УДК 004.85
DOI https://doi.org/10.24144/2616-7700.2023.42(1).181-187

Н. Е. Кондрук

ДВНЗ «Ужгородський національний університет»,
доцент кафедри кібернетики і прикладної математики,
кандидат технічних наук
natalia.kondruk@uzhnu.edu.ua
ORCID: https://orcid.org/0000-0002-9277-5131

АНАЛІЗ ТЕХНІК ЗМЕНШЕННЯ РОЗМІРНОСТІ В МАШИННОМУ НАВЧАННІ

Багато сучасних наборів даних мають високу розмірність, яка може призводити до проблем з перевантаженням моделей, зменшенням ефективності обробки даних та збільшення часу навчання. Тому дослідження застосування технік зменшення розмірності даних є важливою задачею для покращення продуктивності та швидкості аналізу. В роботі проведено огляд та оцінка ефективності сучасних технік для зменшення розмірності високорозмірного ознакового простору даних з метою візуалізації та попередньої обробки даних. Для цього розроблено інформаційно-аналітичну систему на Python, що реалізує PSA, t-SNE, Isomap, UMAP. В якості тестового набору даних був обраний високорозмірний набір «DARWIN» з 451 ознакою. В результаті експерименту всі техніки в цілому показали подібні результати візуалізації даних. t-SNE виявився найефективнішим методом попередньої обробки даних для цього датасету, покращивши точність kNN на 21% і SVC на 4%. Отримані результати доводять, що застосування сучасних методів зменшення розмірності даних може сприяти побудові більш ефективних моделей та прогнозів. Майбутні дослідження передбачають оцінку синергії технік аналізу даних та машинного навчання для вирішення конкретних прикладних задач.

Ключові слова: редукція, зменшення розмірності, візуалізація даних, високорозмірні дані.

1. Вступ. Методи редукції є важливим інструментом для візуалізації даних в машинному навчанні та інтелектуальному аналізі даних. Вони дозволяють зменшити кількість ознак, зберігаючи при цьому значущі характеристики, що приводить до покращення зрозумілості та інтерпретованості даних. Візуалізація є першим етапом аналізу, мета якого полягає в тому, щоб зрозуміти дані, перш ніж перейти до більш цілеспрямованого моделювання та дозволяє: зрозуміти дані швидше та більш ефективно, ніж тільки аналіз числових значень; визначити залежності та взаємозв'язки між ознаками; виявити відхилення та аномалії; прийняти обґрунтовані рішення.

Зменшення розмірності — це перетворення високорозмірних даних в значуще представлення меншої розмірності. Ідеальною має бути зменшена репрезентація з розмірністю, яка відповідає внутрішній розмірності даних. Внутрішня розмірність даних — це мінімальна кількість параметрів, необхідних для пояснення спостережуваних властивостей даних [1]. Традиційно, зменшення розмірності здійснювалося за лінійними техніками, такими як аналіз головних компонент (PCA) [2], факторний аналіз та класичне масштабування [1]. Однак, вони виявились неефективними при обробці складних нелінійних даних. Протягом останнього десятиліття було запропоновано велику кількість нелінійних технік зменшення розмірності: т-розподілене вкладення стохастичної близькості (t-SNE) [3, 4], Isomap [5], UMAP [6] та інші.

Для зменшення розмірності у машинному навчанні використовуються різноманітні методи, які можна розділити на параметричні та непараметричні, лінійні та нелінійні моделі [1]. Кожен з цих підходів має свої переваги та обмеження, тому їх вибір залежить від конкретних потреб та характеристик даних.

Метою роботи є огляд сучасних технік зменшення розмірності та дослідження ефективності їх використання для візуалізації і попередньої обробки високорозмірних даних в задачах машинного навчання та аналізу даних.

2. Моделі і методи. Лінійні параметричні методи — це підхід до зменшення розмірності даних, який базується на їх перетворенні з високорозмірного простору в простір меншої розмірності, зберігаючи при цьому якомога більше інформації. При цьому, застосовується лінійне перетворення даних, яке описується матрицею — параметром моделі, яку можна знайти шляхом мінімізації певної функції втрат, такої як середньоквадратична похибка або сума квадратів похибок.

До лінійних параметричних методів належать: метод головних компонент (PCA), лінійний дискримінантний аналіз (LDA) та канонічний кореляційний аналіз (CCA).

