УДК 004.8
DOI: 10.18523/2617-3808.2022.5.49-53

Кузнєцов В. О., Крак Ю. В., Ляшко В. І., Касянюк В. С.

АНАЛІЗ І СИНТЕЗ ТЕХНОЛОГІЙ КЛАСИФІКАЦІЇ ТЕКСТОВОЇ ІНФОРМАЦІЇ

Розглянуто задачу якісного аналізу процесу перетворення текстової інформації на набір ознак і відповідно перетворення цих ознак на набір, зручний для візуального аналізу. Розглянуто структуру типової технології з аналізу текстової інформації та визначено її основні елементи. Детально наведено опис кожного елементу технології аналізу та класифікації текстової інформації із залученням методів класифікації та групування ознак. Проведено експериментальні випробування окремих компонентів цієї технології.

Ключові слова: оброблення текстів, видобуток ознак, візуальна аналітика, алгоритми класифікації, зменшення розмірності ознак.

Вступ

Розуміння тексту комп'ютером – нетривіальна задача, що потребує залучення багатьох методів і алгоритмів оброблення текстової інформації. Типова технологія має пропонувати засоби підготовки оформлення даних відповідно до певного стандарту, методи представлення тексту у вигляді вектора характеристичних ознак і методи інтелектуального аналізу, візуальної аналітики та прийняття рішень [5].

При дослідженні текстової інформації постає проблема аналізу різнорідних текстів, поданих обмеженою множиною класів (тем текстів). Для їх розділення можливо застосувати різні підходи, зокрема статистичну спорідненість тем, коефіцієнти представлення текстів у прихованому просторі ознак, аномальні елементи даних тощо [10]. Для вирішення таких задач у вільному доступі викладають вибірки текстів, у яких текстову інформацію залежно від поставленого завдання може бути подано в різній формі: у вигляді окремих текстових файлів, гіпертексту (текстів із форматуванням тегами, наприклад HTML, XML), у вигляді баз даних або окремих об'єктів чи кортежей (наприклад, об'єкти типу JSON, дампи пам'яті програм тощо).

У цій роботі запропоновано розробити інформаційну технологію аналізу текстової інформації та провести експериментальні випробування:

- методів пошуку інформативних ознак;
- методів аугментації даних;
- методів перетворення розмірності вектора характерних ознак;
- гіпотез щодо класифікації текстових даних.

У цьому дослідженні становить інтерес формалізація всіх кроків технології з метою вивчити, які кроки технології є необхідними, та відповідно визначити її найважливіші елементи.

Запропонована технологія аналізу текстової інформації

Структура технології

Типова схема аналізу текстів може відрізнятися залежно від підходів, що використовують машинне навчання загального призначення або глибоке навчання [2]. Суть першого – це каскад алгоритмів, на вхід якого надходять дані (як правило, ознаки), а на вихід – гіпотези про належність елемента даних до певного класу. Глибоке навчання передбачає побудову архітектури нейронної мережі, яка поєднує різні базові елементи – звичайні шари, шари згортки, розрідження, шари рекурсії тощо. Відповідно, така мережа може містити каскад алгоритмів, навіть якщо структурні рівні відповідають за різні функції – виявлення ознак, групування та класифікацію [1].

Розглянемо докладніше перший із підвидів навчання, які добре зарекомендували себе, на прикладі інших задач (аналізу жестової інформації). Модель машинного навчання складається з таких методів:

- перетворення розмірності даних – сингулярний розклад [4];
- групування ознак – T-стохастичне вбудовування сусідів [3];
- ущільнення простору ознак – знешумлювальні автокодери [8];

- методи класифікації – дерева рішень, опорні вектори та ін. [6].

Як видно зі структури каскадної моделі навчання, можна поетапно проводити операції з підготовки даних, видобутку ознак, навчання та візуального аналізу ознак. Найбільш ресурсомістким є метод T-стохастичного вбудовування сусідів; другий за часом виконання – метод ущільнення простору ознак на основі глибокого розрідженого автокодувальника. На противагу методам групування ознак і ущільнення простору ознак, методи зменшення розмірності є дуже швидкими і можуть бути швидшими за аналогічні методи глибокого навчання. Найуспішніших результатів щодо часу виконання та ефективності до появи великих даних було досягнуто шляхом об'єднання цих двох груп, де методи класифікації складалися з опорних векторних машин і дерев рішень.

