POZNAN UNIVERSITY OF TECHNOLOGY

Mateusz Lango

Analiza źródeł trudności wieloklasowych danych niezbalansowanych oraz ich wykorzystanie do poprawy metod klasyfikacji

Streszczenie rozprawy doktorskiej

Promotor: prof. dr hab. inż. Jerzy Stefanowski

Wydział Informatyki i Telekomunikacji
Instytut Informatyki
Politechnika Poznańska

Poznań, 2021

Streszczenie

Uczenie klasyfikatorów jest podstawowym obszarem badawczym w nadzorowanym uczeniu maszynowym zajmującym się automatyczną konstrukcją systemów zdolnych do przypisywania instancji testowych do predefiniowanych klas na podstawie zbioru danych [7]. Tak ogólnie postawiony problem doprowadził do opracowania licznych algorytmów, które są szeroko wykorzystywane w wielu obszarach, takich jak rekomendacja produktów [18], identyfikacja autorstwa [10] czy filtrowanie wiadomości [1]. Pomimo znaczących sukcesów w licznych zastosowaniach, pewne problemy wciąż pozostają otwarte i utrudniają powszechne wykorzystanie metod uczenia maszynowego w specyficznych, a zarazem ważnych dziedzinach zastosowań. Jednym z takich problemów jest problem uczenia z danych niezbalansowanych [20, 13].

Zbiór danych nazywamy niezbalansowanym jeśli zawiera on klasy o różnych licznościach, a co najmniej jedna z klas nie jest dostatecznie dobrze reprezentowana [5]. Te niedoreprezentowane klasy nazywamy klasami mniejszościowymi, a ich skuteczne rozpoznawanie jest kluczowe w wielu praktycznych problemach takich jak analiza wydźwięku tekstu, automatyczna konstrukcja grafów czy analiza danych medycznych. Przeciwnie do oczekiwań praktyków preferujących wysoką trafność rozpoznawania klas mniejszościowych, klasyczne metody uczenia maszynowego konstruują klasyfikatory dobrze rozpoznające przede wszystkim klasy większościowe. W ekstremalnych przypadkach skonstruowany klasyfikator całkowicie ignoruje klasy mniejszościowe, nie będąc w stanie zaklasyfikować do nich żadnego przykładu testowego [14].

Początkowo sądzono, że to niepożądane zachowanie systemów uczących się wynika jedynie z poziomu niezbalansowania (ang. *imbalance ratio*), czyli ze znacznej rozbieżności pomiędzy wielkościami klas w zbiorze treningowym. Jednakże późniejsze analizy eksperymentalne wykazały, że dla niektórych prostych problemów klasyfikacyjnych poziom niezbalansowania klas nie ma prawie żadnego wpływu na ostatecznie skonstruowany klasyfikator [11, 17]. Zaobserwowano, że poziom niezbalansowania wpływa na konstrukcję klasyfikatorów i znacząco pogarsza rozpoznawanie klas mniejszościowych jedynie wtedy, gdy występuje razem z innymi czynnikami trudności danych (ang. *data difficulty factors*) [12, 17, 15]. Czynniki trudności danych można podzielić na czynniki globalne, które wpływają na wszystkie przykłady w zbiorze danych oraz czynniki lokalne, które dotyczą tylko pewnego podzbioru instancji. Przykładem globalnego czynnika trudności danych jest omawiany poziom niezbalansowania klas, natomiast nakładanie się klas, dekompozycja klasy na podkoncepty czy znaczna liczba obserwacji odstających to typowe przykłady

czynników lokalnych.

Ze względu na duże znaczenie praktyczne uczenia z danych niezbalansowanych, do jego rozwiązania zaproponowano wiele specjalistycznych metod, które można podzielić na metody algorytmiczne, metody modyfikujące rozkład danych oraz metody kosztowe [2, 8]. Metody algorytmiczne próbują rozwiązać trudności związane z konstrukcją klasyfikatorów z danych niezbalansowanych poprzez dostosowanie do nich istniejących systemów uczących się. W przeciwieństwie do tych metod, które skupiają się na modyfikacji konkretnych algorytmów, metody modyfikujące rozkład danych są uniwersalne i niezależne od wybranego algorytmu uczącego. Metody te modyfikują niezbalansowany zbiór danych w taki sposób, aby uruchomiony na nim algorytm uczący osiągał lepsze wyniki na klasie mniejszościowej. Ostatnia grupa metod wykorzystuje algorytmy uczenia się z kosztami (ang. *cost-sensitive learning*) koncentrując konstrukcję systemu klasyfikacyjnego na klasach mniejszościowych poprzez przypisanie ich przykładom wyższych kosztów błędnej klasyfikacji.

