

Методи стохастичної квантизації для масштабованого кластерного аналізу великих даних

за спеціальністю 113 Прикладна математика
(галузь знань 11 Математика та статистика)

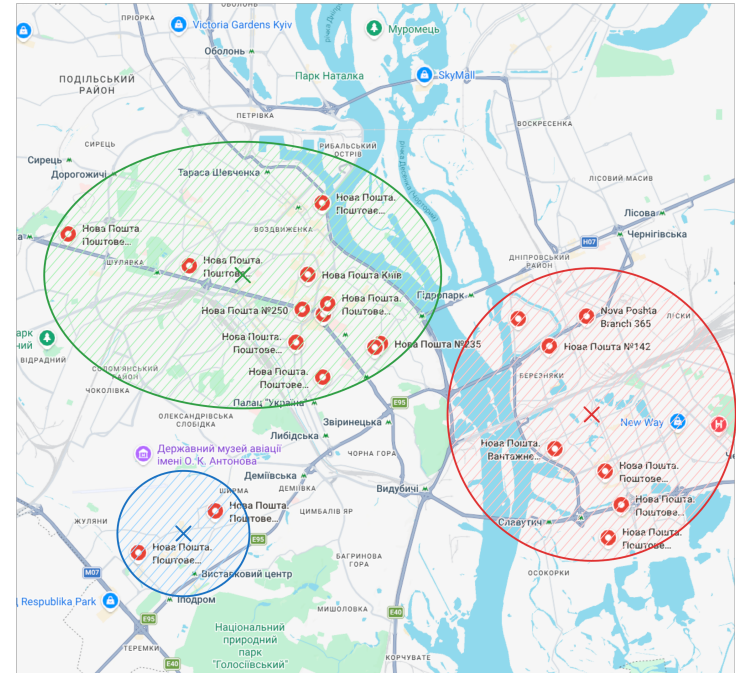
Аспірант: Козирєв А. Ю.

Науковий керівник: др. фіз-мат наук, с.н.с., Норкін В. І.



Актуальність

1. Кластерний аналіз є задачею машинного навчання без учителя, що активно застосовується в задачах сегментації ринку [Tabianan (2022)], логістики, обробки зображень [Jegou (2011)] тощо;
2. В залежності від задачі застосовуються різні типи методів кластеризації:
 - На основі центроїдів (K-means);
 - На основі щільності (DBSCAN, HDBSCAN);
3. Актуальна задача, що розглядається в дослідженні — кластеризація неструктурованих медіа даних (зображень, тексту тощо) за їх латентними (векторними ознаками)



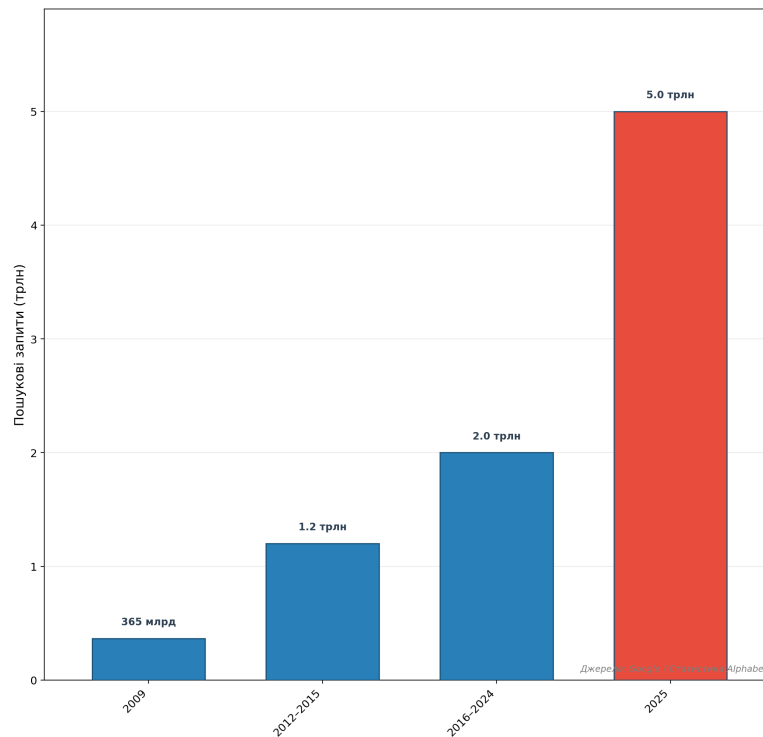
Яскравий приклад застосування методів кластеризації в логістиці — планування розподільних центрів для пунктів доставки за умов обмеження на бюджет.



