Літвінчук Ю.А., Малик І.В.

Розширений алгоритм стратегії еволюції адаптації коваріаційної матриці

В роботі розглянуто розширення алгоритму CMA-ES з використанням сумішей розподілів для знаходження оптимальних гіперпараметрів нейронних мереж. Розроблений алгоритм побудовано згідно припущення багатопіковості щільності розподілу параметрів складних систем. На основі методу Монте Карло було встановлено, що новий алгоритм покращує пошук оптимальних гіперпараметрів в середньому на 12%.

Ключові слова і фрази: суміш розподілів, оптимізація гіперпараметрів, CMA-ES алгоритм, EM алгоритм.

Чернівецький національний університет імені Юрія Федьковича, Чернівці, Україна
e-mail: *j.litvinchuk@chnu.edu.ua, i.malyk@chnu.edu.ua*

Вступ

Оптимізація гіперпараметрів складних систем, сформульована як оптимізація чорної скриньки, є необхідною для автоматизації та високої продуктивності підходів машинного навчання. Гіперпараметри нейронних мереж часто оптимізуються пошуком через сітку, випадковим пошуком [1] або Байєсівською оптимізацією [2]. З точки зору Байєсівського підходу [3] побудови параметрів на основі генетичних алгоритмів [4], важливою є еволюційна стратегія [5] оцінки параметрів. Поняття еволюційної стратегії безпосередньо пов'язане зі зміною розподілу гіперпараметрів між епохами еволюційного алгоритму. Одним із найбільш вдалих методів еволюційних стратегій є метод адаптації коваріаційної матриці (Covariance matrix adaptation, CMA-ES) [6]. Метод CMA-ES має деякі корисні властивості інваріантності та зручний підхід оптимізації чорної скриньки з високим рівнем паралелізму, він показав більш високу продуктивність порівняно з підходами переліченими вище.

Даний метод, як і у випадку Байєсівського алгоритму, полягає у перерахунку коваріаційної матриці розподілів гіперпараметрів між епохами еволюційного алгоритму з подальшим вибором параметрів та врахуванням даної матриці. Очевидним недоліком такого методу є те, що припускається однопіковість щільності розподілу гіперпараметрів (як у нормальному розподілі). Проте на практиці, цільова функція (точності чи

УДК 004.021, 004.032.26, 004.852
2010 *Mathematics Subject Classification:* 30E10, 62F10, 65E05, 68T05.

функції втрат) не є однопіковою, що приводить до збільшення області пошуку за рахунок зміни однієї коваріаційної матриці та включення в область пошуку генетичного алгоритму область зі значеннями, що значно відрізняються від локальних екстремумів.

Насамперед CMA-ES – це стохастичний метод без похідних для чисельної оптимізації нелінійних чи неопуклих задач неперервної оптимізації. Ітераційний алгоритм CMA-ES часто використовує багатовимірний розподіл Гаусса $N(m, \Sigma)$, де $m \in R^d$, $\Sigma \in R^{d \times d}$ — додатно визначена симетрична матриця, d - число змінних. CMA-ES на кожній із своїх ітерацій вибирає λ кандидатів-рішень із багатовимірного нормального розподілу, оцінює ці рішення (послідовно або паралельно), а потім коригує розподіл вибірки, що використовується для наступної ітерації, щоб надати більшу ймовірність хорошим зразкам [7]. Вектор середніх m та коварійна матриця Σ оновлюються відповідно до ранжування рішень в останньому поколінні і CMA-ES навчається вибирати рішення з перспективної області.

CMA-ES, як правило, має тенденцію працювати найкраще для складних алгоритмів оцінки функцій; наприклад, у [8] показано, що CMA-ES дав найкращі результати серед більш ніж 100 класичних та сучасних оптимізаторів у широкому діапазоні функцій чорної скриньки. CMA-ES використовувався для налаштування гіперпараметрів і раніше, наприклад, у роботах [9] або автоматичного розпізнавання мови [10].

