Методи регуляризації в задачах кластеризації

Виконав Міщеня Володимир Володимирович
Науковий керівник Крюкова Галина Віталіївна

Актуальність

Кластеризація – важливий інструмент аналізу даних у багатьох галузях.

Стандартні методи кластеризації мають обмеження: чутливість до шуму, нерелевантних ознак, перенавчання.

Регуляризація дозволяє підвищити стійкість, точність та узагальнюючу здатність моделей кластеризації.

Зростання обсягів даних вимагає ефективних інструментів для їх обробки.

Мета роботи

Дослідження та впровадження методів регуляризації у задачах кластеризації для підвищення їхньої стійкості, точності та здатності до узагальнення при роботі з великими обсягами даних, що характеризуються нерівномірністю розподілу та наявністю нерелевантних даних.

Завдання роботи

01

Аналіз теоретичних основ кластеризації та регуляризації.

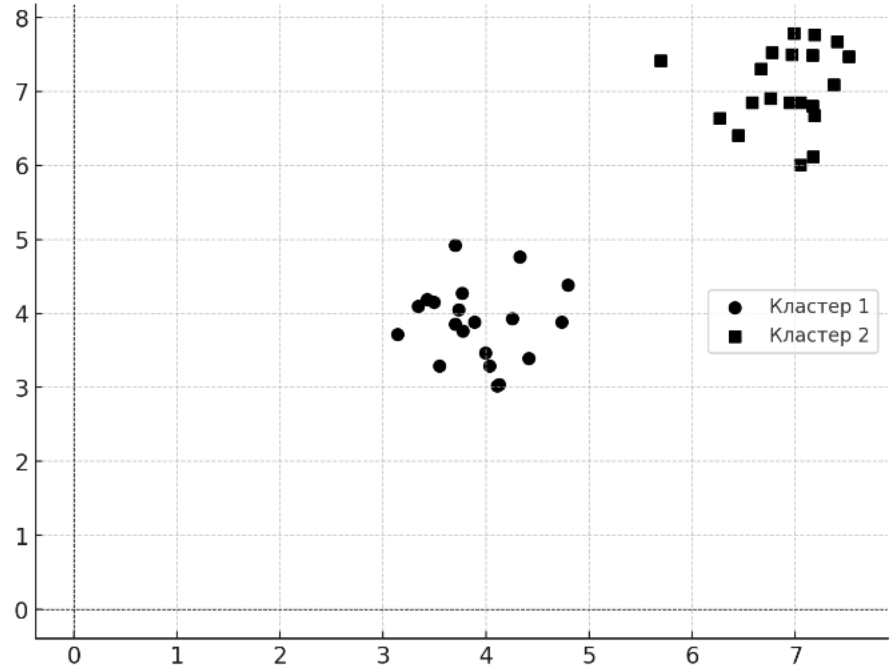
02

Огляд способів впровадження регуляризації в метод k-середніх.

03

Експериментальне порівняння регуляризаційних підходів.