Актуальність

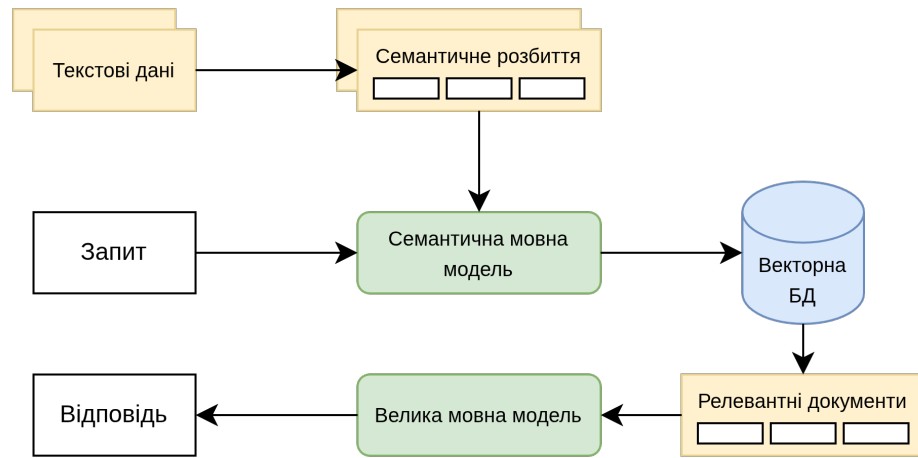
1. Cisco показав у звіті, що глобальний IP-трафік досягнув річного рівня 3,3 зеттабайта у 2021 році, порівняно з 1,2 зеттабайта у 2016 році [\[Cisco Visual Networking Index\]](#);
2. Google обробляє понад 5 трильйонів пошукових запитів на рік. Це перший випадок, коли Google публічно оприлюднив таку цифру з 2016 року, коли компанія підтвердила, що обробляє «понад 2 трильйонів» запитів щорічно [\[Google Commerce\]](#);
3. Сучасні системи пошукових рушіїв використовують алгоритми пошуку за семантичними латентними ознаками у векторних БД.

Кількість пошукових запитів Google за роками



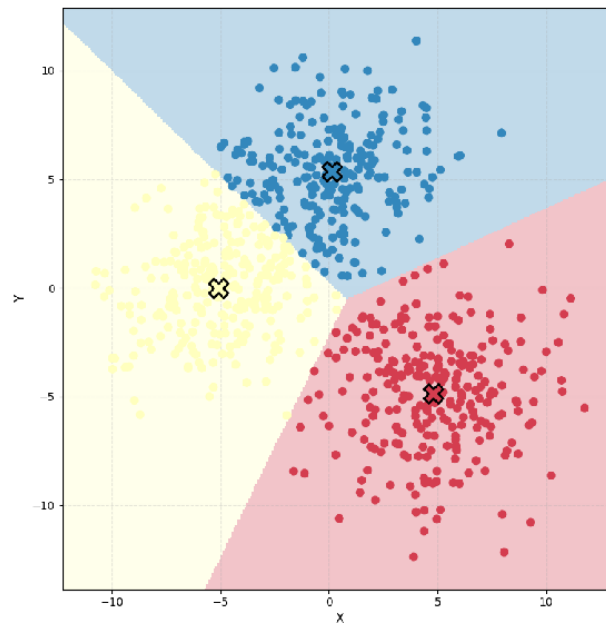
Системи пошукових рушіїв з векторними БД

1. Сучасні архітектури пошукових рушіїв використовують семантичну нейронну мережу для перетворення дискретної множини тексту (або зображень) у метричний простір, де метрика - семантична схожість;
2. Для прискорення пошуку найближчих сусідів використовується індексація даних, так званий інвертований файловий індекс (IVF).
3. Від дозволяє об'єднувати дані по значенню індексу, тому при пошуку даних спочатку знаходиться релевантний запиту індекс, а потім здійснюється пошук в даних з цим індексом;



Інвертований файловий індекс (IVF)

- **Інвертований файловий індекс** - це індекс бази даних, який зберігає відображення (мапінг) від вмісту, наприклад слів або чисел, до їхніх місцезнаходжень у таблиці, документі або наборі документів (названий на протизвагу прямому індексу, який відображає документи на їхній вміст).
- **Структурно інвертований файл** - це індексна структура даних, яка відображає вміст на його розташування в файлі бази даних, у документі або наборі документів. Зазвичай він складається з:
 - а. словника (vocabulary), що містить усі унікальні слова, знайдені в тексті
 - б. для кожного слова t зі словника - інвертованого списку (inverted list), що містить статистичні дані про входження t у тексті.