Prawie wszystkie metody wchodzące w skład wyżej wymienionych kategorii zostały zaprojektowane jedynie dla niezbalansowanych problemów klasyfikacji binarnej. Niemniej jednak zjawisko niezbalansowania klas występuje również w wieloklasowych zbiorach danych. Na przykład w zastosowaniach medycznych możemy tworzyć klasyfikator, który automatycznie wyróżnia nagłe przypadki wśród pacjentów zgłaszających się do Szpitalnego Oddziału Ratunkowego. Jednakże szpital może być zainteresowany nie tylko identyfikacją pilnych przypadków, ale także wykrywaniem przypadków które powinny być leczone przez lekarzy pierwszego kontaktu lub pacjentów których leczenie powinno być prowadzone w innych szpitalach (np. z powodu braku wyspecjalizowanego oddziału). Taki problem byłby niezbalansowany, ponieważ, jak pokazują badania przeprowadzone w USA, pacjenci którzy powinni być leczeni przez lekarzy pierwszego kontaktu stanowią ponad 80% wszystkich przypadków na tego typu oddziałach. Przedstawiony problem klasyfikacji ma dwie klasy mniejszościowe, ale w praktyce występują również problemy z kilkoma klasami większościowymi lub zarówno z wieloma klasami większościowymi jak i mniejszościowymi [4].

Nieliczne zaproponowane metody dla problemów wieloklasowych ograniczają się do dekompozycji problemów na problemy binarne oraz inne wyspecjalizowane metody, w zdecydowanej większości adaptacje binarnych metod przetwarzania wstępnego. Te proste modyfikacje metod binarnych nie uwzględniają jednak bardziej złożonych relacji, jakie pojawiają się między klasami w wieloklasowych problemach niezrównoważonych i nie radzą sobie z dodatkowymi źródłami trudności zidentyfikowanymi przez praktyków. Co więcej, dotychczasowo przeprowadzone teoretyczne i eksperymentalne analizy czynników trudności uczenia z danych niezrównoważonych, z nielicznymi wyjątkami, również dotyczą problemów binarnych, a ich wyników nie można bezpośrednio odnieść do danych wieloklasowych [15, 19, 16].

Na podstawie powyższych przesłanek sformułowano następującą hipotezę niniejszej rozprawy:

> Można zaproponować nowe metody konstrukcji klasyfikatorów dla wieloklasowych danych niezbalansowanych, które będą uwzględniać informacje o źródłach trudności związanych z rozkładem danych zarówno na poziomie lokalnym jak i bardziej globalnym.

Powyższą hipotezę zweryfikowano poprzez przeprowadzenie eksperymentalnej analizy źró-

deł trudności w wieloklasowych niezbalansowanych zbiorach danych oraz zaproponowaniu metody ich identyfikacji w rzeczywistych zbiorach danych. Ponadto, zaproponowano nowe metody klasyfikacji wykorzystujące wykryte czynniki trudności, których skuteczność została zbadana zarówno na typowo wykorzystywanych w pracach badawczych zbiorach danych z repozytoriów UCI i KEEL, jak i na zbiorach danych z wybranego obszaru zastosowań (klasyfikacja wydźwięku). W szczególności zrealizowano następujące zadania:

- W celu wykonania eksperymentalnej analizy źródeł trudności w wieloklasowych danych niezbalansowanych, zaproponowano i zaimplementowano generator sztucznych danych, który pozwolił na zbadanie wpływu różnych czynników trudności na jakość predykcji standardowych algorytmów uczących oraz porównanie ich z wynikiem klasyfikatora optymalnego (tzw. klasyfikator Bayesa) [7].

  Przeprowadzona analiza wskazała na znaczący wpływ nakładania się klas, a w szczególności, że zwiększanie nakładania się klas ma bardziej znaczący wpływ na jakość klasyfikacji niż zwiększanie poziomu niezbalansowania. Wykazano również, że dane z wieloma klasami mniejszościowymi są trudniejsze niż z wieloma większościowymi oraz że nakładanie się klas mniejszościowych i większościowych bardziej obniża jakość predykcji niż nakładanie się klas mniejszościowych. Eksperymenty uwypukliły również szczególną rolę klas pośrednich, dotychczas rzadko omawianych w literaturze, m.in. pokazano zróżnicowany wpływ nakładania się klasy pośredniej na wydajność klasyfikacji, który zależy od tego, na jakiego typu klasę się nakłada.

