Володимир Дубовой[1]

СИНТЕЗ СТАТИСТИЧНО ОПТИМАЛЬНОГО ДАТАСЕТУ ДЛЯ НАВЧАННЯ НЕЙРОННОЇ МЕРЕЖІ

[1]Вінницький національний технічний університет

Для керування складними нелінійними системами все ширше використовують інтелектуальні технології, зокрема, на основі нейронних мереж. Проте навчання таких нейронних контролерів ускладнюється відсутністю на стадії створення системи розмічених датасетів для керованих об'єктів. У цій роботі досліджується модельно-орієнтоване навчання нейроконтролера і створення первинного навчального датасету на основі імітаційної моделі системи. Для цього поставлена і розв'язана задача синтезу оптимальної статистики вхідних даних імітаційної моделі за критерієм мінімуму середніх втрат. Задача генерування датасету з заданими статистичними характеристиками для навчання контролера динамічної системи керування ускладнюється тим, що необхідно одночасно забезпечити як необхідний багатовимірний розподіл ймовірностей у просторі реалізацій векторного впливу на систему, так і його кореляційну структуру у часі. Функцією втрат обрана середня квадратична похибка нейроконтролера. Усереднення здійснюється як по множині реалізацій, так і у часі для врахування динаміки системи. Імітаційна модель децентралізованої системи створена на платформі Scilab/Xcos з використанням попередньо створеної бібліотеки блоків для моделювання розподілених систем. Задача генерування датасету з заданими розподілом ймовірностей і кореляційною структурою розв'язується шляхом генерування некорельованого рівномірно розподіленого датасету з наступним нелінійним динамічним перетворенням. Одночасне отримання заданої виду статистичного розподілу і кореляційної структури отримується ітераційним шляхом з поступовим наближенням до необхідного результату.

Ключові слова: навчання нейронної мережі, синтез датасету, статистична оптимальність

Вступ

Для керування складними нелінійними системами все ширше використовують інтелектуальні технології, зокрема, на основі нейронних мереж. Проте навчання таких нейронних контролерів ускладнюється відсутністю на стадії створення системи розмічених датасетів для керованих об'єктів. Задача генерування датасету з заданими статистичними характеристиками для навчання контролера динамічної системи керування ускладнюється тим, що необхідно одночасно забезпечити як необхідний багатовимірний розподіл ймовірностей у просторі реалізацій векторного впливу на систему, так і його кореляційну структуру у часі.

Для лінійних аперіодичних об'єктів достатньо забезпечити рівномірність наближення до необхідної точності для всіх параметрів, а для цього достатньо рівномірний розподіл даних навчальної вибірки. Для динамічних (інерційних) об'єктів дані повинні надходити як часовий ряд з експоненціальною кореляційною функцією.

Проте для нелінійних об'єктів задача складніша, оскільки ціна похибки у різних частинах діапазону зміни виходу координатора може суттєво відрізнятися [1].

Задача генерування датасету з заданими розподілом ймовірностей і кореляційною структурою розв'язується шляхом генерування некорельованого рівномірно розподіленого датасету з наступним нелінійним динамічним перетворенням. Синтез необхідного нелінійного динамічного перетворення є надскладною математичною задачею, яка не має універсального розв'язку. До того ж таке перетворення складно реалізувати у імітаційній моделі через відсутність необхідних стандартних елементів у середовищі моделювання. Тому звичайно воно розкладається на окреме нелінійне перетворення і лінійне динаічне перетворення. Проте нелінійне перетворення випадкового процесу може суттєво змінити його автокореляційну функцію. Основні аспекти впливу нелінійного перетворення на автокореляцію включають [2]:

1) **Зміна форми кореляційної структури**: Це пов'язано з тим, що кореляційна структура у вихідному процесі трансформується під впливом нелінійних операцій, які зазвичай змінюють взаємозв'язки між окремими значеннями процесу.
2) **Втрата або поява кореляції**: У випадках, коли вихідний процес є некорельованим (наприклад, білий шум), нелінійне перетворення може породити кореляції між віддаленими значеннями.

3) **Зміна коефіцієнтів автокореляції**: Зокрема, піднесення до степеня, більшого за 1, посилює кореляцію між значеннями, які вже мали високий зв'язок у вихідному процесі.
4) **Зміна стаціонарності**: Якщо вихідний процес стаціонарний, то після нелінійного перетворення він може стати нестаціонарним. Це, в свою чергу, вплине на автокореляційну функцію, яка може почати залежати не лише від інтервалу, але й від абсолютного часу.
5) Автокореляційна функція є лінійною характеристикою, яка описує тільки лінійну складову залежності між значеннями процесу. Нелінійне перетворення вносить додаткові нелінійні залежності.