Нейронні мережі із зовнішнім механізмом пам'яті використовують джерела інформації (БД, текстові документи тощо) для покращення власного контексту та підвищення точності генерування відповідей



Алгоритм K-means [Lloyd (1982)]

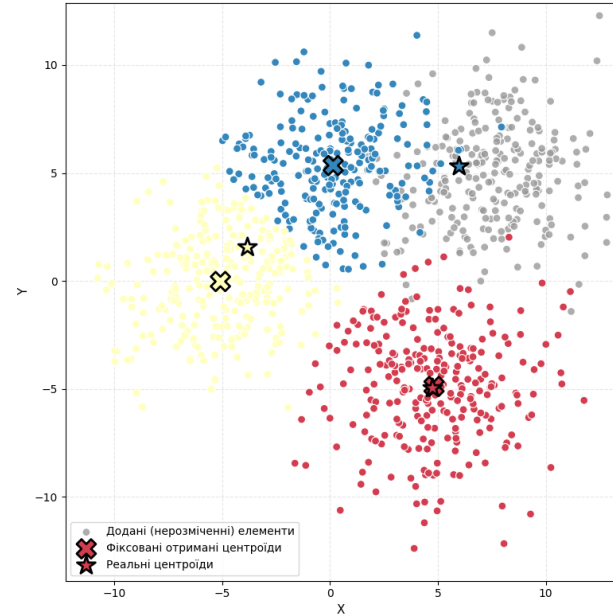
1. **Ініціалізація:** випадковий вибір початкових k центроїдів $\{Y_j\}$;
2. **Групування:** кожен елемент ξ_i призначений кластеру його найближчого поточного центроїда

$$\xi_i \in C_j, \quad j = \arg \min_{1 \leq \ell \leq k} \|\xi_i - Y_\ell\|^2$$

3. **Оновлення:** Обчислення положення $\{Y_j\}$ як середнє кожної точки в кластері C_j :

$$Y_j \leftarrow \frac{1}{\|C_j\|} \sum_{\xi_i \in C_j} \xi_i$$

Основна проблема класичного метода: необхідність переобчислення центроїдів на всій вибірці при надходженні нових даних.



Проблема алгоритму K-means — неадаптивність в умовах потоків даних. Даний приклад демонструє зміщення очікуваних центроїдів від отриманих при додаванні даних до вибірки (сезонні фактори, соціологічні змінні тощо)



Актуальна задача дисертаційного дослідження

Актуальна наукова задача, що розв'язується в дисертації, - це розробка методів стохастичної векторної квантизації для ефективного кластерного аналізу Big Data (кількість екземплярів 10^6 – 10^9 і більше)

Метою роботи є розробка стохастичного методу кластерного аналізу на основі центроїдів з використанням стохастичної апроксимації та субградієнтної оптимізації, який вимагає менше пам'яті обчислювального пристрою та має кращу швидкість збіжності порівняно з існуючими методами.

Об'єктом дослідження є методи кластеризації на основі центроїдів, методи стохастичної негладкої оптимізації для розв'язку задачі.

Предметом дослідження є метод стохастичної квазі-градієнтної кластеризації (стохастичної квантизації) та модифікацій з адаптивним кроком, а також показники оцінки точності та швидкості збіжності методів стохастичної оптимізації.



Постановка задачі досліджень

Для досягнення вказаної мети у дисертаційній роботі вирішуються такі **задачі**:

1. Розробка метода стохастичної векторної квантизації на основі адаптивних стохастичних градієнтних методів (метод Поляка, метод Нестерова, Adagrad, RMSProp, ADAM):
2. Розробка методів оптимізації негладких неопуклих цільових функції на основі згладжування (усереднення) функцій;
3. Теоретичне обґрунтування умов збіжності розроблених методів та практичне порівняння швидкості збіжності на прикладних задачах кластерного аналізу використовуючи дані з відкритих джерел: семантична кластеризація тексту і зображень та кольорова квантизація.



Наукова новизна

1. Розроблено **новий метод** стохастичної квазі-градієнтної кластеризації, який відрізняється від існуючих методів векторної квантизації використанням субдиференційованої цільової функції транспортної відстанні та стохастичної апроксимації, що дозволяє оцінювати оптимальні положення центроїдів лише з використанням підмножини навчальних даних, що зменшує необхідну кількість обчислювальних ресурсів обчислювального пристрою (оперативної пам'яті та обчислювального часу центрального процесора);
2. Розроблено **новий метод** оптимізації негладких неопуклих цільових функцій за допомогою методу згладжування (усереднення) функцій для пошуку точок екстремумів, використовуючи кінцево-різницеву апроксимацію градієнтів вздовж стохастичних напрямків;
3. Розроблено **модифіковану структуру IVF** для інкрементного індексування векторних даних, де для індексації використовується запропонований метод стохастичної квазі-градієнтної кластеризації, в умовах динамічної зміни відповідного розподілу даних (сезонні зміни, соціологічні фактори, тощо).