Кластеризація



Евклідова відстань

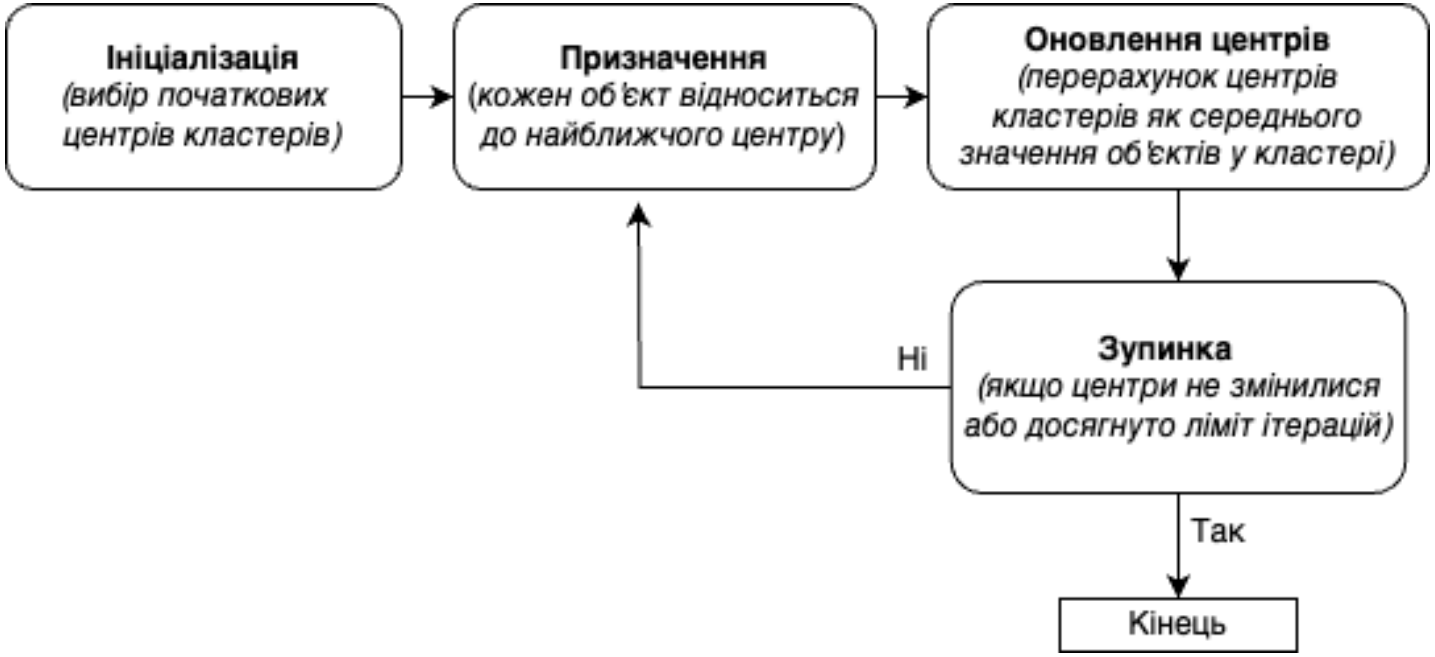
$$\rho(x, x') = \sqrt{\sum_{i=1}^{n} (x_i - x'_i)^2}$$

Метод К-середніх

- Популярний, простий, ефективний
- Мета: мінімізація суми квадратів відстаней від точок до центрів їх кластерів

$$\min_{A_k, C_k} \sum_{k=1}^{K} \sum_{X_i \in A_k} \rho(X_i, C_k)^2$$

- Неконвексна задача



Алгоритм Ллойда

Оцінка якості кластеризації

Внутрішньокластерна відстань

Середня відстань між кожною точкою кластера та центром цього кластера

$$\frac{1}{K}\sum_{k=1}^{K}\frac{1}{|A_k|}\sum_{X_i\in A_k}\rho(X_i,C_k)^2$$

Середня міжкластерна відстань

Середня відстань між усіма парами центрів кластерів

$$\frac{2}{K(K-1)}\sum_{1\le k<j\le K}\rho(C_k,C_j)$$

Індекс Давіса-Болдіна

Середнє відношення внутрішньокластерної відстані до міжкластерної відстані

$$\frac{1}{K}\sum_{k=1}^{K}\max_{j\neq k}\left(\frac{S_k+S_j}{\rho(C_k,C_j)}\right)\ ;\quad S_k=\frac{1}{|A_k|}\sum_{X_i\in A_k}\rho(X_i,C_k)$$

Силуетний коефіцієнт

Наскільки добре кожна точка вписується у свій кластер, порівняно з найближчим альтернативним кластером

$$\frac{1}{n}\sum_{i=1}^{n}\frac{b_i-a_i}{\max(a_i,b_i)},\qquad a_i=\frac{1}{|A_k|-1}\sum_{\substack{X_j\in A_k\\ j\neq i}}\rho(X_i,X_j)$$

$$b_i=\min_{k'\neq k}\left(\frac{1}{|A_{k'}|}\sum_{X_j\in A_{k'}}\rho(X_i,X_j)\right)$$