Навпаки, під час вивчення даних у «глибинному» підході рівні містять приховану інформацію, яку не завжди можна використовувати для прийняття рішень як щодо набору ознак, так і щодо наборів даних. Рішення приймають у кінцевому підсумку за якісними показниками – за збіжністю алгоритму і великою кількості прогонів процедур навчання, щоб визначити оптимальну архітектуру. За винятком пошарового навчання, яке передбачає залучення глибоких автокодерів для ініціалізації мережевих ваг (навчання без учителя), вибір ознак і самих елементів даних визначається вдалим відбором даних (балансування класів у наборі даних), використанням великих вибірок, сформованих афінними перетвореннями та доповненням даних.

Інформаційна зосередженість у тексті

Набори даних мають приховані властивості, які можна простежити лише за умови використання певних підходів. Точки в оригінальному розмірі можуть бути дуже віддаленими й мати велику розмірність у просторі ознак. В результаті зменшення розмірності даних приховані залежності стають більш очевидними, проте їх скупченість залежить від розміру вибірки зразків.

Використовуючи метод T-стохастичного вбудовування, можна побачити, що точки даних, порівняно з оригінальним простором ознак, починають зміщуватися одна до одної і формують певні скупченості, які можна трактувати не інакше, як області інформаційної зосередженості.

Цей метод, по суті, відшукує такі залежності між ознаками, що є в новому, трансформованому просторі ознак, із порівняно великою смугою розділення, однак недоліком такого представлення є час оброблення, який потенційно потребує прискорювачів машинного навчання для побудови нелінійного представлення даних.

Незважаючи на недоліки, цей підхід дуже корисний для візуального аналізу даних і побудови методів класифікації даних. Оскільки всі ознаки представлені в просторі низької розмірності, це допомагає висвітлити одночасно і корисну інформацію, і неінформативні елементи даних (аномалії). Знаючи їх розташування, можна застосувати фільтрацію даних за наявності доволі великої кількості представників даних, а також – визначити основні осі даних шляхом побудови регресії. Отже, можна визначити області з більшою щільністю даних і, відповідно, області з меншою кількістю аномалій. На рис. 1 показано, яким чином у представленні методу T-стохастичного групування відбувається побудова регресії методом консенсусу випадкових зразків (RANSAC) [9] і пошук інформативних елементів даних.

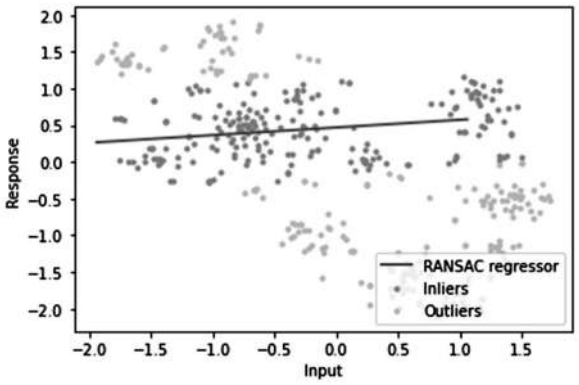

Рис. 1. Головна вісь і аномалії даних у зменшеному просторі ознак

Аугментація даних та її вплив на ефективність розпізнавання

Важливим етапом підготовки даних є створення набору даних, у якому представники класів даних збалансовані як за кількістю елементів даних, так і за їх взаємним розташуванням у випадку, якщо кластери даних розподілені нерівномірно (табл. 1).

Таблиця 1

Показники класифікації оригінального набору даних

Точність	Відгук	f1-міра	Підтримка	
Клас 1	0.59	1.00	0.74	366
Клас 2	0.98	0.95	0.96	454
Клас 3	0.90	0.64	0.75	332
Клас 4	1.00	0.59	0.74	333
Загалом			0.81	1485
Макросер.	0.87	0.79	0.80	1485
Зважене сер.	0.87	0.81	0.81	1485

Для цього можна штучно генерувати вибірки даних, які мають однакове розташування в просторі ознак, тож форма кластерів і розподіл вибірок даних залишатимуться незмінними. Цього можна досягти шляхом застосування певних методів збалансованого навчання, які генерують вибірки на підставі оригінального набору даних [7].

Досліджуючи в такий спосіб наукові тексти з метою вивчення впливу розміру вибірки на якість алгоритму, було виявлено, що використання методів доповнення для генерації нових вибірок, а також балансування вибірок загалом позитивно впливає на якість оброблення, зберігаючи конфігурацію простору ознак, розподіл даних у просторі (табл. 2).

