

МЕТОДИ БАЛАНСУВАННЯ НАБОРІВ ДАНИХ ПРИ РОЗВ'ЯЗАННІ ЗАДАЧ КАТЕГОРИЗАЦІЇ

Лановий О. Ф., Сорокіна А. С.

Харківський національний університет радіоелектроніки, Харків, Україна

Незбалансованість наборів даних є дуже поширеною проблемою, що потребує пильної уваги. Це мотивується тим, що проблема дисбалансу класів значно погіршує ефективність класифікації [1]. Однією з найважливіших складових вирішення цієї проблеми є вибір підходу та застосування методів, які найбільш ефективно розв'язують поставлену задачу. Існує декілька основних підходів, що допомагають вирішити надану проблему: це методи попередньої обробки (методи на рівні даних), алгоритмічно-центровані підходи та гібридні підходи, що є комбінацією перших двох [2]. В даній доповіді найбільше уваги приділяється саме методам модифікації наборів даних, що дозволяє досягати поставленої мети не змінюючи алгоритми машинного навчання, що використовують наданий датасет, на відміну від алгоритмічних підходів, які працюють з незмінними вихідними даними та основну задачу перекладають саме на алгоритм, змінюючи певним чином його ваги.

Метою доповіді є детальний опис, аналіз та порівняння різних підходів та методів усунення незбалансованості у наборах даних при розв'язанні задач категоризації з використанням машинного навчання.

В доповіді наводяться результати порівняння категоризації на наборах даних, збалансованих різними методами, до яких входить як недовідбір (undersampling) основного класу – тобто створення підмножини вихідних даних шляхом видалення вибраних вибірок із класу, так і надвідбір (oversampling) меншого класу – тобто створення надмножини вихідного набору даних шляхом створення нових вибірок з існуючих або шляхом тиражування існуючих [3]. Для розширення класів меншості було використано такі підходи, як об'єднання класів, викидання класів меншості, копіювання даних, копіювання даних з їх подальшим зашумленням, генерація додаткових зразків тощо.

Навчання моделі було виконано на модифікованих наборах тренувальних даних з подальшим порівнянням результатів виконання задач категоризації на тестових наборах даних.

Список літератури

1. Dudjak M., Martinović G. An empirical study of data intrinsic characteristics that make learning from imbalanced data difficult. Expert Systems with Applications. 2021. T. 182, C. 115297. DOI: <https://doi.org/10.1016/j.eswa.2021.115297>
2. Harsurinder Kaur, Husanbir Singh Pannu, Avleen Kaur Malhi. A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions. ACM Computer Surveys. T. 52, № 4. С. 1-36. DOI: <https://doi.org/10.1145/3343440>
3. Bach M, Werner A., Żywiec J., Pluskiewicz W. The study of under- and over-sampling methods' utility in analysis of highly imbalanced data on osteoporosis. Information Sciences. 2017. T. 384. С. 174-190. DOI: <https://doi.org/10.1016/j.ins.2016.09.038>