PCA — це метод зменшення розмірності даних, який використовується для їх відображення у простір меншої розмірності. Він дозволяє перетворити вхідні дані в нові ортогональні змінні, які називаються головними компонентами (principal components), кожна з яких є лінійною комбінацією вхідних ознак. В процесі метод головних компонент формує головні компоненти в порядку спадання їх дисперсії. Перша головна компонента описує напрям, в якому дисперсія даних максимальна. Друга головна компонента обирається таким чином, щоб максимізувати залишкову дисперсію після вилучення першої, і так далі. Цей процес продовжується до тих пір, поки не будуть обрані всі головні компоненти або досягнута певна попередньо визначена їхня кількість [1]. Отримані головні компоненти можна використовувати для зменшення розмірності даних. Для цього вилучаються менш важливі компоненти, які не містять значимої інформації і залишаються більш суттєві.

Лінійні параметричні методи зазвичай працюють добре, якщо дані незашумлені та мають лінійну структуру. В інших випадках використовують нелінійні непараметричні методи. Такі методи називаються непараметричними, оскільки вони не припускають жодної апріорної інформації про розподіл даних [1].

Один з найбільш популярних непараметричних методів — це т-розподілене вкладення стохастичної близькості t-SNE (t-Distributed Stochastic Neighbor Embedding). В основі лежить ідея, що точки високовимірного простору, які близькі одна до одної, повинні відповідати таким же близьким точкам в низькорозмірному просторі. Цей метод використовує ймовірнісну модель для знаходження розподілу ймовірностей сусідства між точками високовимірного простору та низькорозмірного простору. За допомогою градієнтного спуску вирішується задача мінімізації відстаней між відповідними точками у двох просторах [3, 4].

Ще один альтернатиний непараметричний метод — це UMAP (Uniform Manifold Approximation and Projection). UMAP використовує геометричну конструкцію, яка називається "єдиним гладким наближенням до множини щоб знайти більш точне низькорозмірне відображення кожної точки та ієрархію кластерів. Він використовує ідею топологічного аналізу та ріманової геометрії для зна-

ходження нижньої межі відстаней між точками високорозмірного простору та проекції їх на низькорозмірну площину з мінімальною втратою інформації [6]. UMAP враховує локальну структуру даних, щоб створити карту, яка зберігає відносну відстань між точками та ієрархічну структуру даних. Метод застосовує випадкові перетворення, щоб зменшити перенавчання та покращити якість проекції. Він є потужним та досить швидким методом зменшення розмірності, який може бути застосований до великих обсягів даних.

Isomap (Isometric Feature Mapping) — є одним з методів нелінійного зниження розмірності даних. Він використовується для відображення взаємного розташування точок у високорозмірному просторі на низькорозмірний, зберігаючи при цьому геометричні властивості даних.

Основна ідея Isomap полягає в тому, щоб побудувати граф сусідства, де кожна точка представлена як вузол, а ребра відображають взаємну близкість між точками. Потім використовують алгоритм Флойда-Уоршелла геодезичного відстаневого перетворення, який обчислює найкоротші шляхи між усіма парами точок у графі. Геодезична відстань — це найкоротший шлях між двома точками на поверхні многовиду. За допомогою цих підходів обчислюється низькорозмірне представлення, де відстані між точками якомога ближчі до геодезичних відстаней у високорозмірному просторі [5].

Isomap є потужним інструментом для знаходження нелінійних структур у даних і використовується в різних областях, таких як комп'ютерний зір, обробка сигналів, біоінформатика та інші. Він допомагає знизити розмірність даних, зберігаючи при цьому їх внутрішню структуру і геометричні відношення між ними. У порівнянні з PCA, Isomap здатний знайти більш складні форми многовидів, зокрема з урахуванням нелінійних форм. Також Isomap може бути використаний для заповнення пропущених даних.

3. Експерименти.

1. Постановка задачі. Обраний набір даних для порівняльної характеристики методів зменшення розмірності є «DARWIN» [7–9], який містить дані про почерк 174 учасників, що описано 451 атрибутами. Завдання класифікації полягає в тому, щоб відрізнити хворих на хворобу Альцгеймера від здорових людей.

2. Візуалізація даних. Для розв'язання цього завдання було зменшено розмірність простору ознак до 2D методами PCA, T-SNE, UMAP, Isomap.

При використанні методу головних компонент (рис. 1а) пояснювальна дисперсія становила всього 16,8%, що вказує на втрату великої кількості інформації при такій редукції. Щоб пояснювальна дисперсія становила не менше 90% необхідно було б взяти 79 головних компонент, але це не дозволить візуалізувати дані.

На рис. 1б показано візуалізацію згенеровану t-SNE при значенні перплексії — 30. Даний параметр визначає баланс між врахуванням глобальної та локальної структури даних. t-SNE не враховує геометричну структуру даних високорозмірного простору.