Наукова новизна

4. **Теоретично доведено** властивості запропонованих методів:
 - a. сублінійна швидкість збіжності стохастичного кінцево-різницевого метода для опуклих Ліпшицевих функцій;
 - b. асимптотична збіжність запропонованого метода стохастичної квазі-градієнтної кластеризації (SQC);
5. **Показано переваги** запропонованого метода SQC на задачах кластеризації текстових даних на основі семантичних ознак у формі їх векторних репрезентацій;
6. **Удосконалено метод кольорової квантизації** запропонованим методом для більш ефективного стиснення зображень з втратами з меншим використанням пам'яті обчислювального пристрою.



1. Математична модель векторної квантизації: транспортна задача з рухомими центрами

Постановка задачі

Прикладні задачі: індексація даних центроїдами, задачі логістичних центрів, розміщення сервісних центрів, дискретна апроксимація імовірнісних розподілів та ін.

Задача квантизації розглядається як мінімізація транспортної відстанні (або метрики Вассерштейна-Канторовича-Рубінштейна) між розподілами елементів $\Xi = \{\xi_i, i = 1, \dots, I\}$ з вагами p_i та рухомих центрів $Y = \{y_k, k = 1, \dots, K\}$ з вагами q_k :

$$\min_{y=\{y_1, \dots, y_K\} \subset \mathbb{R}^{nK}} \min_{q=\{q_1, \dots, q_K\}} \left[\min_{x=\{x_{ik}\}} \sum_{i=1}^I \sum_{k=1}^K d(\xi_i, y_k)^r x_{ik} \right]$$
$$\sum_{k=1}^K q_k = 1; \quad \sum_{k=1}^K x_{ik} = p_i, \quad i = 1, \dots, I; \quad \sum_{i=1}^I x_{ik} = q_k, \quad k = 1, \dots, K;$$
$$x_{ik} \geq 0, \quad y_k \in Y, \quad q_k \in Q \in [0, 1]$$

1. якщо y_k, q_k фіксовані - транспортна задача;
2. $y_k \in Y = \{\xi_i, i = 1, \dots, I\}$ - нелінійна змішана задача [\[Kuzmenko, Uryasev \(2019\)\]](#);
3. в даному випадку маємо $Q = [0, 1], Y = \mathbb{R}^n$ - задачу стохастичної квантизації для Big Data.



1. Математична модель векторної квантизації: транспортна задача з рухомими центрами

Зведення до задачі неопуклого стохастичного програмування

$$\min_{y=\{y_1, \dots, y_K\} \in Y^K \subset \mathbb{R}^{nK}} \min_{q=\{q_1, \dots, q_K\} \in \mathbb{R}_+^K} \min_{x=\{x_{ik} \geq 0\}} \sum_{i=1}^I \sum_{k=1}^K d(\xi_i, y_k)^r x_{ik} \quad \sum_{k=1}^K x_{ik} = p_i$$

$$\min_{y=\{y_1, \dots, y_K\} \in Y^K \subset \mathbb{R}^{nK}} \min_{q=\{q_1, \dots, q_K\} \in \mathbb{R}_+^K} \sum_{i=1}^I p_i \sum_{k=1}^K d(\xi_i, y_k)^r =$$

$$\min_{y=\{y_1, \dots, y_K\} \in Y^K \subset \mathbb{R}^{nK}} \sum_{i=1}^I p_i \left(\min_{1 \leq k \leq K} d(\xi_i, y_k)^r \right) \implies$$

$$\min_{y=\{y_1, \dots, y_K\} \in Y^K \subset \mathbb{R}^{nK}} F(y_1, \dots, y_K),$$

$$F(y) = \sum_{i=1}^I p_i \min_{1 \leq k \leq K} d(\xi_i, y_k)^r = \mathbb{E}_{i \sim p} \min_{1 \leq k \leq K} d(\xi_i, y_k)^r$$



1. Математична модель векторної квантизації: транспортна задача з рухомими центрами

Розв'язок за допомогою квазіградієнтного методу

Оскільки цільова функція задачі стохастичної квантизації $F(y)$ є узагальнено диференційованою, то існує субдиференціал $\partial F(y)$:

$$F(y) = F(y_1, \dots, y_K) = \sum_{i=1}^I p_i \min_{1 \leq k \leq K} \|y_k - \xi_i\|^r \rightarrow \min_y$$

$$\partial F(y) = \sum_{i=1}^I p_i \partial \min_{1 \leq k \leq K} \|y_k - \xi_i\|^r = \sum_{i=1}^I p_i \partial \|y_{k^*} - \xi_i\|^r, \quad k^* \in S(\xi_i) = \{\arg \min_{1 \leq k \leq K} \|y_k - \xi_i\|^r\}$$