Регуляризація

Метод, метою якого є стабілізація розв'язків, усунення неоднозначності в некоректно поставлених задачах та покращення здатності моделей до узагальнення, що досягається шляхом включення до задачі додаткових припущень, обмежень або штрафів

$$\hat{\theta} = \arg\min_{\theta} \left[\mathcal{L}(\theta; X, y) + \lambda \cdot \Omega(\theta) \right]$$

$\mathcal{L}(\theta; X, y)$ - *основна функція втрат*

$\Omega(\theta)$ - *регуляризаційний штраф*

Регуляризація в к-середніх

Основна ідея: відсіювати зайві ознаки та стабілізувати положення центрів кластерів, покращуючи точність і стійкість результатів.

$$\min_{A_k, C_k} \sum_{k=1}^{K} \left(\sum_{X_i \in A_k} \rho(X_i, C_k)^2 \right) + \lambda P(C)$$

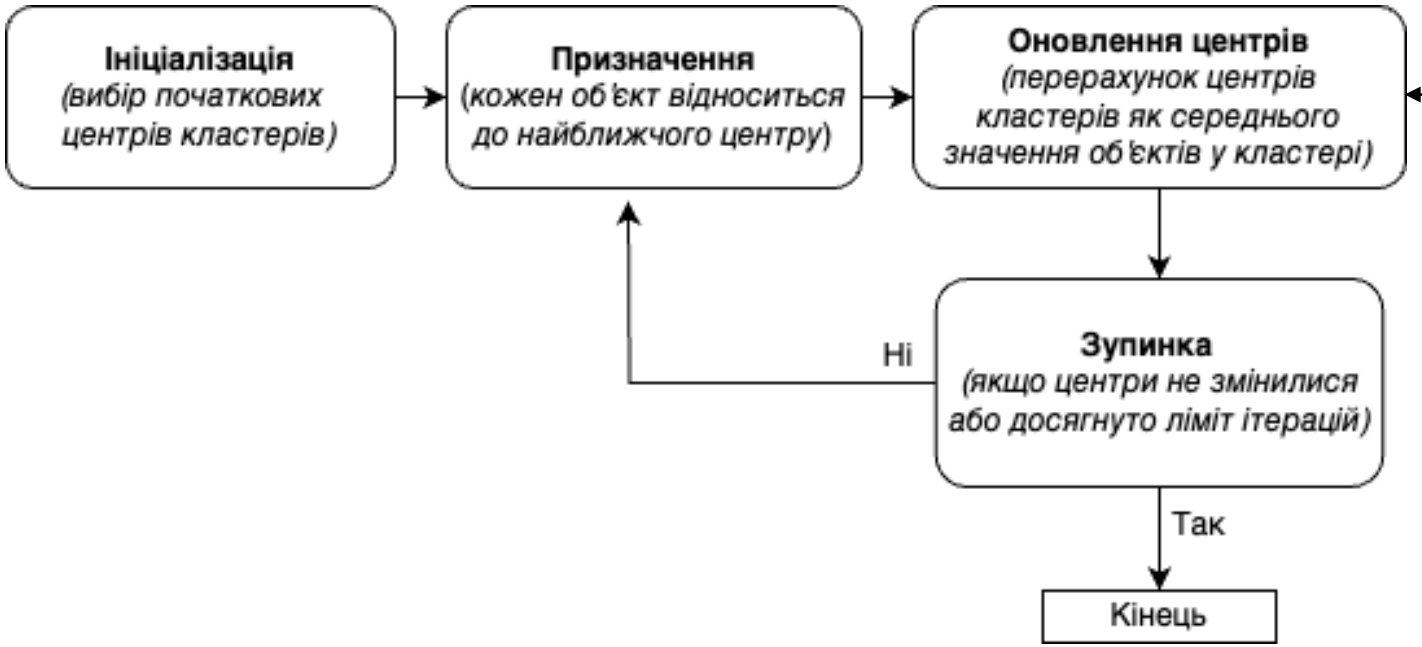
$P(C)$ — штрафна функція, що залежить від матриці центроїдів C

$\lambda \geq 0$ — параметр регуляризації, що контролює силу штрафу.