На рис. 1в представлено візуалізацію методом Isomap, а на рис. 1г методом UMAP із кількістю сусідів — 10 (в обох методах) та мінімальною відстаню — 0,1. Маленькі значення параметра «n_neighbors» означають, що, намагаючись оцінити простір, в якому розподілені дані, алгоритм обмежується малим околом навколо кожної точки, тобто намагається вловити локальну структуру даних.

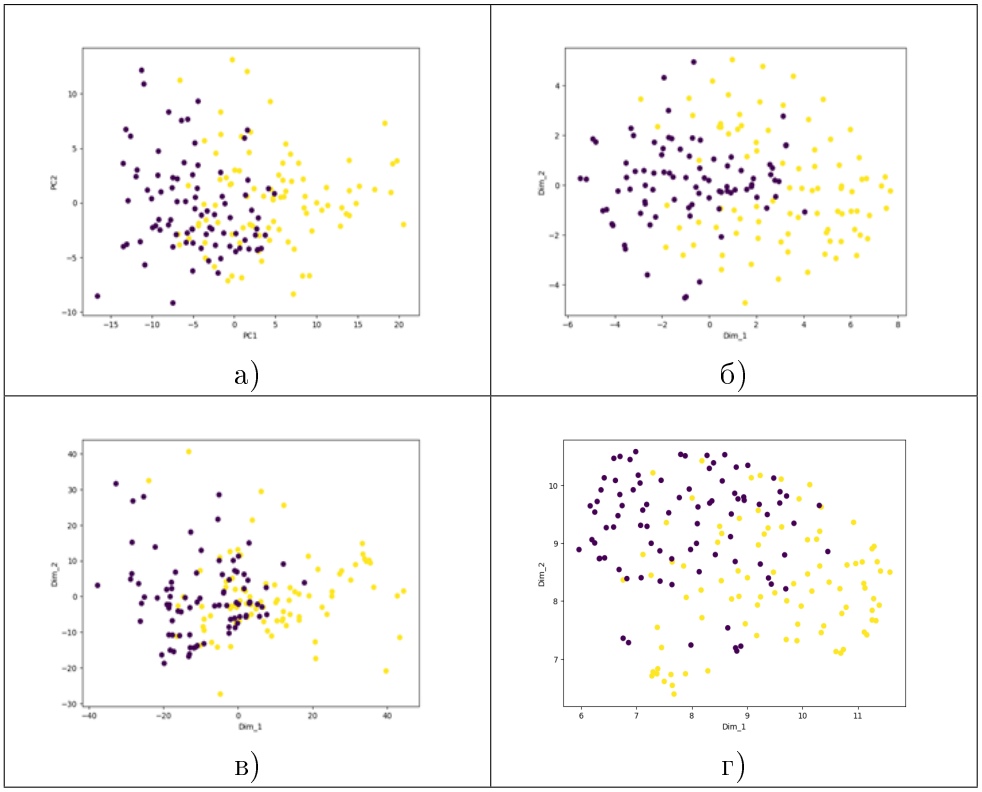

Рис. 1. 2D візуалізація набору даних «DARWIN» методами: а) PSA, б) t-SNE, в) Isomap, г) UMAP.

З іншого боку, великі значення змушують алгоритми враховувати точки у більшому околі, зберігаючи глобальну структуру даних, але упускаючи деталі. Слід зазначити, що метод UMAP є досить сильно варіативним.

Проаналізувавши отримані графічні результати можна стверджувати, що в цілому всі алгоритми відображають схожу кластерну структуру даних: кластери не мають чіткого відокремлення, тобто їх границі розмиті. Наявність великої кількості «граничних» об'єктів буде знижувати ефективність алгоритмів кластеризації та класифікації.

3. Дослідження застосування методів редукції, для попередньої обробки даних. З метою підвищення ефективності роботи методів навчання з учителем в даній частині буде досліджено чи підвищиться точність моделей застосованих до редукованого простору ознак. Дослідження проведено для методів найближчих сусідів (kNN) [10] та опорних векторів (SVM) [11]. Для порівняльного аналізу гіперпараметри моделей взято за замовчуванням. Також попередньо проведена стандартизація даних, тобто приведення всіх ознак до однієї шкали (Standard Scaler). В якості індексів оцінки ефективності моделей взято середню похибку роботи методу кросвалідації при розбитті даних на 5 сукупностей.

Проалізувавши дані табл. 1 очевидно, що найефективнішим методом попередньої обробки для методів найближчих сусідів та опорних векторів є редукція простору ознак t-SNE. Крім того, для методу kNN ефективність підвищувалась завжди із використанням будь-якого підходу редукції. Для методу SVC тільки одна техніка (t-SNE) покращила результат.