Тоді існує рекурентне правило оцінки оптимальних параметрів задачі стохастичної квантизації:

$$y_k^{t+1} = y_k^t - \rho_t g_k(y^t) = \begin{cases} y_k^t - \rho_t r \|y_k^t - \xi_i^t\|^{r-2} (y_k^t - \xi_i^t), & k \in S(\xi_i^t), \\ y_k^t, & k \notin S(\xi_i^t), \end{cases}$$



1. Математична модель векторної квантизації: транспортна задача з рухомими центрами

Алгоритм стохастичної квазіградієнтної кластеризації

Дано множину елементів $\{\xi_i, i = 1, \dots, I\}$, задано гіперпараметри: $\rho, K, T, r=3$.

1. $y^0 = \{y_k^0, k = 1, \dots, K\}$ - ініціалізація центроїдів;

2. **for** $t \in [0, T-1]$ **do**:

a. $\xi^t \in \{\xi_i, i = 1, \dots, I\}$ - вибір випадкового елемента;

b. $y_k^t, k \in S(\xi^t)$ - пошук найближчого центра;

c. $g_k^t = r \|\xi^t - y_k^t\|^{r-2} (y_k^t - \xi^t)$ - оцінка квазіградієнта цільової ф-ї
для $k \in S(\xi^t)$;

d. $y_k^t := \Pi_Y(y_k^t - \rho g_k^t)$ - оновлення обраного центра для $k \in S(\xi^t)$;

3. **end for**



1. Математична модель векторної квантизації: транспортна задача з рухомими центрами

Оцінка збіжності запропонованого метода

Теорема. Нехай існує:

1. послідовність $\{y^t = (y_1^t, \dots, y_k^t)\}$, отримана рекурентним співвідношенням:

$$y_k^{t+1} = \Pi_Y(y_k^t - \rho_t g_k(\tilde{\xi}^t)), \Pi_Y(\cdot) = \arg \min_{y \in Y} \|\cdot - y\|$$

2. підмножина елементів $\{\xi_i\}$ та послідовність $\{\rho_t\}$:

$$\rho_t > 0, \sum_{t=0}^{\infty} \rho_t = \infty, \sum_{t=0}^{\infty} \rho_t^2 < \infty$$

Позначимо через $F(Y^*)$ множину значень F у критичних (стаціонарних) точках Y^* задачі, де $Y^* = \{y = (y_1, \dots, y_k); 0 \in \partial F(y) + N(y)\}$ і $N_Y(y)$ позначає нормальний конус до множини Y у точці y . Якщо $F(Y^*)$ не містить проміжків і множина Y є опуклим компактом, то з ймовірністю один $\{y^t\}$ збігається до зв'язної компоненти Y^* , а послідовність $\{F(y^t)\}$ має границю.

Оскільки цільова функція є узагальнено-диференційованою, то доведення збіжності отриманого метода напряду виводиться з теореми про збіжність методу узагальненого градієнта. [Norkin (2024)]

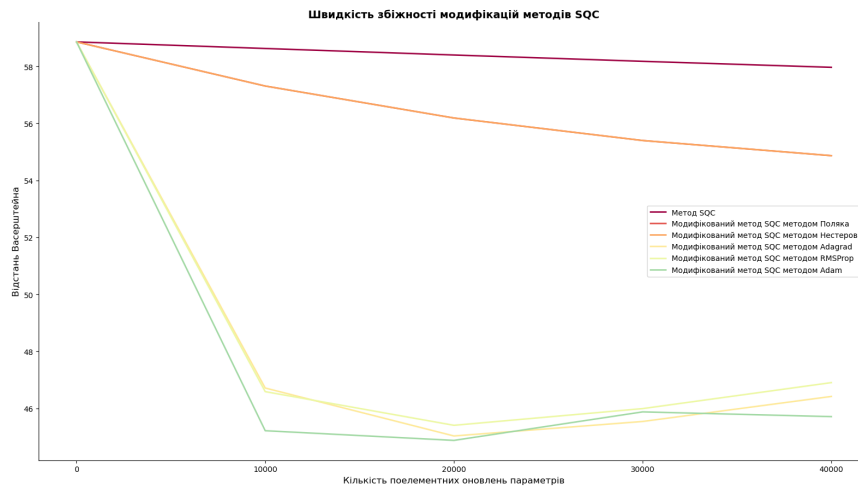


1. Математична модель векторної квантизації: транспортна задача з рухомими центрами

Модифікації з адаптивним кроком

Використання оцінки квазіградієнта цільової функції дозволяє модифікувати запропонований алгоритм для покращення швидкості збіжності з використанням:

1. модифікацій з множниками прискорення:
 - a. Метод Поляка;
 - b. Метод Нестерова;
2. модифікацій з адаптивним кроком спуску:
 - a. Adagrad;
 - b. RMSProp;
 - c. ADAM.