З іншого боку, динамічне перетворення також впливає на статистичний розподіл датасету, оскільки динамічне перетворення реалізується через операцію інтегрального перетворення з певним ядром, тобто дискретного додавання з ваговими коефіцієнтами, а це в свою чергу внаслідок центральної граничної теореми теорії ймовірностей наближає розподіл ймовірностей результату до гаусівського.

Задача генерування випадкових числових послідовностей широко обговорюється в літературі. В роботі [3] обговорюється задача генерування некорельованих послідовностей з декількома стандартними розподілами ймовірностей.

У роботі [4] запропонований «стохастичний синтетичний підхід» до генерування даних для моделювання енергетичних систем, яка враховує добові, тижневі, сезонні і річні тренди. Техніка врахування трендів поєднується з модифікованим методом фракційного гаусового шуму для роботи зі складними багатоперіодичними сезонними трендами.

У роботі [5] для моделювання впливу вітру на вітроелектростанцію генерується часовий стохастичний ряд як білий шум з розподілом Вейбула, а кореляційні характеристики отримуються шляхом лінійного перетворення типу «ковзне середнє».

У роботі [6] для навчання нейронної моделі медичної діагностики генерування даних використовується для збільшення навчального датасету. Генерується нормально розподілена послідовність із збереженням кореляції експериментальної послідовності за допомогою лінійної регресії і додавання шуму.

У роботі [7] для тестування систем машинного навчання генеруються часові ряди з різними статистичними характеристиками без прив'язки до реальних статистичних характеристик даних певного об'єкта.

У роботі [8] розглядається генерування вектору даних з взаємно корельованими компонентами, проте не розглядається динаміка. Кореляція між компонентами забезпечується шляхом попереднього аналізу двовимірних ймовірностей для кожної пари компонент у експериментальних даних.

У роботі [9] автори генерують векторні часові ряди з нормальним розподілом ймовірностей і заданою кореляціною структурою, а потім здійснюють нелінійне перетворення для отримання необхідного розподілу ймовірностей. При цьому вони виходять з гіпотези, що «since these transformations are monotonous, the rank correlation values do not change», що не завжди відповідає дійсності.

Таким чином, можна зробити висновок, що задача генерування стохастичних часових рядів для навчання нейронних мереж ще не отримала повного розв'язку і залишається актуальною.

Метою цієї роботи є розробка підходу до синтезу статистично оптимального датасету для навчання нейронних контролерів систем керування за допомогою імітаційної моделі системи.

Результати дослідження

Для досягнення поставленої мети необхідно розв'язати два завдання:

- визначення оптимального розподілу ймовірностей даних у датасеті за критерієм мінімуму середнього ризику;

- розробка методу генерування векторного часового ряду з визначеним розподілом ймовірностей і кореляційною структурою.

Для визначення оптимального розподілу ймовірностей введемо функцію втрат системи керування

$$q(u) = \max_u \left[y(\mathbf{X}, u) \right] - y(\mathbf{X}, u), \quad (1)$$

де $u \in U$ - керування; $\mathbf{X}$ - вектор входів нейроконтролера ; $y(\mathbf{X}, u)$ - виробнича функція об'єкта [1]. Відповідно контролер повинен реалізовувати функцію пошуку оптимального значення u, тобто $u = \varphi(\mathbf{X}) = \arg\max_u \left[y(\mathbf{X}, u) \right]$. Оцінимо ризик похибки нейроконтролера.

Якщо розподіл ймовірностей вектора вхідних даних контролера $f_{\mathbf{X}}(\mathbf{X})$, ці дані некорельовані і об'єкт статичний, то розподіл ймовірностей виходу контролера

$$f_U(u) = f_{\mathbf{X}}\left(\varphi^{-1}(u)\right) \cdot \left| \frac{d\varphi^{-1}(u)}{du} \right| \quad (2)$$

де $\varphi^{-1}(u)$ - обернена функція керування контролера. Відповідно ризик похибки контролера

$$R = \int_U q(u) f_U(u) du = \int_U \left[\max_u \left[y(\mathbf{X}, u) \right] - y(\mathbf{X}, u) \right] f_{\mathbf{X}}\left(\varphi^{-1}(u)\right) \cdot \left| \frac{d\varphi^{-1}(u)}{du} \right| du . \quad (3)$$

Задача пошуку оптимальної статистики вхідної навчальної вибірки полягає таких параметрів розподілу $\tilde{f}_{\mathbf{X}}(\mathbf{X})$, які мінімізують ризик R. Для параметризації цієї задачі доцільно використовувати розкладання розподілу у певному обраному базисі [10, 11].

