

ЗАСТОСУВАННЯ МЕТОДУ UMAP ДЛЯ АНАЛІЗУ СТРУКТУРИ БАГАТОВИМІРНИХ МЕДИЧНИХ ДАНИХ

Мірошніченко Н.С., Перова І.Г.

Харківський національний університет
радіоелектроніки, кафедра – системотехніки,
м. Харків, Україна.

E-mail: nelia.miroshnychenko@nure.ua,
rikywenok@gmail.com

Abstract

This paper explores the application of the UMAP (Uniform Manifold Approximation and Projection) method for analyzing the structure of multidimensional medical data. UMAP is a modern dimensionality reduction technique that effectively captures hidden relationships and patterns in large datasets while preserving the local topology of the feature space. Its use enables efficient visualization of high-dimensional medical indicators, providing deeper insights into the internal structure of the data and facilitating the identification of patient groups with similar characteristics. The obtained results demonstrate UMAP's ability to maintain local relationships between observations and enhance the interpretability of medical data, making it a valuable tool for exploratory data analysis, pattern discovery, and data preprocessing for subsequent machine learning tasks.

Сучасні медичні дослідження оперують великими обсягами багатовимірних даних, що ускладнює їхній аналіз і візуалізацію. Методи зниження розмірності, такі як UMAP (Uniform Manifold Approximation and Projection), дозволяють перетворювати дані у простір меншої розмірності без суттєвої втрати інформації. Це дає можливість виявляти приховані зв'язки між змінними, групи пацієнтів та потенційні патерни, що неочевидні у початковому просторі ознак.

Застосування UMAP у медичних дослідженнях сприяє виявленню локальних структур та схожих профілів пацієнтів, що є важливим для ранньої діагностики та персоналізованої медицини.

Метод UMAP ґрунтується на ідеї побудови двох графів: одного у високовимірному просторі, що моделює локальну топологію даних, та іншого у просторі низької розмірності, який апроксимує цю топологію у проєкції.

Першим кроком алгоритму є побудова вагового графа у вихідному просторі. Для кожної точки x_i обчислюється міра близькості до сусідів за допомогою ймовірнісного розподілу:

$$p_{j|i} = \exp\left(-\frac{d(x_i, x_j) - p_i}{\sigma_i}\right) \quad (1)$$

де $d(x_i, x_j)$ – відстань між точками, p_i – мінімальна відстань до найближчого сусіда, що гарантує ненульові ваги, а σ_i – параметр масштабу.

Глобальна ймовірність сусідства визначається як:

$$p_{ij} = p_{j|i} + p_{i|j} - p_{j|i} * p_{i|j} \quad (2)$$

Другим кроком алгоритму є побудова графа у просторі низької розмірності. Для проєкції аналогічно визначається ймовірність зв'язку між точками y_i, y_j :

$$q_{ij} = \frac{1}{1 + a \|y_i - y_j\|^{2b}} \quad (3)$$

де a та b – параметри, що контролюють форму функції відстані.

Третім кроком є оптимізація. Метою алгоритму є мінімізація розбіжності між графами високої та низької розмірності. Для цього використовується функція крос-ентропії:

$$C = \sum_{i \neq j} [p_{ij} \log \frac{p_{ij}}{q_{ij}} + (1 - p_{ij}) \log \frac{1-p_{ij}}{1-q_{ij}}] \quad (4)$$

Оптимізація здійснюється методом стохастичного градієнтного спуску, що дозволяє ефективно обчислювати низькорозмірне представлення навіть для великих вибірок [1].

Для експерименту було використано медичний набір Cardiovascular Disease Dataset [2], який містить 70 000 записів пацієнтів з 11 числовими ознаками та цільовою змінною — наявністю серцево-судинного захворювання. Ознаки включають вік, стать, артеріальний тиск, рівень холестерину, індекс маси тіла та інші медичні показники, що можуть впливати на ризик розвитку захворювання. Дані було стандартизовано для усунення впливу ознак з більшими значеннями та забезпечення коректного обчислення відстаней. Для оцінки розподілу класів у латентному представленні застосовано UMAP з параметрами `n_neighbors=15`, `min_dist=0.1`, `metric='euclidean'`, з проєкцією у тривимірний простір.

На рисунку 1 наведено результати 3D-візуалізації, де точки позначено відповідно до класів (0 – здорові, 1 – хворі).

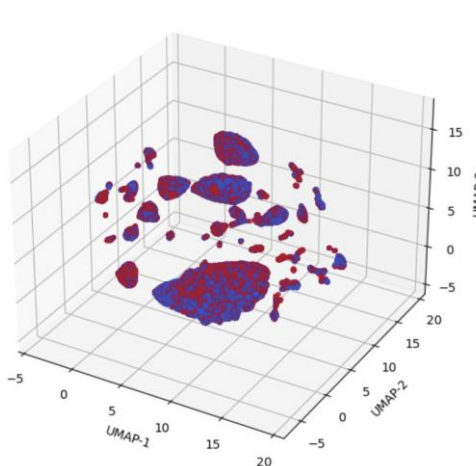


Рис.1 – Розподіл даних у просторі ознак після застосування UMAP

Побудована тривимірна візуалізація за допомогою UMAP демонструє здатність методу ефективно зберігати локальну структуру медичних даних, формуючи компактні «хмари» точок, що відображають схожі профілі пацієнтів за 11 ознаками. Це свідчить про те, що UMAP добре справляється з виявленням локальних груп та закономірностей у складних, багатовимірних вибірках, характерних для медичних даних.

Водночас спостерігається перекриття класів: здорові (Cardio=0, синій) та хворі (Cardio=1, червоний) пацієнти не утворюють чітко відокремлених груп. Це зумовлено багатofакторною природою серцево-судинних захворювань: окремі ознаки (наприклад, високий тиск чи індекс маси тіла) можуть підвищувати ризик, але для точного розділення хворих та здорових недостатньо однієї або двох характеристик.

Отже, UMAP є ефективним інструментом для візуалізації та виявлення локальних структур, однак не гарантує глобального розділення класів. Для підвищення точності прогнозування доцільно застосовувати нелінійні моделі (нейронні мережі, ансамблі) та проводити додаткову інженерію ознак (взаємодії, нормалізація, категоризація за віком/статтю) для підвищення точності прогнозів.

Література

1. Dorrity, M. W., Saunders, L. M., Queitsch, C., Fields, S., & Trapnell, C. (2020). Dimensionality reduction by UMAP to visualize physical and genetic interactions. *Nature Communications*, 11(1), 1537. <https://doi.org/10.1038/s41467-020-15351-4>.
2. Sulianova, A. (n.d.). Cardiovascular Disease Dataset. Kaggle. <https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset>.