- Zaprezentowano metodę wykrywania czynników trudności danych poprzez analizę poziomów bezpieczeństwa przykładów w rzeczywistych zbiorach danych. Proponowana definicja poziomu bezpieczeństwa zawiera współczynniki podobieństwa klas, które mogą modelować wzajemne relacje między klasami. W szczególności mogą one uchwycić zmienną rolę klas pośrednich. Pokazano również, że metoda ta jest pomocna w ocenie trudności rzeczywistych, wieloklasowych zbiorów niezbalansowanych.

- Zaprojektowano metodę wstępnego przetwarzania danych, Similarity Oversampling and Undersampling Preprocessing (SOUP), która wykorzystuje wnioski z przeprowadzonej wcześniej eksperymentalnej analizy czynników trudności. Metoda ta na podstawie obliczonych wcześniej stopni bezpieczeństwa przykładów, ukierunkowuje swoje działanie na konkretne części zbioru danych. SOUP poprzez nadlosowanie bezpiecznych przykładów mniejszościowych i odlosowanie przykładów większościowych ze strefy nakładania się klas, konstruuje nie tylko zbalansowany zbiór danych, ale również zbiór, z którego łatwiej można nauczyć klasyfikatory o wysokiej jakości rozpoznawania klas mniejszościowych.

  Przeprowadzona ocena eksperymentalna na 15 rzeczywistych i 4 sztucznych zbiorach danych wykazała, że SOUP uzyskuje na metryce G-mean lepsze wyniki niż inne metody wstępnego przetwarzania, takie jak Global-CS [21] czy Static-SMOTE [6]. Uzyskane różnice były statystycznie istotnie wg. sparowanych testów rangowych Wilcoxona ze standardowym progiem istotności tj. $\alpha = 0.05$. Co więcej, pojedynczy klasyfikator skonstruowany na przetworzonym przez SOUP zbiorze danych uzyskał niższą (tj. lepszą) średnią rangę w teście Friedmana niż popularne, wyspecjalizowane zespoły klasyfikatorów, które dekomponują problem wieloklasowy do serii problemów

binarnych i stosują w ramach nich binarne techniki nad- i od-losowania.

- Zaproponowano dwa nowe algorytmy konstrukcji zespołów klasyfikatorów, które rozszerzają algorytm bagging [3] dla wieloklasowych niezbalansowanych zbiorów danych. Pierwszy algorytm, Multi-class Roughly Balanced Bagging (MRBBag), jest rozszerzeniem algorytmu Roughly Balanced Bagging [9] dla binarnych danych niezbalansowanych, który okazał się uzyskiwać bardzo dobre wyniki na złożonych binarnych problemach niezbalansowanych, pomimo braku bezpośredniego brania pod uwagę czynników trudności danych. Drugi zaproponowany algorytm wykorzystuje z kolei metodę obliczania poziomów bezpieczeństwa do identyfikacji czynników trudności danych, stosując technikę SOUP i integrując ją z zespołem typu bagging (SOUP-Bagging).

  Eksperymenty przeprowadzone na kilkunastu zbiorach danych wykazały, że obie metody oferują lepszą jakość klasyfikacji pod względem miary G-mean niż metody dekompozycji, klasycznie stosowane dla wieloklasowych problemów niezbalansowanych. Jednakże metoda wykorzystująca informacje o czynnikach trudności tj. SOUP-Bagging uzyskała lepsze wyniki niż MRBBag. Co więcej, nawet pojedynczy klasyfikator wytrenowany na zbiorze danych wstępnie przetworzonym za pomocą SOUP również przewyższył zespół MRBBag w sensie średniej rangi w teście Friedmana.

- Wreszcie, pokazano przydatność proponowanych metod w jednym z wymagających obszarów zastosowań, jakim jest analiza wydźwięku. Przeprowadzono eksperymenty na 12 zróżnicowanych zbiorach tekstów, wśród których były zarówno krótkie opinie z sieci społecznościowych jak i profesjonalne recenzje, oraz na dwóch typach reprezentacji tekstu. Na niskowymiarowej reprezentacji ręcznie stworzonych wyspecjalizowanych cech, SOUP-Bagging uzyskał najlepsze wyniki spośród badanych metod. Z kolei MRBBag uzyskał nieco lepsze wyniki na wielowymiarowej reprezentacji automatycznie uczonych cech metodami semantyki dystrybucyjnej.