Таблиця 2

Показники класифікації зміненого набору даних

Точність	Відгук	f1-міра	Підтримка	Точність
Клас 1	0.81	0.97	0.88	1313
Клас 2	0.92	0.96	0.94	1239
Клас 3	0.95	0.76	0.85	1274
Клас 4	0.98	0.93	0.96	1254
Загалом			0.91	5080
Макросер.	0.92	0.91	0.91	5080
Зважене сер.	0.91	0.91	0.91	5080

Порівняння наведених вище таблиць 1 і 2 може показати різницю між кількістю зразків у кожному класі, а також їх пропорцію один щодо одного. Зміна у відсотках показників для кожного класу даних, а також збільшення розміру набору даних збільшує відклик у найгіршому випадку з 59 % для класу 1 до 81 % і з 87 % до 92 %, що є добрим результатом для використання цієї методики підготовки даних.

Основним результатом застосування такої методики є можливість визначення мінімального розміру набору даних, необхідного для вирішення конкретної проблеми для конкретного типу даних. Це може допомогти виявити можливі ситуації, у яких більш доцільно застосувати методи глибокого навчання. Крім того, використовуючи ту саму стратегію, можна зробити висновок: якщо набір даних можна відокремити з бажаною точністю в просторі ознак зменшеної розмірності, можна спроєктувати мережу глибокого навчання, яка замість навчання на ознаках (подання нижчої розмірності) використовуватиме необроблені дані, які, з іншого боку, також можуть бути доповнені у вихідному просторі.

Визначення практичної швидкодії етапів оброблення тексту

Було виконано різні завдання з оброблення текстів із метою знайти найменш повільне місце в архітектурі навчання (табл. 3).

Таблиця 3

Показники ефективності виконання різних задач оброблення текстів

Назва	Макс, с	Мін,с	Макс/мін	Середнє
Зчитування	1.450	0.139	10.37339	0.6081
Транспонування	0.006	0.001	6.006006	0.0029
Сингулярний розклад	2.173	0.075	28.92936	0.6348
Групування ознак	65.57	10.81	6.06041	39.0700
Кластеризація	0.600	0.072	8.241758	0.2387
Підготовка зразків	0.007	0.002	2.681992	0.0051
Ініціалізація мережі	0.269	0.05	4.829627	0.1585
Навчання мережі	117.643	24.10	4.879832	71.0883
Тестування мережі	2.221	0.583	3.809802	1.2514
Навчання дерева рішень	0.826	0.148	5.587162	0.3859

Згідно з експериментами, наведеними у табл. 3, можна визначити декілька практичних результатів, що мають значення в обробленні тексту:

- завдання, які передбачають прості маніпуляції з пам'яттю (наприклад, зміна форми масиву), демонструють невелику різницю в продуктивності, як у абсолютних, так і у відносних значеннях, і в основному залежать від кількості обчислювальних ядер;
- відносна різниця (28x) у такому завданні, як сингулярний розклад, більш помітна, ніж у групуванні даних або вивченні глибокої мережі. Це означає, що матричні операції, зокрема множення, виконуються краще на певних (новіших) версіях процесорів, хоча це залежить від реалізації;
- практична ефективність навчання залежить від швидкості всіх компонентів, включених у завдання, починаючи від читання з диска до завдань машинного навчання та висновків на основі нових даних, якщо є слабкою ланкою. Наприклад, повільна робота групування ознак або глибокого навчання, які є трудомісткими завданнями, можуть вплинути на загальний час, тому потрібно використовувати правильну архітектуру в контексті кожного експерименту й типу досліджуваних даних.

Визначення ефективності методів класифікації тексту

Числові показники не завжди дають змогу оцінити коректність функціонування певних алгоритмів. У випадку вузької смуги роздільності певні алгоритми працюють гірше, ніж інші. Це може дати уявлення про набір даних і, зокрема, вказати на перенавчання методів зменшення розмірності даних. Наприклад, зменшення шумів у даних і зміщення може бути як позитивним

