

Оптимізація Процесу Навчання Штучних Нейронних Мереж

Олександр Безсонов, Олег Руденко
кафедра комп'ютерних інтелектуальних технологій та систем
Харківський національний університет радіоелектроніки
Харків, Україна
oleksandr.bezsonov@nure.ua, oleg.rudenko@nure.ua

Кирило Олійник, Олександр Романюк
кафедра інформатики
Харківський національний університет радіоелектроніки
Харків, Україна
kirill.oleynik.olegovich@gmail.com, romanyk@gmail.com

Training Process Optimization in Artificial Neural Networks

Oleksandr Bezsonov, Oleg Rudenko
Department of Computer Intelligent Technologies and Systems,
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
oleksandr.bezsonov@nure.ua, oleg.rudenko@nure.ua

Kiril Oliynyk, Oleksandr Romanyk
Department of Informatics,
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
kirill.oleynik.olegovich@gmail.com, romanyk@gmail.com

Анотація—В роботі розглядається питання оптимізації навчання штучних нейронних мереж шляхом вибору оптимальних параметрів, що використовуються в функціоналах (параметр масштабу), та параметрів в алгоритмах оцінювання Качмажа та Нагумо-Ноди (параметри релаксації та регуляризації).

Abstract—In this paper we consider the issue of optimizing the learning of artificial neural networks by selecting the optimal parameters used in the functionalities (scale parameter) and parameters in the Kachmazh and Nagumo-Noda evaluation algorithms (relaxation and regularization parameters).

Ключові слова—штучна нейронна мережа, алгоритм навчання, регуляризація, робастність

Keywords—artificial neural network, learning algorithm, regularization, robustness

I. ВСТУП

Задача побудови нейромережевої моделі нелінійного об'єкта

$$y_{n+1} = f(c^{*T} x_{n+1}) + \xi_{n+1}, \quad (1)$$

де y_{n+1} – вихідний сигнал, що спостерігається;
 $x_{n+1} = (x_{1,n+1}, x_{2,n+1}, \dots, x_{N,n+1})^T$ – вектор вхідних сигналів

$N \times 1$; $c^* = (c_1^*, c_2^*, \dots, c_N^*)^T$ – вектор параметрів $N \times 1$; $f(\bullet)$ – деяка нелінійна функція; ξ_{n+1} – завада; n – дискретний час, и зводиться до мінімізації наперед обраного функціонала якості

$$F[e_n] = \sum_{i=1}^n \rho(e_i), \quad (2)$$

де $e_i = y_i - \hat{y}_i$; $\hat{y}_i = \hat{\theta}_{i-1}^T x_i$ – вихідний сигнал моделі; $\hat{\theta}$ – оцінка вектора c^* ; $\rho(e_i)$ – деяка диференціюєма функція.

Вибір $\rho(e_i)$ суттєво впливає на властивості рішення.

При виборі $\rho(e_i) = 0.5e_i^2$ отримуємо оцінку МНК, яка є нестійкою до викидів і до завад, розподіли яких мають великі хвости. Вибір функції $\rho(e_i)$, відмінної від квадратичної, забезпечує робастність рішення.

Існує досить велика кількість функціоналів, що забезпечують отримання робастних М-оцінок, однак найбільш поширеними є комбіновані функціонали, запропоновані Хьюбером [1] та Хемпелем [2], які складаються з квадратичного, що забезпечує оптимальність оцінок для гауссівського розподілу, і модульного, що дозволяє отримати більш робастний до



Інформаційні системи та технології ICT-2020

Секція 9.

Інтелектуальний аналіз даних, Data Mining та Big Data-технології.

розподілів з важкими «хвостами» (викидами) оцінку. Однак ефективність одержуваних робастних оцінок істотно залежить від численних параметрів, використовуваних в цих функціоналах і які обираються зазвичай на основі досвіду дослідника.

II. ВИБІР ОПТИМАЛЬНИХ ПАРАМЕТРІВ КРИТЕРІЇВ НАВЧАННЯ

При М-оцінюванні, внаслідок неоднорідності функції ρ , одержувані оцінки не зберігають властивості інваріантності. Для того щоб властивість інваріантності масштабу виконувалося, в функціоналах замість помилки $e(i, \theta)$ слід брати $\tilde{e}(i, \theta) = (e(i, \theta) - m)/S$, де S - завадостійка оцінка параметра масштабу або міра розсіювання залишкових різниць (в разі нормального розподілу S є оцінкою σ_ξ); m - математичне сподівання завади $q(x)$, яке в загальному випадку відмінно від нуля.

При нульовому математичному сподіванні завади оцінка параметра масштабу використовуються в якості константи, що входить в розглянуті вище функціонали і функції впливу.

Якщо навчання мережі відбувається в режимі off-line, то в якості оцінки параметра масштабу слід взяти будь-яку MAD-оцінку (Median of Absolute Deviations, медіана абсолютних відхилень) [3].

