ПОКРАЩЕННЯ ЧУТЛИВОСТІ ОНЛАЙН-ЕКСПЕРИМЕНТІВ ЧЕРЕЗ ВИКОРИСТАННЯ ВЕКТОРНИХ ПРЕДСТАВЛЕНЬ СЛІВ

П.О. БІЛІНСЬКИЙ

Робота присвячена дослідженню задачі редукції дисперсії в контрольованих онлайн-експериментах (A/B тестах).

Модель контрольованого онлайн-експерименту застосовується великими цифровими компаніями для оптимізації продукції та покращення досвіду користувача по всьому світу [2]. Для перевірки гіпотези про наявність ефекту зазвичай використовується *середній ефект втручання* τ:

$$\tau = \frac{1}{n}\sum_{i=1}^{N}\mathbb{E}\left[Y_i(1) - Y_i(0)\right] \tag{1}$$

Стандартна статистична оцінка τ в експерименті - це різниця середніх між контрольною та тестовою групами $\Delta = \overline{Y}_1 - \overline{Y}_0$. Чутливість експерименту (здатність виявляти істинний ефект при певному рівні шуму) найбільше залежить від дисперсії оцінки, тому важливо її мінімізувати.

Задача зменшення дисперсії полягає в знаходженні статистичної оцінки для 1, що мінімізує дисперсію в класі можливих оцінок. Існуючі методи редукції дисперсії включають метод контрольних варіатів (CUPED), метод сурогатів, метод стратифікації [1][3][2]. Найбільш ефективним та універсальним є CUPED, котрий використовує оцінку 2. Проте його застосування обмежене лише випадками, коли для спостережень в експерименті наявна апріорна інформація, що на практиці часто не справджується.

$$\widehat{Y}_{cv} = \overline{Y} - \theta\overline{X} + \theta\mathbb{E}\left[X\right] \tag{2}$$

Метою дослідження є модифікація існуючих методів для застосування в задачах, де не виконується припущення про наявність апріорної інформації. Використовується інформація у вигляді нечітких індикаторів, що, за припущенням, містяться в даних текстових пошукових запитів. Створення нечіткої стратифікації спостережень на основі текстових даних дає можливість отримати коваріати X, які є оцінкою апріорної поведінки користувача.

В ході дослідження було розроблено методи генерації синтетичних наборів даних та симуляції контрольованих онлайн-експериментів методом

Монте-Карло. Набори даних симулюють реальні дані пошуку роботи в мережі та містять значний рівень шуму, що перешкоджає виявленню коваріат X.

Побудовано моделі, що виявляють інформативні коваріати X із даних та здійснюють редукцію дисперсії методом CUPED. Перший кроком є обчислення векторних представлень слів, що дозволяє згрупувати семантично схожі запити користувачів (модель sBERT). На основі векторів представлень визначаються кластери, які використовуються методом CUPED для редукції дисперсії. Для випробування якості роботи моделей були проведені експерименти із застосуванням до синтетичних наборів даних. В результаті найкраща модель зменшила дисперсію на 11.4% (Таб. 1).

Модель	**Редукція дисперсії,** %
Binary indicators	11.5%
sBERT + K-Means	6.2%
sBERT + Fuzzy K-Means	6.9%

Табл. 1. Результати експериментів

Запропонований метод не досягає високих показників редукції дисперсії впорівнянні з існуючими методами (CUPED та метод сурогатів досягають 50%) [3][1]. Проте він має застосування в більш широкому колу задач, де існують перешкоди для застосування стандартних методів.

Література

[1] Deng A., Xu Y., Kohavi R., Walker T. *Improving the Sensitivity of Online Controlled Experiments by Utilizing Pre-Experiment Data* // Proceedings of the Sixth ACM International Conference on Web Search and Data Mining. — 2013. — С. 123–132.

[2] Larsen N., Stallrich J., Sengupta S., Deng A., Kohavi R., Stevens N. *Statistical Challenges in Online Controlled Experiments: A Review of A/B Testing Methodology* // The American Statistician. — 2023. — Т. 78, № 2. — С. 135–149.

[3] Deng A., Du M., Matlin A., Zhang Q. *Variance Reduction Using In-Experiment Data: Efficient and Targeted Online Measurement for Sparse and Delayed Outcomes* // Proceedings of the Sixth ACM International Conference on Web Search and Data Mining. — 2023. — С. 3937–3946.

Національний університет "Києво-Могилянська академія", Київ, Україна
Email address: p.bilinskyi@ukma.edu.ua