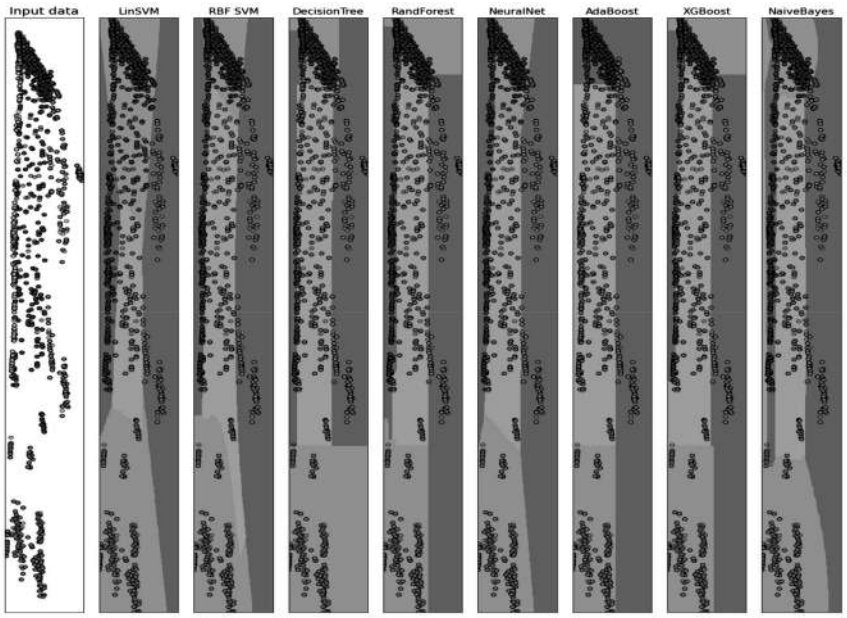

Рис. 2. Полоса розділення за наявності даних, що погано розділяються

моментом, так і негативним, оскільки впливає на можливість як узагальнення, так і побудови гіпотези за умов вузької полоси розділення, що треба завжди враховувати при побудові моделей машинного навчання. На рис. 2 показано, як полоса розділення впливає на ефективність алгоритмів класифікації.

Висновки

У результаті проведеного дослідження наведено типову технологію аналізу текстової інформації для задач класифікації. Проведено реалізацію та експериментальні випробування технології аналізу текстової інформації. Було розглянуто деякі аспекти застосування методів аугментації даних, пошуку скупченостей зразків даних та аномальних викидів, проведено експериментальні випробування алгоритмів класифікації текстової інформації, що дало змогу досягти точності розпізнавання на рівні 92 %. Окремо розглянуто вплив типу апаратного забезпечення на швидкодію алгоритму класифікації.

У подальших дослідженнях запропоновано випробувати інші архітектури й підходи до аналізу текстової інформації, зокрема ймовірнісні моделі – марківські ланцюги, рекурентні моделі на основі нейронних мереж із рекурентними зв'язками між шарами – рекурентні обмежені моделі, довго-короткочасову пам'ять та інші.

Список літератури

1. Aloysius N. A review on deep convolutional neural networks / N. Aloysius, M. Geetha // International Conference on Communication and Signal Processing (ICCSP). – 2017. – Pp. 0588–0592. https://doi.org/ 10.1109/ICCSP.2017.8286426
2. Alzubaidi L. Review of deep learning: concepts, CNN ar-chitectures, challenges, applications, future directions / L. Alzubaidi, J. Zhang, A. J. Humaidi, et al. // Journal of Big Data. – 2021. – No. 8 (53). – Pp. 1–74. https://doi.org/10.1186/s40537-021-00444-8
3. Chan D. M. T-SNE-CUDA: GPU-Accelerated T-SNE and its Applications to Modern Data / D. M. Chan, R. Rao, F. Huang, J. F. Canny // 30th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD). – 2018. – Pp. 330–338. https://doi.org/10.1109/CAHPC.2018.8645912. arXiv:1807.11824v1 [cs.LG]
4. Kirichenko N. Synthesis of systems of neurofunctional transducers in solving classification problems / N. Kirichenko, Yu. Krivonos, N. Lepekha // Cybernetics and systems analysis. – 2007. – Vol. 3. – Pp. 47–57.
5. Krak I. Using visual analytics to develop human and machine-centric models: A review of approaches and proposed information technology /I. Krak, O. Barmak, E. Manziuk // Computational Intelligence. – 2020. – Pp. 1–26. https://doi.org/10.1111/coin.12289
6. Krak I. Data Classification Based on the Features Reduction and Piecewise Linear Separation / I. Krak, O. Barmak, E. Manziuk, and A. Kulias // Advances in Intelligent Systems and Computing. – 2020. – Vol. 1072. – Pp. 282–289. https://doi.org/10.1007/978-3-030-33585-4_28
7. Menardi G. Training and assessing classification rules with imbalanced data / G. Menardi, N. Torelli // Data Mining and Knowledge Discovery. – 2014. – No. 28. – Pp. 92–122.
8. Vahdat A. NVAE: A Deep Hierarchical Variational Autoencoder / A. Vahdat, J. Kautz // 34th Conference on Neural Information Processing Systems (NeurIPS 2020). – 2020. https://doi.org/10.48550/arXiv.2007.03898
9. Zhang G. More Informed Random Sample Consensus / G. Zhang, Y. Chen. – 2020. https://doi.org/10.48550/arXiv.2011.09116
10. Zhou N. TextRank Keyword Extraction Algorithm Using Word Vector Clustering Based on Rough Data-Deduction / N. Zhou, W. Shi, R. Liang, N. Zhong // Computational Intelligence and Neuroscience. – 2022. https://doi.org/10.1155/2022/5649994