- Warto nadmienić, że podjęto również pewne wysiłki mające na celu popularyzację zaproponowanych metod. Jednym z rezultatów tych starań jest biblioteka opensource dla języka Python, kompatybilna z biblioteką sklearn[1].

[1] www.cs.put.poznan.pl/mlango/multiimbalance.php

Bibliografia

[1] Hamzah Al Najada and Xingquan Zhu. isrd: Spam review detection with imbalanced data distributions. In *Proceedings of the 2014 IEEE 15th international conference on information reuse and integration (IEEE IRI 2014)*, pages 553–560. IEEE, 2014.

[2] Paula Branco, Luís Torgo, and Rita P. Ribeiro. A survey of predictive modeling on imbalanced domains. *ACM Comput. Surv.*, 49(2), August 2016.

[3] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[4] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

[5] Nitesh V Chawla, Nathalie Japkowicz, and Aleksander Kotcz. Special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter*, 6(1):1–6, 2004.

[6] Francisco Fernández-Navarro, César Hervás-Martínez, and Pedro Antonio Gutiérrez. A dynamic over-sampling procedure based on sensitivity for multi-class problems. *Pattern Recognition*, 44(8):1821 – 1833, 2011.

[7] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[8] Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

[9] Shohei Hido, Hisashi Kashima, and Yutaka Takahashi. Roughly balanced bagging for imbalanced data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 2(5-6):412–426, 2009.

[10] John Houvardas and Efstathios Stamatatos. N-gram feature selection for authorship identification. In *International Conference on Artificial Intelligence: Methodology, systems, and applications*, pages 77–86. Springer, 2006.

[11] Nathalie Japkowicz. The class imbalance problem: Significance and strategies. In *Proceedingsroc. of the International Conference on Artificial Intelligence*, volume 56, 2000.

[12] Nathalie Japkowicz. Class imbalance: Are we focusing on the right issue? In *Proceedings of the ICML'03 Workshop on Learning from Imbalanced Data Sets*, ICML '03, 2003.

[13] Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016.

[14] Mateusz Lango, Zdeněk Žabokrtský, and Magda Ševčíková. Semi-automatic construction of word-formation networks. *Language Resources and Evaluation*, pages 1–30, 2020.

[15] Krystyna Napierala and Jerzy Stefanowski. Identification of different types of minority class examples in imbalanced data. In *7th International Conference on Hybrid artificial intelligent systems*, Lecture Notes in Computer Science, pages 139–150. Springer, 2012.

[16] Krystyna Napierała. *Improving Rule Classifiers For Imbalanced Data*. PhD thesis, Poznan University of Technology, Poznań, Poland, 2013.

[17] Ronaldo C. Prati, Gustavo Enrique de Almeida Prado Alves Batista, and Maria Carolina Monard. Class imbalances versus class overlapping: an analysis of a learning system behavior. In *MICAI 2004: Advances in Artificial Intelligence*, pages 312–321. Springer, 2004.

[18] Mojtaba Salehi and Isa Nakhai Kamalabadi. A hybrid recommendation approach based on attributes of products using genetic algorithm and naive bayes classifier. *International Journal of Business Information Systems*, 13(4):381–399, 2013.

[19] Jerzy Stefanowski. Dealing with data difficulty factors while learning from imbalanced data. In Stan Matwin and Jan Mielniczuk, editors, *Challenges in Computational Statistics and Data Mining*, pages 333–363. Springer International Publishing, 2016.

[20] Qiang Yang and Xindong Wu. 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making*, 5(04):597–604, 2006.

[21] Zhi-Hua Zhou and Xu-Ying Liu. On multi-class cost-sensitive learning. *Computational Intelligence*, 26(3):232–257, 2010.

Instytut Informatyki, Wydział Informatyki i Telekomunikacji
Politechnika Poznańska

Skład przy użyciu systemu LaTeX.

BibTeX:

```
@phdthesis{ Lango2021,
   author = "Mateusz Lango",
   title = "{Analiza źródeł trudności wieloklasowych danych niezbalansowanych oraz ich wykorzystanie do
poprawy metod klasyfikacji}",
   school = "Poznan University of Technology",
   address = "Pozna{\'n}, Poland",
   year = "2021",
}
```