Таблиця 1.

Оцінка точності моделей шляхом перехресної перевірки.

Точність (стандартне відхилення) моделі	Без попередньої обробки методами редукції	Попередня 2D редукція ознакового простору за PCA	Попередня 2D редукція ознакового простору за t-SNE	Попередня 2D редукція ознакового простору за UMAP	Попередня 2D редукція ознакового простору за Isomap
kNN	0,64(0,12)	0,68(0,21)	0,85(0,14)	0,73(0,12)	0,73(0,11)
SVC	0,83(0,14)	0,69(0,19)	0,87(0,12)	0,75(0,2)	0,77(0,17)

4. Висновки та перспективи подальших досліджень. Дане дослідження є розвитком напрямку прикладного аналізу даних [12–16].

Досліджено ефективність використання різних технік (PSA, t-SNE, Isomap, UMAP) редукції високорозмірного ознакового простору даних, як для візуалізації даних так і для застосування цих технік до їх попередньої обробки. Зроблено порівняльний аналіз отриманих результатів. Очевидно, що необхідно застосувати ряд різних технік зменшення розмірності для визначення найефективнішої до кожної окремої прикладної задачі. Розроблена інформаційно-аналітична система на мові програмування Python та бібліотек scikit-learn, umap-learn, що реалізує описаний підхід. В якості апробаційної моделі обрано високорозмірний (451 ознака) набір даних «DARWIN». В ході експериментального дослідження для його візуалізації всі техніки в загальному показали однаковий результат. Найефективнішим методом попередньої обробки даних виявився t-SNE, що покращив точність kNN на 21%, а SVC на 4%.

Отже, застосування сучасних технік редукції може значно полегшити аналіз та розуміння даних у машинному навчанні, а також допомогти у побудові більш ефективних моделей та прогнозів.

Перспективні дослідження полягають у дослідженні ефективності застосування та поєднання різних технік аналізу даних та машинного навчання до розв'язання прикладних задач.

Список використаної літератури

1. Gisbrecht A., Hammer B. Data visualization by nonlinear dimensionality reduction. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery.* 2015. No. 5. P. 51–73. DOI: https://doi.org/10.1002/widm.1147
2. Bro R., Smilde A. Principal component analysis. *Analytical methods.* 2014. Vol. 6, No. 9. P. 2812–2831. DOI: https://doi.org/10.1039/c3ay41907j
3. Maaten L., Hinton G. Visualizing Data using t-SNE. *Journal of Machine Learning Research.* 2008. No. 9. P. 2579–2605.
4. Hinton G., Roweis S. Stochastic Neighbor Embedding. *Neural Information Processing Systems.* 2002. No. 15. P. 1–8.
5. Tenenbaum J., Silva V., Langford J. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science.* 2000. Vol. 290. P. 2319–2323. DOI: https://doi.org/10.1126/science.290.5500.2319
6. McInnes L., Healy J., Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426 [stat.ML].* 2020. P. 1–63. DOI: https://doi.org/10.48550/arXiv.1802.03426

7. Fontanella F. DARWIN. *UCI Machine Learning Repository.* URL: https://archive-beta.ics.uci.edu/dataset/732/darwin
8. Cili N. D. An experimental protocol to support cognitive impairment diagnosis by using handwriting analysis. *Procedia Computer Science.* 2018. No. 141. P. 466–471. DOI: https://doi.org/10.1016/j.procs.2018.10.141
9. Cili N. D. Diagnosing Alzheimer's disease from online handwriting. *Engineering Applications of Artificial Intelligence.* 2022. Vol. 111. DOI: https://doi.org/10.1016/j.engappai.2022.104822
10. Cover T., Hart P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory.* 1967. No. 13. P. 21–27.
11. Cortes C., Vapnik V. Support-vector networks. *Machine Learning.* 1995. No. 20. P. 273–297.
12. Кондрук Н. Е. Використання мір подібності в методах класифікаціїі. *Науковий вісник Ужгородського університету. Серія: «Математика і інформатика».* 2021. Вип. 1, № 38. С. 143–148. DOI: https://doi.org/10.24144/2616-7700.2021.38(1).143-148
13. Kondruk N. E. Use of length-based similarity measure in clustering problems. *Radio Electronics. Computer Science. Control.* 2018. Vol. 3, No. 48. P. 98–105. DOI: https://doi.org/10.15588/1607-3274-2018-3-11
14. Kondruk N. E., Malyar M. M. Analysis of Cluster Structures by Different Similarity Measures. *Cybernetics and Systems Analysis.* 2021. Vol. 57. P. 436–441. DOI: https://doi.org/10.1007/s10559-021-00368-4
15. Kondruk N., Malyar M. Dimensionality Reduction of the Criterion Space in Some Optimization Problems. International Conference "Computational Intelligence", 28–30 September 2021. Kyiv-Uzhhorod, 2021. P. 112–121. URL: https://ceur-ws.org/Vol-3018/Paper_11.pdf
16. Кондрук Н. Е. Моделі багатофакторного прогнозування. *Науковий вісник Ужгородського університету. Серія : «Математика і інформатика».* 2022. Т. 40, № 1. С. 168–174. DOI: https://doi.org/10.24144/2616-7700.2022.40(1).168-174