Розкладання розподілу ймовірності $f_{\mathbf{X}}(\mathbf{X})$ і кореляційної структури $K_{\mathbf{XX}}(\tau)$ дозволяє також параметризувати варіації цих функцій для процесу оптимізації.

Нехай $A = \{a_0, a_1 \ldots a_n\}$ - вектор коефіцієнтів розкладання розподілу ймовірностей параметра $x_i \in \mathbf{X}$. Тоді вся сукупність коефіцієнтів розкладання розподілів вхідних даних $\mathbf{X}$ утворює матрицю $\mathbf{A}[m,n]$. Аналогічно сукупність коефіцієнтів розкладання кореляційних функцій вхідних даних утворює матрицю $\mathbf{B}[m,n]$.

Якщо визначений оптимальний розподіл вхідних даних $\tilde{f}_{\mathbf{X}}(\mathbf{X})$ і кореляційна функція $\tilde{K}_{\mathbf{XX}}(\tau)$, то відповідні матриці коефіцієнтів $\tilde{\mathbf{A}}[m,n]$ і $\tilde{\mathbf{B}}[m,n]$. Для перетворення вхідного датасету з метою забезпечення оптимальної статистики пропонується ітераційний алгоритм:

1) *Статистичний аналіз вхідних впливів* $\mathbf{X}$, *отримання* $f_{\mathbf{X}}$ *і* $K_{\mathbf{XX}}$;
2) *Визначення оптимального розподілу* $\tilde{f}_{\mathbf{X}}$;
3) *Розкладання* $\tilde{f}_{\mathbf{X}}$ *і* $K_{\mathbf{XX}}$, *отримання* $\tilde{\mathbf{A}}[m,n]$ *і* $\tilde{\mathbf{B}}[m,n]$;
4) *Генерування датасету;*
5) *Визначення його* $\hat{f}_{\mathbf{X}}$ *і отримання відповідних* $\hat{\mathbf{A}}[m,n]$;
6) *Знаходження варіації розподілу* $\Delta\mathbf{A} = h \cdot \left(\hat{\mathbf{A}}[m,n] - \tilde{\mathbf{A}}[m,n] \right)$, *де* $0 < h \le 0{,}5$ - *крок ітерації і відповідного розподілу* f_Δ *з коефіцієнтами розкладання* $\mathbf{A}_\Delta = \hat{\mathbf{A}} + \Delta\mathbf{A}$;
7) *Знаходження необхідного нелінійного перетворення для виконання кроку ітерації шляхом розв'язання функціонального рівняння відносно* $\varphi(\mathbf{X})$;
8) *Нелінійне перетворення датасету, визначення його* $\hat{K}_{\mathbf{XX}}$ *і відповідних* $\hat{\mathbf{B}}[m,n]$;
9) *Знаходження варіації кореляції* $\Delta\mathbf{B} = h \cdot \left(\hat{\mathbf{B}}[m,n] - \tilde{\mathbf{B}}[m,n] \right)$ *і відповідного кореляційної структури* R_Δ *з коефіцієнтами розкладання* $\mathbf{B}_\Delta = \hat{\mathbf{B}} + \Delta\mathbf{B}$;

10) Знаходження необхідного лінійного перетворення для виконання кроку ітерації шляхом розв'язання функціонального рівняння $R_{\Delta}(\tau) = \int_0^{\infty} \hat{R}(\theta) g(\theta - \tau) d\theta$ *відносно* $g(\theta)$;

11) Виконання лінійного перетворення з імпульсною передатною функцією $g(\theta)$;

12) Повернення до п. 5) доки $\left(\left|\Delta \mathbf{A}\right| > \varepsilon\right)$ *or* $\left(\left|\Delta \mathbf{B}\right| > \varepsilon\right)$, *де* ε *- допустима похибка.*

Висновки

Запропонований підхід дозволяє синтезувати статистично оптимальний датасет для навчання нейронних контролерів систем керування, проте необхідно провести додаткові дослідження збіжності алгоритму перетворення даних для різних типів розподілів ймовірностей та кореляційних структур.

