Streszczenie

W ninejszej pracy podjęto problem klasyfikacji z wykorzstaniem zbioru otwartego w danych wysokowymiarowych. Jest to zadanie polegające na wykrywaniu nowych danych – wartości odstających lub przykładów spoza rozkładu (ang. *out-of-distribution*, OOD), które znacząco różnią się od wszelkich wcześniej dostępnych/znanych próbek, czyli danych treningowych, wykorzystywanych do uczenia klasyfikatora zamkniętego. Chociaż zadanie to jest mocno ugruntowane i opisane na solidnym tle statystycznym w ogóle, okazuje się ono trudne i wciąż nierozwiązane w przypadku przestrzeni cech o wysokim wymiarze, gdzie wszystkie obecne podejścia i środki okazują się niewystarczające. Co więcej, chociaż zaproponowano wiele metod wykrywania danych odstających w takim zadaniu, to w obszernej literaturze można znaleźć sprzeczne rekomendacje na temat opisywanych rozwiązań.

Motywacją do podjęcia tego problemu badawczego jest znaczący postęp i zdumiewająca skuteczność modeli głębokiego uczenia się (CNN lub ViT) w zadaniach klasyfikacji obrazów, jakie raportowane są w analizach porównawczych. Stosowane modele obejmują wysokowymiarowe wektory cech ($d \sim 10^3$), jednak nadal opierają się na rozpoznawaniu w ramach zbiorów zamkniętych. Popularyzacja narzędzi sztucznej inteligencji i ostatnie postępy w dziedzinie uczenia maszynowego, takie jak pojawienie się złożonych technik głębokiego uczenia oraz rosnące zainteresowanie samochodami autonomicznymi, a także innymi pojazdami tego typu - wszystko to sprawia, że zagadnienia niezawodności i bezpieczeństwa są obecnie niezwykle ważnymi tematami. Problem wykrywania danych odstających wpisuje się w tę tematykę, ponieważ jednym z kluczowych aspektów związanych z niezawodnością modeli uczenia maszynowego jest zdolność modeli do adaptacji do nowych danych i sytuacji. W każdej rzeczywistej implementacji systemów opartych na uczeniu maszynowym, gdzie bezpieczeństwo ma charakter krytyczny, niezawodne wykrywanie OOD jest podstawowym wymogiem, zapewniającym stabilność działania w przypadkach przeoczonych lub pominiętych podczas uczenia modeli. Jednak, jak podkreślają czołowi naukowcy w tej dziedzinie, takie aspekty są jak dotąd ciągle niedostatecznie zbadane w literaturze.

Niniejsza rozprawa składa się z trzech zasadniczych części. W pierwszej części omówiono niezbędne tło dotyczące technik wykrywania danych odstających, koncentrując się na rozróżnieniu głównych podejść zaproponowanych już w literaturze. Podano szczegółowy opis wybranych metod *post-hoc*, a także wymagany formalizm i notację dla zadania klasyfikacji z uwzględnieniem zbioru otwartego. Główne zainteresowania badawcze skupiają się w pracy na metodach pracujących w przestrzeniach cech, ze względu na ich uniwersalność – możliwość zastosowania ich do dowolnego już istniejącego, wcześniej wytrenowanego modelu.

Następnie, w dominującym objętościowo rozdziale, opisano wyniki przeprowadzonych badań numerycznych na symulowanych rozkładach danych. Przeanalizowano skuteczność wybranych metod *post-hoc* w wykrywaniu danych odstających, biorąc pod uwagę takie czynniki, jak wymiary wektorów cech, liczba próbek uczących i odległość do przykładów odstających – aby sprawdzić, jak dobrze różne metody potrafią odróżnić zarówno próbki treningowe, jak i testowe, od danych odstających. Dodatkowo analizowany jest wpływ obecności korelacji w danych na działanie metod oraz zachowanie się metod, gdy cechy charakteryzują się niejednorodnymi wariancjami, czyli gdy dane są niezestandaryzowane. Przeprowadzone badania wykazują nieoczywiste zachowania niektórych z badanych metod, co jest szczególnie widoczne przy wysokich wymiarach wektorów cech. W pracy wykazano, że metody te posiadają znacząco różne możliwości rozróżnienia danych znanych (ang. *in-distribution*, ID) od danych nieznanych (ang. *out-of-distribution*, OOD). Ponadto w badaniu określono warunki wymagane, aby metody mogły zapewnić wierne odwzorowanie danych uczących.

Na koniec przeprowodzana jest konfrontacja metod z danymi ze świata rzeczywistego. Szeroki zakres wstępnie wytrenowanych algorytmów reprezentacji danych jest wykorzystywany do uzyskania wektorów cech z dokumentów tekstowych i danych obrazowych, które następnie są badane pod kątem ich potencjału w zadaniu klasyfikacji ze zbiorem otwartym w odniesieniu do danych uczących. Zaobserwowano i omówiono wiele znaczących różnic pomiędzy technikami reprezentacji. Okazuje się, że właściwości reprezentacji mają duży wpływ na skuteczność detektorów OOD w postawionym zadaniu. Można zatem sformułować wytyczne dotyczące wyboru metod odpowiednich dla konkretnej reprezentacji. Pokazano, że dla wszystkich analizowanych przypadków istnieje znaczna liczba klas zawierających próbki znacznie trudniejsze do odróżnienia od wartości odstających, dlatego też dla zastosowań krytycznych dla bezpieczeństwa proponuje się analizę separowalności ID-OOD dla poszczególnych klas. Taka ocena pozwala zidentyfikować luki bezpieczeństwa i ryzyka związane z klasami o słabej generalizacji w zadaniu wykrywania OOD; klasy te mogą wymagać dokładniejszych analiz w postawionym zadaniu.

Przeprowadzone badania stanowią wkład dziedzinę, dostarczając nowego wglądu w zachowania i właściwości wybranych metod *post-hoc* w odniesieniu do wysokowymiarowych przestrzeni cech. Podano zalecenia dotyczące wyboru, stosowania i kalibracji metod OOD dla poszczególnych reprezentacji danych w zadaniu wykrywania danych odstających, uwzględniając zastosowania w danych obrazowych i tekstowych.

Słowa kluczowe

– Klasyfikacja z wykorzystaniem zbioru otwartego.

– Wykrywanie danych spoza rozkładu.

– Miary odstania danych.

– Wysokowymiarowe wektory cech.

– Techniki reprezentacji danych.