Типи штрафних функцій

Регуляризація	Формула штрафу	Механізм впливу	Особливості
L0 (Hard-Thresholding)	$\sum_{j=1}^{p} I(\|C_{\cdot,j}\|_2 > 0)$	Повне включення або виключення ознаки	Чіткий відбір ознак, висока інтерпретованість
L1 (Lasso)	$\sum_{j=1}^{p} \|C_{\cdot,j}\|_1$	М'яке обнулення окремих координат центрів	Розрідженість без повного виключення ознак
L2 (Ridge)	$\sum_{j=1}^{p} \|C_{\cdot,j}\|_2^2$	Гладке зменшення центрів до нуля	Не здійснює відбору ознак, стабілізує модель
Group Lasso	$\sum_{j=1}^{p} \|C_{\cdot,j}\|_2$	Стискає координати ознаки	Може обнулити всю ознаку

Алгоритм Ллойда для регуляризованого k-середніх



Hard-Thresholding

$$C_{k,j} = \begin{cases} C^*_{k,j}, & \text{якщо} \quad \|X\|_2^2 > \rho(X, MC^*_{\cdot,j})^2 + n\lambda \\ 0, & \text{інакше.} \end{cases}$$

Lasso

$$C_{k,j} = \max\left(0, 1 - \frac{n\lambda}{2|A_k|\left|C^*_{k,j}\right|}\right) C^*_{k,j}$$

Ridge

$$C_{k,j} = \frac{1}{1 + \frac{n\lambda}{|A_k|}} C^*_{k,j}$$

Group Lasso

$$C_{k,j} = \frac{1}{1 + \frac{n\lambda}{2|A_k|\|C^*_{\cdot,j}\|_2}} C^*_{k,j}$$

Вибір параметра λ

Інформаційні критерії AIC/BIC

$$\mathrm{AIC} = \mathrm{WCSS} + 2Kp_{\mathrm{sel}},$$

$$\mathrm{BIC} = \mathrm{WCSS} + Kp_{\mathrm{sel}} \ln(n).$$

Gap-метод

$$\Delta_{q+1} = \frac{1}{n}\mathrm{WCSS}_{q+1} - \frac{1}{n}\mathrm{WCSS}_q$$

$$\downarrow$$

$$D_{q+1} = \frac{m - \Delta_{q+1}}{s}$$

q — кількість вже включених ознак

m і s — середнє та стандартне відхилення отриманих Δ^*_{q+1}

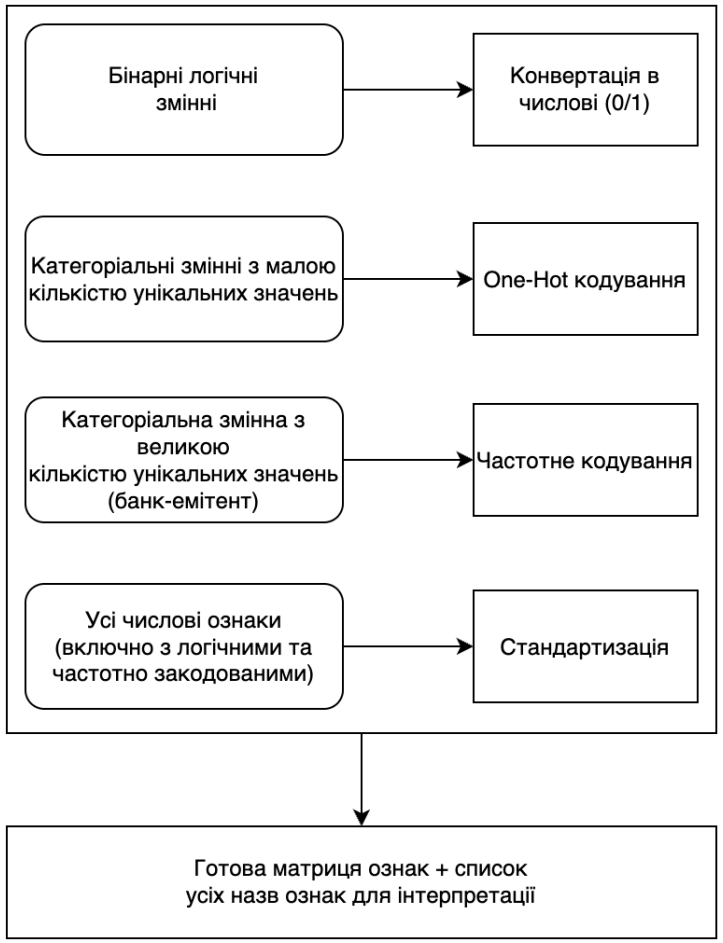
WCSS — сумарна внутрішньокластерна дисперсія