Порівняння кількості обчислень функції відстанні для кожної модифікації метода. Адаптивні методи мають перевагу у швидкості збіжності використовуючи нормалізацію кроку спуску



2. Метод оптимізації негладких цільових функцій використовуючи кінцево-різницеву апроксимацію

Визначення кінцево-різницевої апроксимації

Ліпшицева функція $f_\varepsilon(x)$ є диференційованою, її субдиференційована множина дорівнює очікуваному значенню диференційованого стохастичного наближення:

$$\nabla f_\varepsilon(x) = \frac{1}{\nu_n} \int_{B_h(x)} \partial f(y) dy$$

Ми можемо оцінити поверхневий інтеграл за допомогою алгоритму Монте-Карло на множині K рівномірно розподілених стохастичних векторів y на одній сфері $S_1(0) = \{y \in R^n: \|y\| = 1\}$. Тепер ми можемо подати згладжену функцію як кінцево-різницеву суму вздовж стохастичних векторів:

$$\begin{aligned} \nabla f_\varepsilon(x) &= \frac{n}{h} \mathbb{E}_{\bar{y}}[f(x + h\bar{y}) - f(x)]\bar{y} \\ &= \frac{n}{2h} \mathbb{E}_{\bar{y}}[f(x + h\bar{y}) - f(x - h\bar{y})]\bar{y} \\ &= \mathbb{E}_{\bar{y}}\left[\left|\frac{1}{2h}(f(x + h\bar{y}) - f(x - h\bar{y}))\bar{y}\right|^2\right] \leq L^2 \end{aligned}$$

Використовуючи це наближення, ми можемо замінити значення градієнта на згладжену функцію і отримати узагальнений алгоритм спуску за стохастичним градієнтом:

$$x_{k+1} = x_k - \rho_k \eta_k, \quad \eta_k = \frac{1}{2h_k} (F(x_k + h_k \bar{y}) - F(x_k - h_k \bar{y})) \bar{y}$$

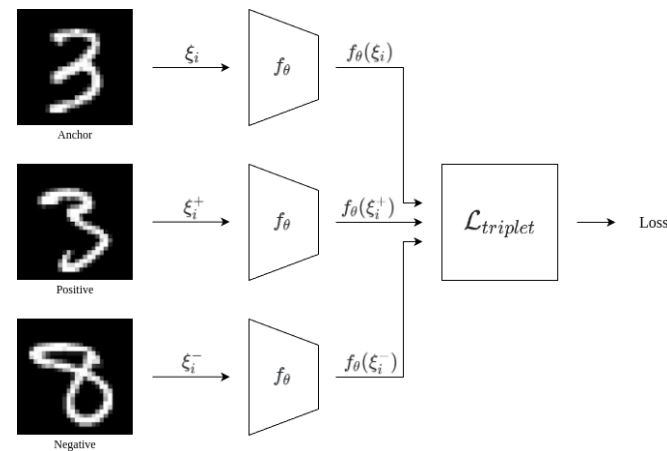


3. Експериментальне дослідження точності отриманого метода у порівнянні з класичним K-means