References

Aloysius, N., & Geetha, M. (2017). A review on deep convolutional neural networks. In *International Conference on Communication and Signal Processing (ICCSP)*, 0588–0592. https://doi.org/10.1109/ICCSP.2017.8286426

Alzubaidi, L., Zhang, J., Humaidi, A. J., et al. (2021). Review of deep learning: concepts, CNN ar-chitectures, challenges, applications, future directions. *Journal of Big Data*, *8* (53), 1–74. https://doi.org/10.1186/s40537-021-00444-8

Chan, D. M., & Rao, R. & Huang, F., & Canny, J. F. (2018). T-SNE-CUDA: GPU-Accelerated T-SNE and its Applications to Modern Data. In *30th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD)*, 330–338. https://doi.org/10.1109/CAHPC.2018.8645912

Kirichenko, N., & Krivonos, Yu., & Lepekha, N. (2007). Synthesis of systems of neurofunctional transducers in solving classification problems. *Cybernetics and systems analysis*, *3*, 47–57.

Krak, I., & Barmak, O., & Manziuk, E. (2020). Using visual analytics to develop human and machine-centric models: A review of approaches and proposed information technology. *Computational Intelligence*, 1–26. https://doi.org/10.1111/coin.12289

Krak, I., Barmak, O., Manziuk, E., & Kulias, A. (2020). Data Classification Based on the Features Reduction and Piecewise Linear Separation. *Advances in Intelligent Systems and Computing*, *1072*, 282–289. https://doi.org/10.1007/978-3-030-33585-4_28

Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. In *Data Mining and Knowledge Discovery*, *28*, 92–122. https://core.ac.uk/download/pdf/41172947.pdf

Vahdat, A., & Kautz, J. (2020). NVAE: A Deep Hierarchical Variational Autoencoder. In *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*. https://doi.org/10.48550/arXiv.2007.03898

Zhang, G., & Chen, Y. (2020). *More Informed Random Sample Consensus*. https://doi.org/10.48550/arXiv.2011.09116

Zhou, N., & Shi, W., & Liang, R., & Zhong, N. (2022). TextRank Keyword Extraction Algorithm Using Word Vector Clustering Based on Rough Data-Deduction. *Computational Intelligence and Neuroscience*. https://doi.org/10.1155/2022/5649994

V. Kuznetsov, Iu. Krak, V. Liashko, V. Kasianiuk

ANALYSIS AND SYNTHESIS OF TECHNOLOGY FOR TEXTUAL INFORMATION CLASSIFICATION

The task of developing effective text information classification systems requires the thoughtful analysis and synthesis of variable components of technology. These components strongly affect the practical efficiency and the requirements to the data. For this purpose, a typical technology was discussed, comparing the regular "learning from features" approach versus the more advanced "deep learning" approach, that studies from data. In order to implement the technology, the first approach was tested, which included the means (methods, algorithms) for analysis of the features of the source text, by applying the dimensionality transformation, and building model solutions that allow the correct classification of data by a set of features. As a result, all the steps of the technology are described, which allowed to determine the way of presenting data in terms of hidden features in data, their presentation in a standard visual form and evaluate the solution, as well as its practical efficiency, based on this set of features. In a depth study, the informational core of the document was studied, using the regression and T-stochastic grouping of features for dimensionality reduction.

The separate results contain estimation of practical efficiency of the algorithms in terms of time and relative performance for each step of the proposed technology. This estimation gives a possibility to obtain the best algorithm of intelligent data processing that is useful for a given dataset and application. In order to estimate the best suited algorithm for separation in reduced dimension an experiment was carried out which allowed the selection of the best range of data classification algorithms, in particular boosting methods. As a result of the analysis of the technology, the necessary steps of this technology were discussed and the classification on real text data was conducted, which allowed to identify the most important stages of the technology for text classification.

Keywords: text processing, feature extraction, visual analytics, classification algorithms, feature dimensionality reduction.

Матеріал надійшов 10.09.2022