10.1007/978-1-4471-5013-8. 4. *Bengio Y., Le Cun Y., Hinton G.* Deep learning// Nature. № 521(7553). 2015. P.436-444. 5. *Schmidhuber J.* Deep learning. Neural networks: An overview//Neural Networks. № 61. 2015. P.85-117. 6. *Goodfellow I., Bengio Y., Courville A.* Deep Learning//MIT Press. 2016. P. 800. 7. *Specht D.F.* Probabilistic neural networks// Neural Network. № 3. 1990. pp. 109-118. 8. *Specht D. F.* Probabilistic neural networks and polynomial ADALINE as complementary techniques to classification// IEEE Trans. on Neural Networks. №1. 1990. P. 111-121. 9. *Kung S.Y.* Kernel Methods and Machine Learning// Cambridge: University. 2014. P. 591. DOI: 10.1017/ CBO9781139176224. 10. *Zahirniak D.R., Chapman R., Rogers S.K., Suter B.W., Kabriski M., Pyatti V.* Pattern recognition using radial basis function network//Aerospace Application of Artificial Intelligence, Proceedings, Dayton, Ohio. 1990. P. 249-260. 11. *Nelles O.* Nonlinear Systems Identification// Berlin: Springer. 2001. P.786. DOI: 10.1007/978-3-662-04323-3. 12. *Moody J., Darken C.J.* Fast learning in networks of locally tuned processing units// Neural Computation. №1. 1989. P. 281-294. 13. *Poggio T., Girosi F.* Networks for approximation and learning// Proceedings of the IEEE. № 78(9). 1990. P. 1481-1497. 14. *Park J., Sandberg I. W.* Universal Approximation Using Radial-Basis-Function Networks// Neural Computation. №3(2). 1991. P. 246-257. 15. *Leonard J. A., Kramer M. A., Ungar L. H.* Using radial basis functions to approximate a function and its error bounds// IEEE Transactions on Neural Networks. №3(4). 1992. P. 624-627. 16. *Amirian M., Schwenker F.* Radial Basis Function Networks for Convolutional Neural Networks to Learn Similarity Distance Metric and Improve Interpretability// IEEE Access. № 8. 2020. P. 123087-123097. 17. *Bodyanskiy Ye., Deineko A., Pliss I., Chala O., Nortsova A.* Matrix Fuzzy-Probabilistic Neural Network in Image Recognition Task// 2020 IEEE Third International Conference on Data Stream Mining and Processing (DSMP). 2020. P. 33-36. 18. *Bodyanskiy Ye., Tyshchenko A., Deineko A.* An evolving radial basis neural network with adaptive learning of its parameters and architecture// Aut. Control Comp. Sci. №49. 2015. P. 255-260. 19. *Bodyanskiy Ye., Kolodyazhniy V., Stephan A.* An Adaptive Learning Algorithm for a Neuro-fuzzy Network. Computational Intelligence. Theory and Applications. Fuzzy Days // Lecture Notes in Computer Science, Springer, Berlin, Heidelberg. № 2206. 2001. P. 68-75. 20. *Epanechnikov V.A.* Non-Parametric Estimation of a Multivariate Probability Density// Theory Probab. №14(1). 1967. P. 153-158. DOI: 10.1137/1114019

Надійшла до редколегії 29.06.2021

Чала Ольга Сергіївна, молодший науковий співробітник ПНДЛ АСУ ХНУРЕ. Наукові інтереси: нейронні мережі та інші засоби штучного інтелекту. Адреса: Україна, 61166, м. Харків, пр. Науки, 14, тел. (057) 702 18 90.