Порівняння швидкості індексації на датасеті ArXiv

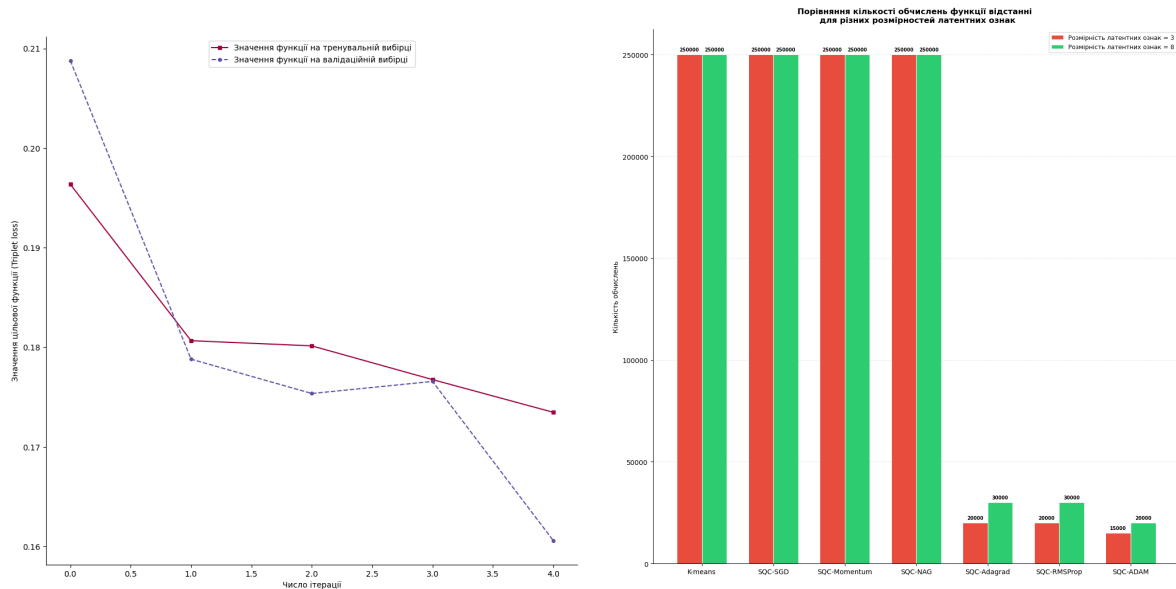
- Для порівняння швидкості кластеризації текстових даних (індексації IVF) було використано відкритий набір даних [\[ArXiv Metadata\]](#);
- Запропонований метод SQC та K-means проводили кластеризацію на латентних ознаках анотації статей ArXiv, згенерованих натренованою мережею Triplet Network;
- В даному експерименті Triplet Network - нейронна мережа архітектури Encoder-only Transformer, натренована на контрастній цільовій функції з гіперпараметрами $epoch = 5, batch = 1024, \rho = 10^{-3}, \lambda = 10^{-5}, \alpha = 1.0$:

$$\mathcal{L}_{triplet}(\theta) = \max(0, d(f_{\theta}(\xi_i), f_{\theta}(\xi_i^+)) - d(f_{\theta}(\xi_i), f_{\theta}(\xi_i^-)) + \alpha)$$



3. Експериментальне дослідження точності отриманого метода у порівнянні з класичним K-means

Тренування Triplet Network на датасеті ArXiv



Результати процесу індексації текстових даних ArXiv з використанням запропонованого метода SQC та Triplet Network: (1) зміна значень цільової функції triplet loss для тренувальної та валідаційної вибірки з кожною ітерацією, для запобігання перенавчання було застосовано early stopping (patience=2); (2) векторні репрезентації текстів, отримані Triplet Network, використані для кластеризації методом SQC, серед представлених модифікацій адаптивні методи мають найкращу швидкість збіжності



4. Прикладна задача: кольорова квантизація

Визначення задачі

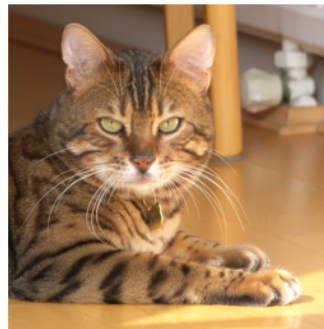
Кольорова квантизація - метод стиснення зображення із втратами використовуючи кластеризацію на основі центроїдів:

1. Представлення пікселів зображення в тривимірному просторі (вісі - інтенсивність кольорових каналів RGB);
2. Встановлення кількості кольорів оптимальної палітри та знаходження їх положення методом кластеризації;
3. Перефарбування кожного пікселя до кольору найближчого центроїда.

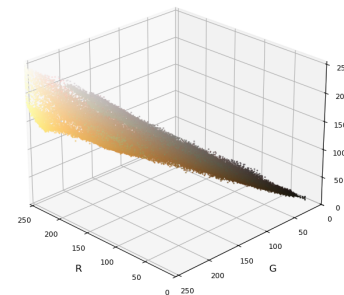
Сфери застосування:

4. Оптимізація веб-зображень;
5. Сегментація зображень;
6. Кольоровий друк.

Оригінальне зображення



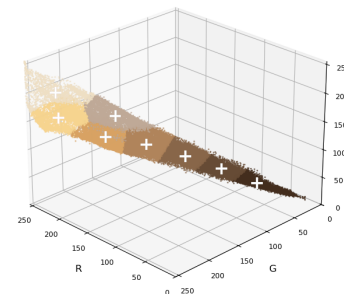
Представлення зображення як множини пікселів у просторі RGB



Стиснуте зображення з втратами (палітра в 8 кольорів)



Представлення стиснутого зображення як множини пікселів у просторі RGB (палітра в 8 кольорів)