Практична задача

В IT-компанії, що розробляє підписочні додатки, виникає необхідність сегментувати користувачів за рівнем ризику створення дружнього фроду, щоб покращити управління підписками та мінімізувати фінансові ризики.

Назва поля	Опис
user uuid	Унікальний ідентифікатор користувача
payment method	Метод оплати користувача
trial period	Тривалість пробного періоду
group country	Група країни користувача
gender	Стать користувача
age group	Вікова група користувача
login	Факт логіну в додаток
source	Джерело реєстрації (наприклад, реклама)
has cons	Придбання додаткової консультації
has otp upsell	Придбання додаткового матеріалу
rebill number	Кількість повторних платежів
card type	Тип картки (дебетова/кредитна)
issuing bank	Банк-емітент картки
card brand	Бренд картки
fraud count	Кількість шахрайських повідомлень
income count	Кількість здійснених платежів

Попередня обробка та додаткова оцінка



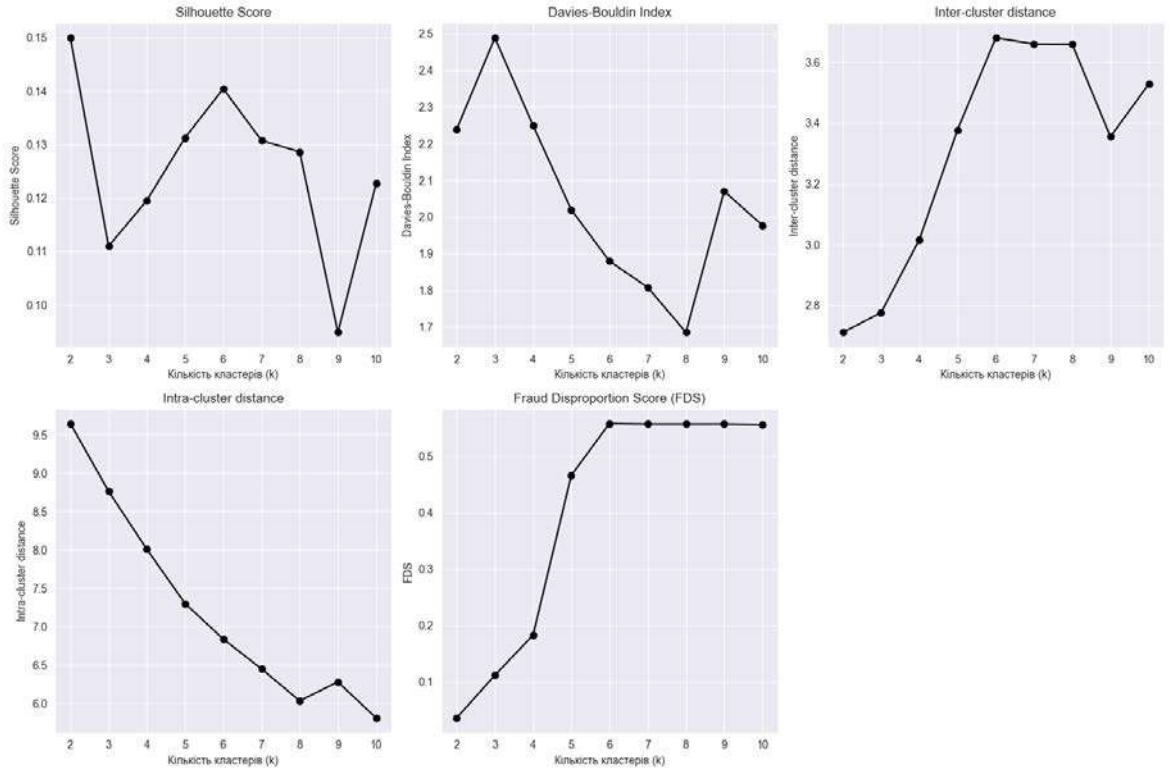
Fraud Disproportion Score

$$\text{FDS} = \sum_{k=1}^{C} \left| \frac{p_k}{TP} - \frac{f_k}{TF} \right|$$

- TP — загальна кількість платежів усіх користувачів.
- TF — загальна кількість випадків фроду серед усіх користувачів.
- p_k — кількість платежів в кластері k.
- f_k — кількість випадків фроду в кластері k.