У ранніх роботах по робастному оцінюванню в якості стійких оцінок параметра масштабу розглядалися медіанні. Так в [4] відзначається, що в якості оцінки S можна використовувати медіану модуля

$$S = \text{med}\{|e|\}.$$

Більш точна MAD-оцінка

$$S = \frac{\text{med}\{|e|\}}{0.6745}$$

розглядалася в роботах [4-5].

В [6] використовувалася медіанна оцінка виду

$$S = \frac{\text{med}\{|e - \text{med}(e)|\}}{0.6745}$$

Дещо інший підхід оцінювання S вивчався в [6]. Тут розглядався запропонований Хьюбером підхід отримання оцінки параметра S шляхом рішення S рівнянь

$$\frac{1}{n-2p-1} \sum_{i=p+1}^n \psi^2 \left[\frac{e_i}{S_i} \right] = \beta,$$

де константа β обирається так, що S є достовірною оцінкою σ_ξ в разі $\xi \in N(0, \sigma_\xi^2)$, тобто $\beta = E_\Phi\{\psi^2(e)\}$,

$$\beta = \frac{1}{\sqrt{2\pi}} \int [\Phi(x)]^2 e^{-\frac{x^2}{2}} dx,$$

де Φ - стандартний нормальний розподіл. Значення константи β для різних функцій впливу $\psi'(e)$ наведені в [7].

При вирішенні практичних завдань для оцінювання S зазвичай використовують протягом якогось певного числа вимірювань. Наприклад, в роботі [8] розглядається послідовність помилок $e(i)$, $i=1,2,\dots,n$, з яких k останніх ($p < k < n$) використовуються для отримання оцінки параметра масштабу.

Запропонована в цій роботі оцінка має вигляд

$$S_k = \frac{d_k}{\Phi^{-1}[0.5(1+k/n)]},$$

де d_k - половина ширини вікна, що включає k останніх помилок; $\Phi^{-1}[\bullet]$ - аргумент загальної (інтегральної) функції розподілу.

Оптимальне значення k відповідає мінімуму дисперсії нормалізованої помилки ε_k^2 , для визначення якої використовуються тільки різниці q_k точок:

$$\varepsilon_k^2 = \frac{1}{q_k - p} \sum_{i=1}^{q_k} \left(\frac{e_k(i)}{S_k} \right)^2 = \frac{\hat{\sigma}_k^2}{S_k^2}$$

При вирішенні задач в режимі on-line природним є рекуррентне оцінювання параметрів перешкоди.

Завадостійке оцінювання параметра положення на основі алгоритму стохастичної апроксимації Роббінса-Монро для випадку, коли безліч розподілів перешкоди належить класу асиметричних $\mathcal{E}$ -засмічених нормальних розподілів, розглядалося в [8], а для розподілів довільного виду - в роботі [9]. Такий підхід є досить привабливим в обчислювальному аспекті, тому що оцінки визначаються за допомогою простих рекурентних формул і не використовують в явному вигляді даних минулих спостережень.

III. ВИБІР ОПТИМАЛЬНИХ ПАРАМЕТРІВ ПРОЦЕДУР НАВЧАННЯ

Слід зазначити, що зазвичай в якості алгоритмів навчання, що мінімізують (2), обирають градієнтні, серед яких найбільш простими та досить ефективними є алгоритми, зокрема регуляризовані, Качмажа

$$\hat{\theta}_{n+1} = \hat{\theta}_n + \gamma \frac{e_{n+1} x_{n+1}}{\|x_{n+1}\|^2 + \delta},$$

та Нагумо-Ноди



$$\hat{\theta}_{n+1} = \hat{\theta}_n + \gamma \frac{e_{n+1} \text{sign} x_{n+1}}{|x_{n+1}| + \delta},$$

де γ - деякий параметр (параметр релаксації), що впливає на швидкість збіжності алгоритму; де $\delta > 0$ - параметр регуляризації (метою використання параметра регуляризації δ є підвищення обчислювальної стійкості алгоритмів).

Як і для будь-якого параметра, який штучно вводиться в критерій або оцінку, для параметра регуляризації δ не існує загальних рекомендацій по його вибору. Вибір оптимального значення δ залежить від наявності апіорної інформації щодо властивостей досліджуваного об'єкта, статистичних властивостей корисних сигналів і завад тощо. Неважко бачити, що з метою збільшення швидкості збіжності даний параметр повинен прагнути до нуля, а з метою підвищення стійкості оцінки повинен бути відмінним від нуля.

Однак практично всім відомим класичним процедурам вибору параметра регуляризації притаманний спільний недолік - вони не враховують статистичних властивостей одержуваних оцінок. Тому на практиці слід, швидше за все, виходити з конкретної ситуації і задавати параметр δ досить малим, але в той же час дозволяє отримувати оцінки в умовах мультиколінеарності.