Kondruk N. E. Analysis of Dimensionality Reduction Techniques in Machine Learning.

Many modern datasets have high dimensionality, which can lead to issues such as model overload, decreased data processing efficiency, and increased training time. Therefore, researching the application of data dimensionality reduction techniques is an important task for improving productivity and analysis speed. This work provides an overview and evaluation of the effectiveness of contemporary techniques for reducing the dimensionality of high-dimensional feature spaces in data, aiming at data visualization and preprocessing. To accomplish this, an information analytics system was developed in Python, that implements PCA, t-SNE, Isomap, and UMAP. The "DARWIN" dataset with 451 features was selected as the test dataset. The experimental results showed similar data visualization outcomes for all techniques overall. t-SNE proved to be the most effective data preprocessing method for this dataset, improving the accuracy of kNN by 21% and SVC by 4%. The obtained results demonstrate that modern data dimensionality reduction methods can contribute to constructing more effective models and predictions. Future research will involve evaluating the synergy between data analysis techniques and machine learning to address specific applied problems.

Keywords: reduction, dimensionality reduction, data visualization, high-dimensional data.

References

1. Gisbrecht, A., & Hammer, B. (2015). Data visualization by nonlinear dimensionality reduction. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* 5, 51–73. https://doi.org/10.1002/widm.1147
2. Bro, R., & Smilde, A. (2014). Principal component analysis. *Analytical methods,* 6(9), 2812–2831. https://doi.org/10.1039/c3ay41907j
3. Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE, *Journal of Machine Learning Research,* 9, 2579–2605.
4. Hinton, G., & Roweis, S. (2002). Stochastic Neighbor Embedding. *Neural Information Processing Systems,* 15, 1–8.

5. Tenenbaumm, J., Silva, V., & Langford, J. (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science,* 290, 2319–2323. https://doi.org/10.1126/science.290.5500.2319
6. McInnes, L., Healy, J., & Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426 [stat.ML],* 1–63. https://doi.org/10.48550/arXiv.1802.03426
7. Fontanella, F. DARWIN. *UCI Machine Learning Repository.* Retrieved from https://archive-beta.ics.uci.edu/dataset/732/darwin
8. Cilia, N. D. (2018). An experimental protocol to support cognitive impairment diagnosis by using handwriting analysis. *Procedia Computer Science,* 141, 466–471. https://doi.org/10.1016/j.procs.2018.10.141
9. Cilia, N. D. (2022). Diagnosing Alzheimer's disease from online handwriting. *Engineering Applications of Artificial Intelligence,* 111, 104822. https://doi.org/10.1016/j.engappai.2022.104822
10. Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory,* 13, 21–27.
11. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning,* 20, 273–297.
12. Kondruk, N. E. (2021). Use of similarity measures in classification methods. *Scientific Bulletin of Uzhhorod University. Series of Mathematics and Informatics,* 1(38), 85–91. https://doi.org/10.24144/2616-7700.2021.38(1).143-148
13. Kondruk, N. E. (2018). Use of length-based similarity measure in clustering problems. *Radio Electronics. Computer Science. Control,* 3(46), 98–105. https://doi.org/10.15588/1607-3274-2018-3-11
14. Kondruk, N. E., & Malyar, M. M. (2021). Analysis of Cluster Structures by Different Similarity Measures. *Cybern. Syst. Anal.,* 57, 436–441. https://doi.org/10.1007/s10559-021-00368-4
15. Kondruk, N., & Malyar, M. (2021). Dimensionality Reduction of the Criterion Space in Some Optimization Problems, Kyiv-Uzhhorod. Retrieved from https://ceur-ws.org/Vol-3018/Paper_11.pdf
16. Kondruk, N. E. (2022). Models of multivariate forecasting. *Scientific Bulletin of Uzhhorod University. Series of Mathematics and Informatics,* 40(1), 168–174. https://doi.org/10.24144/2616-7700.2022.40(1).168-174

Одержано 02.05.2023