У зв'язку з цим пропонується використати деяке розширення CMA-ES алгоритму, використовуючи багатопікові моделі. Для цього буде використано поняття суміші (суміші розподілів), зокрема, на суміші нормальних розподілів, оскільки вибір цих розподілів є найбільш ілюстративний і легко може бути реалізований на практиці.

1 Оцінка параметрів щільностей сумішей

Надалі будемо припускати, що щільність суміші належить до одного сімейства розподілів [11], визначається як зважена сума k щільностей компонентів. Щільності компонентів обмежені деяким параметричним класом щільностей, який вважається придатним для наявних даних та обчислювальних цілей. Позначимо $p(x; \theta_s)$ - щільність s-го компонента, де θ_s - параметри компонента; π_s ваговий коефіцієнт s-го компонента суміші. Ваги повинні бути невід'ємними $\pi_s \geq 0$ та в сумі давати одиницю: $\sum_{s=1}^{k} \pi_s = 1$. Ваги π_s також відомі як «пропорції змішування» і їх можна розглядати як ймовірність $p(s)$ того, що вибірка даних буде вилучатися з компонентів суміші s. Тоді щільність суміші компонентів k визначається як:

$$p(x) = \sum_{s=1}^{k} \pi_s p(x; \theta_s), \tag{1}$$

де $\theta = \{\theta_1, ..., \theta_k, \pi_1, ..., \pi_k\}$ - параметри суміші.

Щільність суміші можна інтерпритувати як моделювання процесу, в якому спочатку вибирається «джерело» s відповідно до поліноміального розподілу $\pi_1, ..., \pi_k$, а потім береться вибірка з відповідної щільності компонентів $p(x; \theta_s)$. Таким чином, можливість вибору джерела s і даних x дорівнює $\pi_s p(x; \theta_s)$. Тоді гранична можливість вибору даних

x визначається формулою (1). Важливою похідною величиною є «апостеріорна ймовірність» компонента суміші, заданого вектором даних, яка використовується для оцінки параметрів суміші. Апостеріорна ймовірність компонентів суміші визначається за допомогою правила Байєса і має вигляд

$$p(s \mid x) := \frac{\pi_s p(x;\theta_s)}{p(x)} = \frac{\pi_s p(x;\theta_s)}{\sum_{\acute{s}} \pi_{\acute{s}} p(x;\theta_{\acute{s}})}. \tag{2}$$

Першим кроком під час використання моделі суміші є визначення її архітектури: відповідний клас щільностей компонентів і кількість щільностей компонентів у суміші. Після того, як ці варіанти дизайну зроблені, оцінюються вільні параметри в моделі суміші таким чином, щоб щільність (2) якомога точніше наближала щільність реальних даних.

Оцінка параметрів моделі для заданих даних стає пошуком параметрів максимальної правдоподібності для даних у наборі ймовірнісних моделей, визначених вибраною архітектурою. Логарифм правдоподібності для набору даних $X_N = \{x_1, ..., x_N\}$ можна записати як:

$$\mathcal{L}(X_N, \theta) = \log p(X_N;\theta) = \log \prod_{n=1}^{N} p(x_N;\theta) = \sum_{n=1}^{N} p(x_N;\theta). \tag{3}$$

Знайти параметри максимальної правдоподібності для щільності одного компонента легко і часто це можна зробити в аналітичному вигляді. Це стосується, наприклад, компонентів нормальної суміші. Однак, якщо ймовірнісна модель є сумішшю, оцінка часто стає значно складнішою, оскільки логарифмічна правдоподібність як функція параметрів може мати багато локальних екстремумів. Отже, для отримання оцінок параметрів необхідна деяка нетривіальна оптимізація, і тут добре спрацьовує ітераційний EM-алгоритм (expectation-maximization algorithm). EM-алгоритм знаходить параметри в локальних екстремумах логарифмічної функції правдоподібності, заданих деякими початковими значеннями параметрів.

