

ДИНАМІЧНА ВІЗУАЛІЗАЦІЯ ДАНИХ ВЕЛИКОЇ РОЗМІРНОСТІ

Вінницький національний технічний університет

Анотація

В роботі розглянуто метод обробки даних великої розмірності на основі яких будується їх візуалізація у двох, або трьох вимірах. Метод базується на алгоритмах зменшення розмірності та може бути застосованим для розробки біотехнічної акустичної системи динамічної візуалізації серцевого м'яза людини.

Ключові слова: зменшення розмірності, обробка даних, динамічна візуалізація.

Abstract

The paper considers a method for processing high-dimensional data, on the basis of which their visualization in two or three dimensions is built. The method is based on dimensionality reduction algorithms and can be applied to the development of a biotechnical acoustic system for dynamic visualization of the human heart muscle.

Keywords: dimensionality reduction, data processing, dynamic visualization.

Вступ

Зменшення розмірності (ЗР) є стандартною практикою аналізу даних великої розмірності. На додаток до зменшення шуму та полегшення подальшого обчислювального аналізу, методи ЗР широко використовуються для візуалізації даних у двох або трьох вимірах. Фактично, багато методів ЗР були спеціально розроблені для генерації інформативних візуалізацій даних великої розмірності.

Деякі популярні методи ЗР для візуалізації даних включають лінійний метод головних компонент (МГК) і нелінійне t-розподілене вкладення стохастичної близькості (t-SNE, з англ. t-distributed Stochastic Neighbor Embedding) і алгоритм “Uniform Manifold Approximation and Projection” (UMAP). Багато інших методів ЗР були розроблені для усунення недоліків широко використовуваних підходів t-SNE та UMAP. Методи ЗР для візуалізації знайшли специфічне застосування в широкому діапазоні різних дисциплін. Деякі приклади включають перевірку ідентичності типу клітини в одноклітинній біології, дослідження вхідних даних із моделей глибокого навчання, дослідження закономірностей у геноміці людини та аналізі хімічного складу зоряних об'єктів [1].

Результати дослідження

Незважаючи на популярність методів ЗР для візуалізації даних великої розмірності, ці методи схильні до спотворень і неоднорідності якості низьковимірної візуалізації. Таким чином, просте використання методів ЗР для перевірки, підтвердження або інформування результатів досліджень і вказівок може бути сприйнятливим до неправильного тлумачення через ці спотворення. Наприклад, у галузі одноклітинної біології візуалізація t-SNE або UMAP часто використовується для підтвердження ідентичності типів клітин кластерів, інтеграції різних одноклітинних наборів даних та обчислення клітинних траєкторій за допомогою показників швидкості РНК. Для кожного з вищезазначених випадків використання наявні спотворення відстаней між спостереженнями та неоднорідності в якості візуалізації ЗР, що може вплинути на отримані інтерпретації. Як правило, через ці спотворення візуалізації ЗР можуть призвести до неправильної перевірки кластерів (тобто недостатньої кластеризації або надмірної кластеризації), штучного виявлення або відсутності виявлення сполучних зв'язків між кластерами та штучної присутності впорядкування або втрати порядку впорядкування спостережень уздовж осей метаданих. Занепокоєння щодо ЗР посилюються статичним характером поточних методів візуалізації ЗР, які зазвичай показують зображення, що

представляє одну ініціалізацію методу ЗР. Цей підхід може маскувати випадковість у візуалізації через дані та/або метод ЗР, і може бути вразливим до вибору.

Щоб усунути описані обмеження використовують структуру для створення динамічних візуалізацій даних шляхом узгодження кількох початкових візуалізацій ЗР. У результаті динамічна візуалізація надає значно більше інформації, ніж будь-яка окрема статична візуалізація. Ці динамічні візуалізації забезпечують інтуїтивну структуру для розуміння користувачами чутливості візуалізації ЗР до збурень у даних і до будь-яких стохастичних компонентів методу ЗР. Цей інструмент можна використовувати для діагностики різноманітних помилок інтерпретації, які охоплюють багато з вищезгаданих проблем. Щоб кількісно визначити ступінь флуктуації вибірки у візуалізаціях завантажувального ЗР, вводиться оцінка дисперсії, яка відстежує мінливість у репліках у реальному світі та може використовуватися для оптимізації конвеєра візуалізації ЗР. На відміну від попередніх показників якості для оцінки візуалізацій ЗР, які вимірюють узгодженість візуалізації з даними великої розмірності, оцінка дисперсії може виміряти вплив шуму вибірки при генеруванні даних на спотворення, що спостерігаються у візуалізації ЗР [2].