Вибір параметрів

Оптимальна кількість кластерів k=6



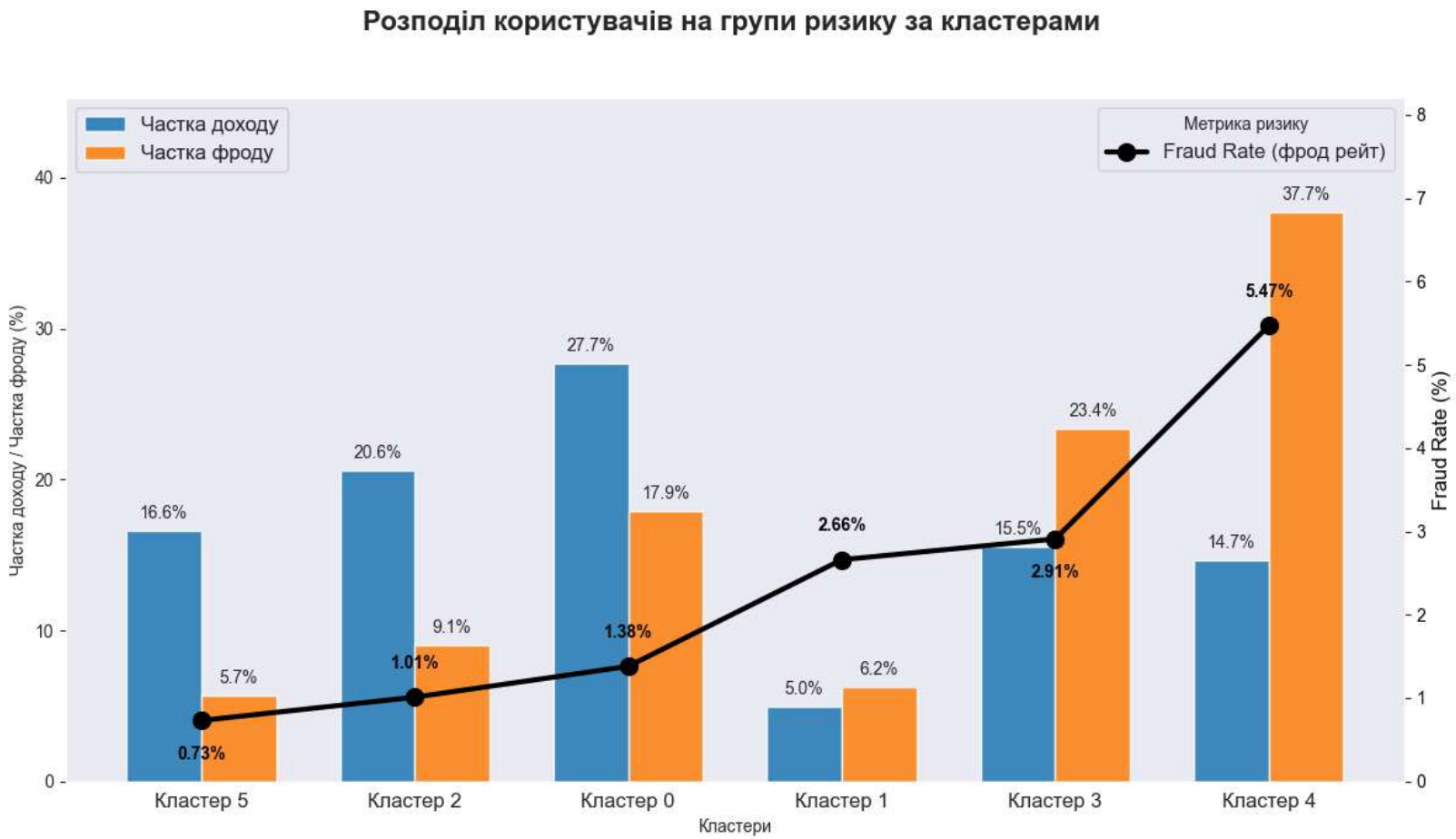
Оптимальний параметр λ

Метод	**Параметр регуляризації (λ)**
Hard-Thresholding	0.2
Lasso	0.15
Ridge	0.13
Group Lasso	0.1

Порівняльний аналіз

Метод	Силуетний коеф.	Давіса–Болдіна Індекс	Внутрішньокластерна відстань	Міжкластерна відстань	FDS
k-середніх	0.1404	1.8793	6.8393	3.6802	0.5579
Hard-Thresholding	0.1046	1.8945	5.0475	3.6814	0.5702
Lasso	0.1536	1.1294	3.5848	3.4052	0.6199
Ridge	0.1417	1.5522	4.9604	1.9444	0.5781
Group Lasso	0.1466	1.5336	4.2568	3.4556	0.6135

Результат сегментації користувачів



Список літератури

[1] J. A. Hartigan, M. A. Wong *Algorithm AS 136: A K-Means Clustering Algorithm,* // Journal of the Royal Statistical Society. Series C (Applied Statistics), 1979. — Vol. 28, — P. 100–108.

[2] Abiodun M. Ikotun, Absalom E. Ezugwu, Laith Abualigah, Belal Abuhaija, Jia Heming *K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data,* // Information Sciences 2023. — Vol. 622, — P. 178–210.