2 Вибір базового розподілу суміші

Суміш розподілів - це суміш двох або більше ймовірнісних розподілів, зазвичай з одного сімейства. Випадкові змінні беруться з однієї батьківської популяції до створення нового розподілу. Батьківські популяції можуть бути одновимірними або багатовимірними і повинні мати однакову розмірність. Розподіли можуть складатися з різних розподілів (наприклад, нормальний розподіл та розподіл Стьюдента) або з одного й того ж розподілу з різними параметрами. Нові розподіли ймовірностей розглядаються як справжні функції щільності ймовірності і тому можуть використовуватися для знаходження очікуваних значень, оцінок максимальної правдоподібності та інших статистичних даних.

Суміш нормальних розподілів найчастіше використовуються для моделювання неперервних даних [12]. Перша причина такої популярності полягає в тому, що оцінка

параметра максимальної правдоподібності може бути виконана в закритій формі і вимагає лише обчислення середнього значення даних і коваріації. Друга причина полягає в тому, що з усіх щільностей із певною дисперсією щільсть Гаусса має найбільшу ентропію [13].

Можна виділити основні сімейства суміші розподілів.

- Пуасонові суміші [14], [15]

$$P_g(x|\varphi) = \int_0^\infty \frac{e^{(-\theta)}\theta^x}{x!} dG(\theta|\varphi).$$

- Суміші експоненційних розподілів [16], [17]

$$F(t) = \int_0^\infty (1 - e^{-xt}) dH(x).$$

- Суміші Вейбулла [18]

$$S(x) = \sum_{i=1}^{n} a_i e^{(-b_i x^{c_i})}, x > 0,$$

де $a_i \in \mathbb{R}$, $i = 1, ..., n$ та $\sum_{i=1}^{n} a_i = 1$.

- Суміш нормальних розподілів [19]

$$p(x) = \sum_{s=1}^{k} \pi_s p(x|\mu_s; \theta_s), \tag{4}$$

де $p(x|\mu_s; \theta_s)$ щільність багатовимірного нормального розподілу з $\mathbb{R}^d$ та параметрами $(\mu_s; \theta_s)$.

3 ЕМ-алгоритм

Оцінку параметрів розподілу суміші легко провести за допомогою ЕМ-алгоритму. Технічно ідея алгоритму полягає в ітераційному визначенні нижньої межі логарифмічної ймовірності та максимізації цієї нижньої межі. Алгоритм на прикладі суміші нормальних розподілів (4), виконує Е-крок, оцінюючи до якого компонента належить кожна точка даних, і М-крок переоцінює параметри на основі цієї оцінки.

Е-крок. Обчислення допоміжних величин:

$$r_{ij} = \frac{\pi_i p(x_i|\mu_j; \theta_i)}{\sum_{s=1}^{k} \pi_s p(x_j|\mu_s; \theta_j)}, \tag{5}$$

де r_{ij} – ймовірність того, що об'єкт x_i був отриманий з j-ї компоненти суміші при поточному наближенні параметрів π_i, θ_i.

М-крок. Переоцінка нового наближення параметрів суміші:

$$\pi_j = \frac{1}{n} \sum_{i=1}^{n} r_{ij}, \tag{6}$$

$$\mu_i = \frac{\sum_{i=1}^{n} r_{ij} \cdot x_i}{\sum_{i=1}^{n} r_{ij}}, \tag{7}$$

$$\theta_i = \frac{\sum_{i=1}^{n} r_{ij}(x_i - \mu_i)^T(x_i - \mu_i)}{\sum_{i=1}^{n} r_{ij}}. \tag{8}$$

Критерій зупинки. Ітерації методу здійснюються до збіжності варіаційної нижньої оцінки на логарифмічну функцію правдоподібності моделі.

$$\mathcal{L}(\pi, \mu, \theta, r|X) = \sum_{i=1}^{n}\sum_{j=1}^{k} r_{ij}(\ln \pi_j + \ln p(x_i|\mu_i; \theta_i) - \sum_{i=1}^{n}\sum_{j=1}^{k} r_{ij} \ln r_{ij}). \tag{9}$$

Під збіжністю можна розуміти, наприклад, змінення не більше ніж на заздалегідь задану точність $\varepsilon > 0$.

