

Streszczenie

Niniejsza rozprawa rozwija dziedzinę uczenia dla danych Positive-Unlabeled (PU), rozwiązując fundamentalne wyzwania w scenariuszach, gdzie tylko niektóre pozytywne przykłady są oetykietowane, podczas gdy wszystkie negatywne przykłady pozostają nieoetykietowane. Badania koncentrują się na scenariuszu no-SCAR, gdzie etykietowanie zależy od cech próbki, czyniąc klasyfikację znacznie trudniejszą niż w przypadku tradycyjnego założenia Selected Completely At Random (SCAR).

Praca przedstawia szereg osiągnięć naukowych. W pierwszej kolejności wykazano fundamentalne niekompatybilności między scenariuszami single sample i case control w uczeniu PU. Poprzez formalną analizę udowodniono, że stosowanie metod dla niepoprawnego scenariusza prowadzi do nieoptymalnej wydajności. Opracowane warianty uwzględniające scenariusz osiągają znaczące poprawy metryk (do 20 punktów procentowych) przy jednoczesnym ograniczeniu overfittingu.

Drugą kluczową innowacją jest wprowadzenie VAE-PU+OCC – podejścia generatywnego, które wzbogaca uczenie PU oparte na wariacyjnych autoenkoderach o klasyfikację jednoklasową. Metoda eliminuje problematyczne bezpośrednie wykorzystanie sztucznie wygenerowanych przykładów, zastępując je identyfikacją rzeczywistych obserwacji positive-unlabeled w oryginalnym zbiorze danych, i usuwa konieczność obcinania funkcji ryzyka, zachowując jednocześnie gwarancje teoretyczne.

Kolejnym wkładem jest opracowanie pierwszej empirycznej procedury kontroli False Omission Rate w detekcji obserwacji odstających, dostarczając matematycznych podstaw oraz praktycznych algorytmów umożliwiających kontrolowanie proporcji niewykrytych wartości odstających wśród obserwacji zaklasyfikowanych jako normalne.

W dalszej części zidentyfikowano i sformalizowano *rozszerzone* uczenie PU, gdzie informacja o statusie etykietowania jest dostępna podczas predykcji. Wyprowadzona została optymalna reguła decyzyjna Bayesa d_B^{PU} , która została zintegrowana z metodą VAE-PU-Bayes, skutkując konsekwentnymi poprawami wydajności również w klasycznych zadaniach PU.

Ostatnim elementem badań jest analiza label shift dla rozszerzonych danych PU, gdzie prawdopodobieństwa a priori klas różnią się między zbiorem treningowym i testowym. Przeprowadzona analiza teoretyczna wykazuje, że optymalna klasyfikacja wymaga jedynie modyfikacji progu decyzyjnego, prowadząc do opracowania praktycznych algorytmów wraz z udowodnieniem ich skuteczności.

Zaproponowane metody demonstrują istotne empiryczne usprawnienia na zbiorach danych zarówno syntetycznych, jak i rzeczywistych, przy czym VAE-PU-Bayes wyłania się jako najbardziej efektywna spośród opracowanych metod. Przedstawione badania ustanawiają nowe fundamenty teoretyczne, dostarczając jednocześnie praktycznych narzędzi, które znacząco poszerzają możliwości uczenia się z danych PU w wymagających, realistycznych warunkach.

Abstract

This thesis advances the field of Positive-Unlabeled (PU) learning by addressing fundamental challenges in scenarios where only some positive examples are labeled while all negative examples remain unlabeled. The research focuses on the complex no-SCAR setting, where labeling depends on features, making classification significantly more challenging than under traditional Selected Completely At Random (SCAR) assumption.

The thesis makes several key contributions. First, it establishes fundamental incompatibilities between single sample and case control scenarios in PU learning, proving through formal analysis that cross-scenario application of methods leads to suboptimal performance. This work introduces scenario-aware variants achieving substantial improvements up to 20 percentage points while reducing overfitting.

Moreover, the thesis develops VAE-PU+OCC, a novel generative approach that enhances variational autoencoder-based PU learning through one-class classification. This method replaces problematic direct use of generated examples with identification of true positive-unlabeled examples within the original dataset, eliminating the need for risk truncation while ensuring theoretical guarantees.

Furthermore, the research introduces the first empirical procedure for False Omission Rate control in outlier detection, providing mathematical foundations and practical algorithms for controlling the proportion of undetected outliers among observations classified as inliers – critical for applications requiring high sensitivity and controlled false negative rates.

In the next step, the thesis identifies and formalizes augmented PU learning, where labeling status is available at prediction time. The optimal Bayesian decision rule d_B^{PU} is derived and integrated into VAE-PU-Bayes, achieving consistent performance improvements in classic PU task.

Finally, the work addresses label shift in *augmented* PU data, where class prior probabilities differ between training and test datasets. Theoretical analysis establishes that optimal classification requires only threshold modification, leading to practical algorithms with proven effectiveness.

The proposed methods demonstrate substantial empirical improvements across synthetic and real-world datasets, with VAE-PU-Bayes emerging as the strongest performer. This research establishes new theoretical foundations while providing practical tools that significantly advance the state-of-the-art in learning from partially labeled data under realistic, challenging conditions.