[3] Christopher Holder, Anthony Bagnall, Jason Lines *On time series clustering with k-means,* // School of Electronics and Computer Science, University of Southampton, Southampton, UK, 2024.

[4] David MacKay *An Example Inference Task: Clustering,* // Information Theory, Inference, and Learning Algorithms, 2003. — Vol. 20, — P. 284–292.

[5] David Arthur, Sergei Vassilvitskii *k-means++: The Advantages of Careful Seeding* // Stanford University 2020.

[6] Ali Idrus, Nafan Tarihoran, Ucup Supriatna, Ahmad Tohir, Suwarni Suwarni, Robbi Rahim *Distance Analysis Measuring for Clustering using K-Means and Davies Bouldin Index Algorithm* // TEM Journal, 2022. — Vol. 11, — P. 1871–1876.

[7] Peter J. Rousseeuw *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis* // Journal of Computational and Applied Mathematics, 1987. — Vol. 20, — P. 53–65.

[8] Jakob Raymaekers, Ruben H. Zamar *Regularized K-means through hard-thresholding* // University of British Columbia, Department of Statistics 2020.

[9] Rafail Ostrovsky, Yuval Rabani, Leonard J. Schulman, Chaitanya Swamy *The Effectiveness of Lloyd-Type Methods for the k-Means Problem* // University of Waterloo 2020.

Дякую за увагу!