Вхідні дані для динамічної візуалізації такі ж, як і для вибраного статичного методу ЗР, за винятком одного додаткового параметра, що відповідає кількості початкових зразків, які використовуються для візуалізації, що залежить від передбачуваного варіанту використання та обчислювальних можливостей доступних користувачеві. Динамічні візуалізації можуть забезпечити надійне підтвердження попередніх висновків і створення нових дослідницьких гіпотез.

Розглянемо вхідні дані для динамічної візуалізації як $X \in R^{n \times p}$, де n – кількість спостережень, а p – кількість ознак. Наприклад, у випадку одноклітинних даних транскриптомії вхідними даними є нормалізована матриця підрахунку транскриптів, де спостереження відповідають окремим клітинам, а ознаки відповідають зазвичай великій кількості генів [3]. Далі генерується вихідний зразок X шляхом вибірки n рядків X із заміною. Це повторюється B разів, щоб накопичити набір B початкових зразків X , які позначаємо як $\{X^{(1)}, \dots, X^{(B)}\}$. Для даного ЗР методу генерується двовимірна візуалізація X , яку назвемо $Y = f(X) \in R^{n \times 2}$. Використовуючи той самий метод ЗР, візуалізації також генеруються для зразків початкового завантаження, що дає набір $\{Y^{(1)}, \dots, Y^{(B)}\}$ у випадковому порядку. Кожна завантажувальна візуалізація $Y^{(k)}$ з $k \in \{1, \dots, B\}$ вирівнюється до еталонної візуалізації Y за допомогою алгоритму Kabsch для мінімізації середньої квадратичної відстані між узгодженими спостереженнями $Y^{(k)}$ і Y під час операцій тривимірного обертання, які зберігають відношення відстані у візуалізації, отже, не змінюють значення цільової функції, мінімізоване більшістю методів ЗР.

Послідовність випадково впорядкованих початкових візуалізацій $\{Y^{(1)}, \dots, Y^{(B)}\}$ може бути представлена використовуючи три різні модальності динамічної візуалізації, кожна з яких висвітлює деякі аспекти даних і їх візуалізації ЗР. Інтерактивна візуалізація з використанням HTML створює інтерпольовані переходи між випадковою послідовністю початкових візуалізацій і надає користувачам можливість високорівневої взаємодії, такої як увімкнення/вимкнення спостережень за допомогою окремих міток і читання метаданих, вбудованих у візуалізацію. Анімована візуалізація представляє випадкову послідовність початкових візуалізацій у форматах відео та GIF, які можна легко вставити в дослідницькі робочі процеси та презентації. Складена візуалізація накладає вирівняні початкові візуалізації в одне зображення з низькою прозорістю точок розсіювання, щоб візуально представити щільність візуалізацій, що добре підходить для представлення динаміки в одному кадрі.

Метод динамічної візуалізації забезпечує наступні функціональні можливості:

1. Діагностує стабільність сполучних зв'язків між кластерами. З'єднувальні зв'язки часто виникають у візуалізації кластеризованих даних, коли два або більше окремих кластерів спостережень можуть виглядати контактними або мати вузький зв'язок між ними в низьковимірній візуалізації. Ці сполучні зв'язки можуть представляти базові зв'язки між двома кластерами, особливо в одноклітинній біології, де мостові з'єднання можуть фіксувати диференціацію клітин або шляхи розвитку, але вони також можуть бути артефактами, створеними спотвореннями в методі ЗР та/або внутрішнім шумом у даних. Справедливість мостового з'єднання важко виокремити зі статичної візуалізації. Динамічні візуалізації можуть допомогти відрізнити надійні мостові з'єднання, які з'являються в більшості завантажувальних візуалізацій, від випадкових або штучних мостових з'єднань, які з'являються лише в одній або невеликій кількості завантажувальних візуалізацій [4].

2. Оцінює впевненість у гіпотезах кластерів. Візуалізації ЗР кластеризованих даних часто використовуються або для підтвердження міток, отриманих за допомогою алгоритму кластеризації, або для визначення підмножин спостережень, які можуть представляти нові кластери. Динамічні візуалізації можна використовувати в обох випадках для вивчення стабільності ідентифікованих кластерів у різних візуалізаціях початкового завантаження.

3. Розкриває поділ і впорядкування даних за мітками. Як у кластеризованих, так і в некластеризованих даних візуальні шаблони в мітках спостереження корисні для кращого розуміння зв'язків між мітками та для створення нових дослідницьких гіпотез. Динамічні візуалізації можна використовувати для оцінки тонких шаблонів візуалізацій ЗР, таких як ступінь поділу різних міток у некластеризованому середовищі та впорядкування міток уздовж безперервної траєкторії.