Список використаної літератури

[1] V. Dubovoi, M. Yukhimchuk, V. Kovtun, K. Grochla. Functional Dependability of Distributed Control of Multi-zone Objectsunder Failures Conditions. In IEEE Access, vol. 12, pp. 95736-95749, 2024, doi: 10.1109/ACCESS.2024.3421380.

[2] Xin Huang, Han Lin Shang. Nonlinear autocorrelation function of functional time series (2022) DOI: https://doi.org/10.21203/rs.3.rs-1592981/v1

[3] Boyd S, El Ghaoui L., Feron E., BalakrDuffy, Daniel. (2018). Random Number Generation and Distributions. 819-852. 10.1002/9781119170518.ch26.

[4] Liu, Mengwei & Reed, Patrick & Anderson, C.. (2021). Stochastic Synthetic Data Generation for Electric Net Load and Its Application. 10.24251/HICSS.2021.383.

[5] Salim, Omar & Dorrah, Hassen & El-kahawy, Mahmoud. (2018). A Novel Algorithm to Generate Synthetic Data for Continuous-State Stationary Stochastic Process (Wind Data Application). 333-338. 10.1109/MEPCON.2018.8635270.

[6] Wonseok Yang, Woochul Nam/ Data synthesis method preserving correlation of features. Volume 122, 2022, 108241

[7] Kang, Yanfei & Hyndman, Rob & Li, Feng. (2019). GRATIS: GeneRAting TIme Series with diverse and controllable characteristics. 10.48550/arXiv.1903.02787.

[8] Nicklas Javergard, Rainey Lyons, Adrian Muntean1 and Jonas Forsman. Preserving correlations: A statistical method for generating synthetic data. arXiv:2403.01471v1 [cs.LG] 3 Mar 2024

[9] Kai Vahldiek, Libing Zhou, Wenfeng Zhu and Frank Klawonn. Development of a Data Generator for Multivariate Numerical Data with Arbitrary Correlations and Distributions. 2021,Intelligent Data Analysis,25(4) pp.789-807; DOI:10.3233/IDA-205253.

[10] Cavalcante, Charles & Mota, joão & Romano, João. (2004). Polynomial expansion of the probability density function about Gaussian mixtures. Machine Learning for Signal Processing XIV - Proceedings of the 2004 IEEE Signal Processing Society Workshop. 163 - 172. 10.1109/MLSP.2004.1422970.

[11] Gambaro, A.M. Exponential expansions for approximation of probability distributions. *Decisions Econ Finan* (2024). https://doi.org/10.1007/s10203-024-00460-2

[12] Birrell, Jeremiah & Katsoulakis, Markos & Pantazis, Yannis. (2022). Optimizing Variational Representations of Divergences and Accelerating Their Statistical Estimation. IEEE Transactions on Information Theory. 10.1109/TIT.2022.3160659.

Volodymyr Dubovoi[1]

Synthesis of a statistically optimal dataset for training a neural networks

[1] Vinnytsia National Technical University

To control complex nonlinear systems, intelligent technologies, in particular, based on neural networks, are increasingly used. However, training such neural controllers is complicated by the absence of labeled datasets for controlled objects at the stage of creating the system. This work investigates model-oriented training of a neural controller and the creation of an initial training dataset based on a simulation model of the system. For this purpose, the problem of synthesizing optimal statistics of input data of the simulation model according to the criterion of minimum average losses is posed and solved. The problem of generating a dataset with specified statistical characteristics for training a controller of a dynamic control system is complicated by the fact that it is necessary to simultaneously provide both the necessary multidimensional probability distribution in the space of realizations of the vector influence on the system and its correlation structure in time. The mean square error of the neural controller is chosen as the loss function. Averaging is performed both over a set of realizations and over time to take into account the dynamics of the system. The simulation model of the decentralized system is created on the Scilab/Xcos platform using a pre-created library of blocks for modeling distributed systems. The task of generating a dataset with a given probability distribution and correlation structure is solved by generating an uncorrelated uniformly distributed dataset with subsequent nonlinear dynamic transformation. Simultaneously obtaining a given type of statistical distribution and correlation structure is obtained by iterative means with gradual approximation to the required result.

Keywords: neural network training, dataset synthesis, statistical optimality

Dubovoi Volodymyr Mykhailovych - Dr. Tech. of Sciences, professor, professor of the department of computer control systems, e-mail v.m.dubovoy@gmail.com