Оновлення середнього значення використовує нову вагу змішування, а оновлення коваріаційної матриці використовує нове середнє значення μ та ваги π.

4 Розширений алгоритм CMA-ES

Нехай $P(\theta; X_{1:k}, y_{1:k})$ – розподіл гіперпараметрів нейронної мережі на основі значень цільової функції, отриманої на основі k епох, де X_k – значення гіперпараметрів на k-му кроці, y_k – значення цільової функції на k-му кроці. Тоді алгоритм еволюційної стратегії на основі розширеного CMA можна описати наступними кроками:

1. Визначення області зміни гіперпараметрів (a_0), розмірності суміші (n), кількості генів в генетичному алгоритмі (N), точності методу (ε).
2. Задання випадковим чином ($\pi^{(0)}$, $\mu^{(0)}$, $\theta^{(0)}$).
3. Вибір N генів X_k згідно розподілу (4) та обчислень значень цільової функції y_k.
4. Перерахунок параметрів ($\pi^{(k+1)}$, $\mu^{(k+1)}$, $\theta^{(k+1)}$) на основі формул (6)-(8).
5. Якщо задовольняється умова виходу

$$|\mathcal{L}_{k+1} - \mathcal{L}_k| < \varepsilon$$

то перейти до виконання генетичного алгоритму на основі розподілу гіперпараметрів з розподілом

$$p(\theta) = P(\theta; X_{1:k}, y_{1:k}).$$

Якщо,

$$|\mathcal{L}_{k+1} - \mathcal{L}_k| < \varepsilon$$

то перейти до кроку 3.

Слід зауважити, що на кроці 3 виконується одна епоха генетичного алгоритму, тому поряд із оптимізацією параметрів π, μ, θ шукається і оптимальне значення гіперпараметів.

5 Моделювання

Як було зазначено вище, розглянутий розширений CMA-ES алгоритм дозволяє уникати розгляду хромосом із невисокими значеннями цільової функції $\mathcal{L}$. Моделювання багатопікових цільових функцій оптимізації показало, що кількість епох для пошуку глобального мінімуму можна зменшити до 60% при вірній оцінці параметрів суміші. Слід зуважити, що серідній відсоток покращення для розширеного CMA-ES алгоритму складає близько 12%. При моделюванні даних та оцінки глобальних максимумів було використано метод Монте-Карло.

6 Висновки

У роботі розглянуто розширення CMA-ES алгоритму за припущення багатопіковості розподілу хромосом в генетичному алгоритмі. Використовуючи теорію сумішей та оцінку параметрів сумішей на основі EM-алгоритму, розроблено алгоритм для оцінки гіперпараметрів складних систем на основі розширеного CMA-ES алгоритму. Використовуючи моделювання методом Монте-Карло встановлено, що розширений CMA-ES алгоритм покращує пошук оптимального розв'язку складної ситеми в середньому на 12%.

В подальших роботах цього напрямку плануться розглянути ефективність розробленого алгоритму на реальних даних.

Список літератури

[1] Bergstra J., Bengio Y. Random search for hyper-parameter optimization. JMLR, 13:281–305, 2012.

[2] Snoek J., Rippel O., Swersky K., Kiros R., Satish N., Sundaram N., Patwary M., Ali M., Adams R., et al. Scalable bayesian optimization using deep neural networks. arXiv preprint arXiv:1502.05700, 2015

[3] Eggensperger K., Feurer M., Hutter F., Bergstra J., Snoek J., Hoos H., and Leyton-Brown K. Towards an empirical foundation for assessing Bayesian optimization of hyperparameters. In NIPS workshop on Bayesian Optimization in Theory and Practice. - 2013.- 5p.