Динамічна візуалізація та розрахунок оцінки дисперсії корисні для підвищення впевненості в інтерпретаціях візуалізації даних ЗР у широкому діапазоні дисциплін, включаючи, одноклітинну біологію, просторову біологію, геноміку, астрономію та інші. Динамічні візуалізації можуть діагностувати такі проблеми, як стабільність мостових зв'язків між кластерами, стійкість виявлених кластерів і порядок або шаблони у візуалізаціях даних з мітками, які неможливо діагностувати за поточної парадигми статичної візуалізації ЗР. Оцінка дисперсії служить показником якості для візуалізацій ЗР, який доповнює існуючі міри узгодження. Оцінка дисперсії має кілька бажаних властивостей, включаючи позитивну кореляцію з мінливістю між повторами даних і збіжність до оцінки дисперсії за реальної вибірки розподілу даних для достатньо великих n або B та за умов регулярності.

Висновки

Незважаючи на широкий вибір показників, доступних для оцінки якості візуалізацій ЗР, ці показники обмежено застосовуються в дослідницьких процесах і публікаціях. Така недостатня кількість використання та поінформованості спричинена головним чином двома причинами. По-перше, поточні оцінки якості важко інтерпретувати, особливо для користувачів з нетехнічною освітою. По-друге, хоча поточні показники якості вимірюють помилки, внесені методом ЗР, вони не враховують мінливість і шум у вихідних даних великої розмірності, які можуть однаково відповідати за візуалізації, які спотворюються від істини. Таким чином, оцінка дисперсії є першою метрикою ЗР, яка враховує як шум у вихідних даних, так і вплив збурень на спотворення візуалізації ЗР. Оцінка дисперсії доповнює існуючі показники якості, які зосереджені насамперед на вимірюванні зміщення, а не дисперсії процесу генерації даних і візуалізації.

Розробка оптимальних рішень для обробки повторюваних спостережень у візуалізаціях ЗР є перспективним напрямком для майбутніх досліджень. Подібним чином і процедура динамічної візуалізації, і розрахунок балів дисперсії вимагають генерації великої кількості візуалізацій ЗР, яка лінійно масштабується з кількістю початкових зразків і може вимагати багато обчислювальних потужностей для візуалізації даних із багатьма спостереженнями. Також, слід зазначити, що теоретичні гарантії початкової візуалізації дійсні лише за певних умов регулярності і їх ще потрібно підтвердити для окремих алгоритмів ЗР. Усунення цих обмежень у процедурі динамічної візуалізації зробить як динамічні візуалізації так і оцінки дисперсії можливими до застосування у більш широкому діапазоні алгоритмів ЗР, модальностей даних і дослідницьких питань.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Wang, Y., Huang, H., Rudin, C. & Shaposhnik, Y. Understanding how dimension reduction tools work: an empirical approach to deciphering t-SNE, UMAP, TriMap, and PaCMAP for data visualization. *Journal of Machine Learning Research* 22 (201), 1–73 (2021). <https://doi.org/10.48550/arXiv.2012.04456>
2. L.J.P. van der Maaten, G.E. Hinton. Visualizing high-dimensional data using t-SNE. *Journal of machine learning research* 9 (nov), 2579–2605 (2008).

3. Su, Y., Shi, Q. & Wei, W. Single cell proteomics in biomedicine: High-dimensional data acquisition, visualization, and analysis. *PROTEOMICS* 17 (3-4), 1600267 (2017). <https://doi.org/10.1002/pmic.201600267>
4. Kobak, D. & Berens, P. The art of using t-SNE for single-cell transcriptomics. *Nature Communications* 10 (1), 5416 (2019). <https://doi.org/10.1038/s41467-019-13056-x>

Дячук Олексій Олександрович — аспірант кафедри БМІОЕС, факультет інформаційних електронних систем, Вінницький національний технічний університет, Вінниця.

Тимчик Сергій Васильович — канд. техн. наук, доцент кафедри біомедичної інженерії та оптико електронних систем, Вінницький національний технічний університет, e-mail: tymchyksv@ukr.net

Diachuk Oleksii O. — graduate student of the Department of BMIOES, Faculty of Information Electronic Systems, Vinnytsia National Technical University, Vinnytsia, email : dayte2@gmail.com

Tymchyk Serhii V. — Cand. Sc. (Eng), Associate Professor of the Department of Biomedical Engineering and Optical-Electronic Systems, Vinnytsia National Technical University, Vinnytsia, e-mail: tymchyksv@ukr.net