Спроба непрямого обліку цих властивостей була зроблена в роботі [10], в якій в якості критерію запропоновано розглядати відношення сигнал-шум (ENR - echo-to-noise ratio), тобто нормалізовану величину

$$ENR = \frac{\sigma_y^2}{\sigma_\varepsilon^2} = \frac{\theta^T R_x \theta}{\sigma_\varepsilon^2},$$

$$\text{де } \sigma_y^2 = M\{\tilde{y}_n^2\} - \sigma_\varepsilon^2 = M\{\tilde{\varepsilon}_n^2\}; R_x = M\{x_n x_n^T\},$$

Мінімізація $M\{\tilde{\varepsilon}_{n+1}^2\} = \sigma_\varepsilon^2$, де $\tilde{\varepsilon}_{n+1} = y_{n+1}^* - w_n^T x_{n+1}$, в припущенні, що $\|x_n\|^2 \approx N\sigma_x^2$ і

$$M\{\|x_n\|\} = \alpha_x = \sqrt{2(\pi\sigma_x^2)^{-1}} \sigma_x^2 = \alpha_x \sigma_x^2,$$

$$\text{де } \alpha_x = \sqrt{2(\pi\sigma_x^2)^{-1}},$$

дозволяє отримати наступний вираз для нормалізованого параметра регуляризації

$$\delta = \frac{N(1 + \sqrt{1 + ENR})}{ENR} \sigma_x^2;$$

Інші рекомендації пропонуються в роботах [11, 12]

Важливим параметром, який визначає швидкість навчання, є параметр γ . Існує досить багато рекомендацій щодо його вибору, наприклад [13] для алгоритму Качмажа пропонується

$$\gamma^{opt} = 1 - \frac{N\sigma_\varepsilon^2}{(N-2)\sigma_x^2 M\{\|\tilde{\theta}_n\|^2\}},$$

а для алгоритму Нагумо-Ноди

$$\gamma_n^{opt} = \frac{2NM\{\|\tilde{\theta}_n\|^2\}}{\pi NM\{\|\tilde{\theta}_n\|^2\} + 2\omega\sigma_\varepsilon^2},$$

$$\text{де } \omega = (2\sigma_x^2)^{-1}.$$

IV. Висновки

В роботі було розглянуто питання оптимізації навчання ШНМ шляхом вибору оптимальних параметрів, використовуваних в функціоналах (параметр масштабу), та параметрів в алгоритмах оцінювання Качмажа та Нагумо-Ноди (параметри релаксації та регуляризації).

ЛІТЕРАТУРА

- [1] Hampel F.R. The influence curve and its role in robust estimation / F.R. Hampel // J. Amer. Statist. Assoc. – 1974. – June. – 69. – P. 383-393.
- [2] Hampel F.R. Robust Statistics. The Approach Based on Influence Functions / F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, W.A. Stahel – N.Y.: John Wiley and Sons, 1986. – 526 p.
- [3] Andrews D.F. A robust method for multiple linear regression / D.F. Andrews // Technometrics. – 1974. – 16. – P.523-531.
- [4] Mosteller F. Data analysis and regression: a second course in statistics / F. Mosteller, J.W. Tukey – Addison Wesley, 1977. – 588 p.
- [5] Deng G. Sequential and adaptive learning algorithms for M-estimation / G. Deng // EURASIP J. Adv. in Signal Proc., 2008, ID 459586.
- [6] Хогг Р.В. Введение в помехоустойчивое оценивание / Р.В. Хогг // В кн. «Устойчивые стохастические методы оценки данных» – М.: Машиностроение, 1984. – С.12-25.
- [7] Хьюбер П.Дж. Помехоустойчивое сглаживание / П.Дж. Хьюбер // В кн. «Устойчивые стохастические методы оценки данных» – М.: Машиностроение, 1984. – С. 36-46.
- [8] Martin R.D. Robust estimation via stochastic approximation / R.D. Martin, C.J. Masreliez // IEEE Tr. Inf. Theory. – 1975. – 21. – P. 263-271.
- [9] Ванде-Линде В.Д. Метод помехоустойчивого оценивания в задачах передачи сообщений / В.Д. Ванде-Линде // В кн. «Устойчивые стохастические методы оценки данных» - М.: Машиностроение, 1984. – С. 147-164.
- [10] Benesty J. On regularization in adaptive filtering / J. Benesty, C. Paleology, S. Ciochina // IEEE Trans. on Audio, Speech, and Language Proc. – 2011. – V.19. - №6. – P.1734-1742.
- [11] Руденко О.Г., Романюк О.С. Analysis of influence on the statistical properties of estimates of nonstationary parameters / В кн. Інформаційні системи і технології. (под ред. Пономаренка В.С.)- Харків: ФОП Бровін О.В., 2019.-С. 97-107.
- [12] Руденко О.Г., Безсонов О.О., Сердюк, Н.М., Олійник, К.О., Романюк О.С. Робастна ідентифікація об'єктів на основі мінімізації комбінованого функціоналу. Системи обробки інформації, №1 (160), 2020.- С.80-88
- [13] Rudenko O., Bezsonov O., Lebediev O., Lebediev V., Oliynyk K. Studying the properties of a robust algorithm for identifying linear objects, which minimizes a combined functional Eastern-European Journal of Enterprise Technologies, 2020. – № 4/4 (106). – P.37-46.



Інформаційні системи та технології ICT-2020

Секція 9.

Інтелектуальний аналіз даних, Data Mining та Big Data-технології.