УДК 004.85:519.237

ДОСЛІДЖЕННЯ ВПЛИВУ НЕЗБАЛАНСОВАНОСТІ КЛАСІВ НА ЕФЕКТИВНІСТЬ АЛГОРИТМІВ КЛАСИФІКАЦІЇ

Стус М. М.[1]; Озарінська О. В.[2]; Татаренко П. П.[3]
[1,2,3]Національний університет «Одеська політехніка»
[1,2,3]Україна, Одеса
[1]odstus.mm@gmail.com; [2]o.v.ozarinska@op.edu.ua; [3]tatarenko.qw.p@gmail.com

Анотація. *Розглянуто вплив незбалансованого розподілу класів на оцінювання моделей класифікації. Порівняно випадкове збільшення та зменшення вибірки, синтетичне формування прикладів і зважування класів. Обґрунтовано обмеженість загальної точності та доцільність використання precision, recall, F1 і PR-AUC. Визначено умови коректного порівняння моделей без витоку даних.*
Ключові слова: *класифікація; незбалансованість класів; SMOTE; метрики якості; машинне навчання.*

Вступна частина. У задачах виявлення шахрайських операцій, технічних несправностей і рідкісних подій кількість об’єктів одного класу часто суттєво перевищує кількість іншого. За такого розподілу класифікатор може досягати високої загальної точності, переважно передбачаючи більшість, але пропускати саме ті випадки, для яких створено систему. Тому незбалансованість класів є проблемою не лише навчання, а й коректного оцінювання результатів [1].

Мета роботи – порівняти основні способи врахування дисбалансу класів та визначити набір метрик, що адекватно характеризують якість моделей за рідкісного позитивного класу. Робота має порівняльно-аналітичний характер; кількісні результати експерименту не заявляються.

Основна частина. Поширеним способом корекції є випадкове збільшення кількості прикладів меншості (oversampling). Воно зберігає всі вихідні спостереження, проте повторення тих самих об’єктів може посилювати перенавчання. Випадкове зменшення більшості (undersampling) скорочує час навчання, але ризикує вилучити інформативні приклади. Метод SMOTE створює синтетичні об’єкти шляхом інтерполяції між сусідніми прикладами меншості [2]. Його використання потребує обережності за наявності шуму, категоріальних ознак і перекриття класів.

Альтернативою зміні складу вибірки є зважування класів у функції втрат. За такого підходу помилки на рідкісному класі отримують більшу вагу, а навчальні дані залишаються незмінними. Зважування зручне для алгоритмів, що підтримують ваги спостережень, однак вибір коефіцієнтів впливає на співвідношення хибнопозитивних і хибнонегативних прогнозів. Жоден спосіб балансування не гарантує переваги для всіх наборів даних [3].

Показник accuracy є недостатнім за сильного дисбалансу. Наприклад, якщо позитивний клас становить 2 % вибірки, модель, яка завжди прогнозує негативний клас, матиме accuracy 98 %, але не виявить жодного позитивного випадку. Тому доцільно аналізувати precision – частку правильних позитивних прогнозів, recall – частку знайдених позитивних об’єктів, а також F1 як узагальнення precision і recall. Вибір метрики залежить від вартості різних типів помилок.

Для порівняння моделей за різних порогів рішення використовують ROC-AUC та площу під кривою «точність–повнота» (PR-AUC). За рідкісного позитивного класу PR-AUC часто інформативніша щодо якості позитивних прогнозів, оскільки враховує їхню точність за змінної повноти [4]. Порівнювати PR-AUC між вибірками з різною поширеністю позитивного класу слід обережно, адже її базовий рівень залежить від цієї поширеності.

Коректне оцінювання моделей за умов незбалансованості класів потребує належної організації навчальної та тестової вибірок. Одним із найбільш поширених підходів є

стратифікований поділ даних, який забезпечує збереження приблизно однакового співвідношення класів у сформованих підвибірках. Це особливо важливо за невеликої кількості позитивних прикладів, оскільки випадковий поділ без стратифікації може призвести до недостатнього представлення рідкісного класу в окремих частинах даних. Процедури балансування, масштабування ознак і добору гіперпараметрів необхідно виконувати виключно на навчальних даних у межах кожного циклу перехресної перевірки. Застосування SMOTE до всього набору даних перед його розділенням створює ризик витоку інформації, оскільки синтетичні спостереження можуть формуватися з використанням об'єктів, які згодом потраплять до тестової вибірки. У результаті модель може демонструвати штучно завищені показники якості, що не відповідатимуть її ефективності на нових даних. Крім того, важливо забезпечити однакові умови порівняння різних способів балансування, використовуючи спільні навчальні та тестові розбиття й узгоджені параметри оцінювання. Підсумкове тестування слід проводити на незалежній вибірці зі збереженим природним співвідношенням класів, що дозволяє отримати реалістичнішу оцінку практичної придатності класифікатора.

Висновки. Установлено, що незбалансованість класів може приховувати низьку якість розпізнавання рідкісних подій за високої загальної точності. Порівняно переваги й обмеження випадкового балансування, SMOTE та зважування класів. Обґрунтовано необхідність спільного аналізу precision, recall, F1 і PR-AUC та перевірки на тестових даних із природним співвідношенням класів. Подальша робота передбачає експериментальне порівняння підходів за різної частки позитивних прикладів.

Використання інструментів ШІ: не використовувались.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

[2] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. https://doi.org/10.1613/jair.953

[3] Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from imbalanced data sets*. https://doi.org/10.1007/978-3-319-98074-4

[4] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Stus M. M., Ozarinska O. V., Tatarenko P. P.

Investigating the impact of class imbalance on classification algorithm performance

Abstract. *The paper examines class imbalance in classification tasks and compares random oversampling, undersampling, SMOTE, and class weighting. It explains why accuracy can be misleading and discusses precision, recall, F1, and PR-AUC as evaluation measures. Particular attention is given to stratified validation and preventing data leakage during resampling. The study is analytical and does not claim unverified experimental improvements.*

Keywords: *class imbalance; classification; SMOTE; performance metrics; machine learning.*