[4] Venkatesan D., Kannan K., Saravanan R. A genetic algorithm-based artificial neural network model for the optimization of machining processes. Neural Computing and Applications. -· February 2009.- 7p.

[5] Beyer H.-G. The Theory of Evolution Strategies. - Springer; 2001st edition (March 27, 2001).- 401p.

[6] Loshchilov I. , Hutter F. CMA-ES for Hyperparameter Optimization of Deep Neural Networks. arXiv:1604.07269v1 [cs.NE] 25 Apr 2016. – 9p.

[7] Hansen N. and Ostermeier A., Ostermeier A. Completely derandomized self-adaptation in evolution strategies.Evolutionary computation, 9(2):159–195, 2001.

[8] Loshchilov I., Schoenauer M., and Sebag M. Bi-population CMA-ES agorithms with surrogate'models and line searches. In Proc. of GECCO'13, pp. 1177–1184. ACM, 2013.

[9] Loshchilov I., Schoenauer M., and Sebag M. Self-adaptive Surrogate-Assisted Covariance Matrix Adaptation Evolution Strategy. In Proceedings of the 14th annual conference on Genetic and evolutionary computation, pp. 321–328. ACM, 2012.

[**10**] Watanabe Sh. and Le Roux J. Black box optimization for automatic speech recognition. In Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, pp. 3256– 3260. IEEE, 2014.

[**11**] McLachlan G. and Peel J. Finite Mixture Models. Wiley, New-York, NY, USA. 42p., 2000.

[**12**] Lee G. and Scott C. Em algorithms for multivariate gaussian mixture models with truncated and censored data. Computational Statistics & Data Analysis, 56(9):2816–2829, 2012. – 13p.

[**13**] Peter W., Buchen and Michael K. The Maximum Entropy Distribution of an Asset Inferred from Option Prices. The Journal of Financial and Quantitative Analysis Vol. 31, No. 1 (Mar., 1996), pp. 143-159

[**14**] Dempster, A. P., Laird, N., and Rubin D. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 39(1):1–38. Pages: 12, 44, 89. 1977.

[**15**] Karlis D. EM Algorithm for Mixed Poisson and Other Discrete Distributions. https://doi.org/10.1017/S0515036100014033 Published online by Cambridge University Press.May 2005 , pp. 3 - 24

[**16**] KARLIS, D. Estimation and hypothesis testing problems in Poisson mixtures. Phd Thesis, Department of Statistics, Athens University of Economics. 1998.

[**17**] Anaya-Izquierdo K., Marriott P. Local mixture models of exponential families (Submitted).2006.

[**18**] Carta J., Ramirez P. Analysis of two-component mixture Weibull statistics for estimation of wind speed distributions. Renew Energy 32:518:531. 2007.

[**19**] Elmahdy E., Aboutahoun A. A new approach for parameter estimation of of finite Weibull mixture distributions for reliability modeling. Appl Math Model 37:1800:1810. 2013.

Надійшло 14.12.2022

Litvinchuk Yu.A., Malyk I.V. *The extended CMA-ES algorithm*, Bukovinian Math. Journal. **10**, 2 (2022), 137–143.

The paper considers the extension of the CMA-ES algorithm using mixtures of distributions for finding optimal hyperparameters of neural networks. Hyperparameter optimization, formulated as the optimization of the black box objective function, which is a necessary condition for automation and high performance of machine learning approaches. CMA-ES is an efficient optimization algorithm without derivatives, one of the alternatives in the combination of hyperparameter optimization methods. The developed algorithm is based on the assumption of a multi-peak density distribution of the parameters of complex systems. Compared to other optimization methods, CMA-ES is computationally inexpensive and supports parallel computations. Research results show that CMA-ES can be competitive, especially in the concurrent assessment mode. However, a much broader and more detailed comparison is still needed, which will include more test tasks and various modifications, such as adding constraints. Based on the Monte Carlo method, it was shown that the new algorithm will improve the search for optimal hyperparameters by an average of 12%.