Висновки

1. Запропонований новий стохастичний ітераційний метод кластеризації векторних даних на основі центроїдів, який на кожній ітерації виконує оновлення цетрів з використанням лише одного елемента з множини навчальних даних;
2. Запропонований метод є адаптований до задач кластерного аналізу в умовах потоків даних, тобто задач які моделюють сезонні змінні, соціологічні фактори тощо;
3. Запропонований метод кластеризації використовує аналогічні стохастичні апроксимації, що і в глибинному навчанні, що дозволяє масштабувати метод на рівень Big Data. Масштабованість метода продемонстрована на наборі даних ArXiv з вибіркою в $\simeq 10^5$ елементів;
4. В експериментальних результатах запропонований метод потребує в середньому використовує 10% обчислень функції відстані для збіжності за класичний метод кластерного аналізу K-means, який є стандартний методом кластеризації на основі центроїдів;



Основні публікації

Категорія видання, Назва праці	Видавництво, журнал (назва, номер, рік) чи номер авторського свідоцтва, бібліографічне посилання	Прізвища співавторів
Категорія видання: A, Scopus Квартиль видання: Q2 Назва праці: Constrained Global Optimization by smoothing.	Norkin, V., Pichler, A., & Kozyriev, A. (2025). Constrained Global Optimization by smoothing. <i>Lecture Notes in Computer Science</i> , Vol. 14476. P.136–150. Print ISSN 0302-9743. https://doi.org/10.1007/978-3-031-81241-5_10	Norkin, V., Pichler, A.
Категорія видання: A Назва праці: Сучасні стохастичні квазіградієнтні алгоритми оптимізації.	Норкін, В. І., Козирев, А. Ю. і Норкін, Б. В. (2024) «Сучасні стохастичні квазіградієнтні алгоритми оптимізації», Міжнародний науково-технічний журнал "Проблеми керування та інформатики", 69(2), с. 71–83. doi: 10.34229/1028-0979-2024-2-6.	Норкін, В. І., Норкін, Б. В.
Категорія видання: A Назва праці: Robust Clustering on High-Dimensional Data with Stochastic Quantization.	Kozyriev, A., & Norkin, V. (2024). Robust Clustering on High-Dimensional Data with Stochastic Quantization. <i>International Scientific Technical Journal "Problems of Control and Informatics,"</i> 70(1), pp. 32–48. doi: 10.34229/1028-0979-2025-1-3.	Norkin, V.
Категорія видання: Б Назва праці: Lossy image compression with stochastic quantization	Kozyriev, A., & Norkin, V. (2024). Lossy image compression with stochastic quantization. <i>Cybernetics and Computer Technologies</i> , (3), 60–66. https://doi.org/10.34229/2707-451x.24.3.6	Norkin, V.
Категорія видання: Б Назва праці: On shor's R-algorithm for problems with constraints	Norkin, V., & Kozyriev, A. (2023). On shor's R-algorithm for problems with constraints. <i>Cybernetics and Computer Technologies</i> , (3), 16–22. https://doi.org/10.34229/2707-451x.23.3.2	Norkin, V.



Апробації

Категорія видання, Назва праці	Видавництво, журнал (назва, номер, рік), бібліографічне посилання	Прізвища співавторів
Конференція: "Numerical Computations: Theory and Algorithms (NUMTA 2023)" Назва праці: Constrained global optimization by smoothing	Norkin V., Pichler A., Kozyriev A. Constrained global optimization by smoothing. <i>Book of Abstracts of the 4th International Conference and Summer School "Numerical Computations: Theory and Algorithms (NUMTA 2023)"</i> (Calabria, Italy, June 14-20, 2023), edited by Y.D. Sergeyev, D.E. Kvasov, M. Chiara Nasso. P.59. https://www.numta.org/pdf/NUMTA2023_Book.pdf	Norkin V., Pichler A.
Конференція: "Education and Research in the Information Society (ERIS 2023)" Назва праці: Modern quasi-gradient stochastic algorithms	Norkin V., Kozyriev A. Modern quasi-gradient stochastic algorithms. <i>Program of 16th International Conference Education and Research in the Information Society</i> . October 12, 2023. Plovdiv, Bulgaria. https://eris.adis.org/wp-content/uploads/2023/10/ERIS-Program_2023.pdf	Norkin V.
Конференція: АВТОМАТИКА 2024 Назва праці: Стиснення зображення з втратами за допомогою стохастичного квантування	Козирев А.Ю., Норкін В.І. Стиснення зображення з втратами за допомогою стохастичного квантування. <i>Збірник тез доповідей XXVII Міжнародної конференції «АВТОМАТИКА 2024», присвяченої пам'яті академіка НАН України В.М. Кунцевича та академіка НАН України Ю.Г. Кривоноса</i> . 20–22 листопада 2024 р., м. Дніпро (Україна) . ISBN 978-966-2344-74-5 . С.124-125. http://automatika2024.dp.ua/	Норкін В.І.



Дякую за увагу!

