Wydział Elektroniki
Katedra Systemów i Sieci Komputerowych
Politechnika Wrocławska

Classifier selection for imbalanced data stream classification

Selekcja klasyfikatorów w zadaniu klasyfikacji niezbalansowanych strumieni danych
(streszczenie)

Paweł Zyblewski

Słowa kluczowe: *Rozpoznawanie wzorców; uczenie indukcyjne; klasyfikacja; zespół klasyfikatorów; selekcja klasyfikatorów; przetwarzanie wstępne danych; dane trudne; dane niezbalansowane; strumienie danych; dryf koncepcji; uczenie aktywne.*

Wrocław 04.2021

Streszczenie

Rozprawa doktorska koncentruje się na wykorzystaniu algorytmów *Dynamicznej Selekcji Zespołu Klasyfikatorów* w połączeniu z metodami przetwarzania wstępnego w zadaniu klasyfikacji statycznych oraz strumieniowych danych niezbalansowanych. Celem pracy było przedstawione naturalnej zdolności algorytmów selekcji klasyfikatorów do radzenia sobie z niezbalansowaniem danych oraz zaproponowanie nowych, efektywnych rozwiązań rzadko poruszanego w literaturze problemu klasyfikacji wysoce niezbalansowanych strumieni danych. W oparciu o te założenia, w pracy sformułowana została hipoteza, zakładająca, że

> *Istnieją metody wykorzystujące zarówno wstępne przetwarzanie danych, jak i metody selekcji klasyfikatorów, które przewyższają jakość predykcji znanych z literatury metod stosowanych w klasyfikacji danych trudnych.*

Hipoteza została uprawdopodobniona poprzez osiągnięcie poniższych celów:

***Cel 1* – Opracowanie algorytmu selekcji zespołu klasyfikatorów na potrzeby klasyfikacji danych niezbalansowanych oraz zaprojektowanie dedykowanej reguły kombinacji.**
Cel został zrealizowany poprzez opracowanie trzech algorytmów, opartych na grupowaniu modeli bazowych w jednowymiarowej przestrzeni różnorodności klasyfikatorów. Podstawę do utworzenia tej przestrzeni stanowiła zaproponowana miara H, informująca o wpływie poszczególnych klasyfikatorów na różnorodność osiąganą przez cały zespół.

Algorytm *Diversity Ensemble Pruning* (DEP) dokonuje grupowania modeli bazowych w przestrzeni różnorodności, a następnie ocenia jakość klasyfikacji poszczególnych klasyfikatorów w oparciu o *zbalansowaną dokładność*. Do finalnego zespołu wybierany jest, z każdego klastra, model o najwyższej wartości BAC. Algorytm *Two-step majority voting organization* (TSMV), zamiast redukować liczność zespołu, dokonuje klasyfikacji danych niezbalansowanych z wykorzystaniem struktury głosowania dwuetapowego. W pierwszym etapie głosowania, każdy klaster traktowany jest jako osobny zespół klasyfikatorów, który niezależnie podejmuje decyzję w oparciu o *głosowanie większościowe*.
W drugim etapie, ponownie poprzez *głosowanie większościowe*, kombinowane są decyzje

uzyskane przez poszczególne klastry. Algorytm *Random Sampling Multistage Organization* (RSMO), będący modyfikacją TSVM, wykorzystuje dodatkowo operację losowania ze zwracaniem w celu zredukowanie liczby podobnych klasyfikatorów wykorzystywaych w procesie podejmowania decyzji.

***Cel 2* – Opracowanie algorytmu *Dynamicznej Selekcji Klasyfikatorów* opartego o miary dystansu, na potrzeby klasyfikacji danych niezbalansowanych.**
Cel został zrealizowany poprzez opracowanie dwóch algorytmów Dynamicznej Selekcji Klasyfikatorów, które oceniają kompetencje modeli bazowych w zależności od decyzji podjętych przez nie w odniesieniu do przypadków znajdujących się w lokalnym sąsiedztwie klasyfikowanej instancji, jednocześnie uwzględniając odległość Euklidesową do tych przypadków. *Dynamic Ensemble Selection using Euclidean distance* (DESE) wykorzystuje do selekcji wyłącznie decyzje klasyfikatorów oraz odległości, natomiast *Dynamic Ensemble Selection using Imbalance Ratio and Euclidean distance* (DESIRE) dodatkowo modyfikuje otrzymane wagi w oparciu o stopień niezbalansowania klasyfikowanego problemu.

***Cel 3* – Opracowanie opartego o przetwarzanie wsadowe algorytmu klasyfikacji wysoce niezbalansowanych strumieni danych.**
Cel został zrealizowany poprzez zaproponowanie algorytmu *Minority Driven Ensemble* (MDE). Algorytm ten dokonuje klasyfikacji wysoce niezbalansowanych strumieni danych z użyciem reguły decyzyjnej, która wykorzystuje lokalną charakterystykę danych do preferowania klasy mniejszościowej.

***Cel 4* – Zaprojektowanie metody łączącej Dynamiczną Selekcją Klasyfikatorów oraz przetwarzanie wstępne danych, na potrzeby klasyfikacji niezbalansowanych danych strumieniowych.**
Cel został osiągnięty poprzez zaproponowanie dwóch, opartych o przetwarzanie wsadowe, podejść do łączenia algorytmów Dynamicznej Selekcji Klasyfikatorów oraz technik przetwarzania wstępnego na potrzeby klasyfikacji wysoce niezbalansowanych strumieni danych. Metoda *Dynamic Ensemble Selection for Imbalanced Stream Classification* (DESISC) generuje pojedynczy model na każdej nowej porcji danych, podczas gdy podejście *Dynamic Ensemble Selection for Imbalanced Stream Classification using Stratified Bagging* (DESISC-SB) wykorzystuje do tego celu stratyfikowaną wersję *Baggingu*.

***Cel 5* – Zaproponowanie strategii budowania modeli klasyfikacji w przypadku strumieni danych z ograniczonym dostępem do etykiet.**
Cel został osiągnięty poprzez zapoponowanie strategii odpytywania o etykiety, nazwanej *Budget Active Labeling Strategy* (BALS). Algorytm ten łączy w sobie losowe podejście do etykietyzacji z podejściem właściwym algorytmom uczenia aktywnego. Dzięki temu oprócz puli instancji wybranych na podstawie ich odległości od granicy decyzyjnej, etykiety pozyskiwane są również dla małej liczby obiektów losowo wybranych z aktualnej porcji danych.

***Cel 6* – Ewaluacja zaproponowanego wczesniej frameworku klasyfikacji strumieni danych, w przypadku ograniczonego dostępu do etykiet.**
Cel został osiągnięty poprzez połączenie metody DESISC-SB z podejściem do uczenia aktywnego, opartym na przekazywaniu do etykietyzacji przypadków znajdujących się w określonej odległości od granicy decyzyjnej problemu.

***Cel 7* – Przeprowadzenie ewaluacji eksperymentalnej, porównującej zaproponowane algorytmy z podejściami stanowiącymi *state-of-the-art*.**
***Cel 8* – Opracowanie biblioteki języka *Python*, pozwalającej na analizę trudnych strumieni danych.**
Cele 7 i 8 zostały osiągnięte dzięki zaprojektowaniu oraz implementacji środowiska eksperymentalnego w języku *Python*, które posłużyło do przeprowadzenia badań związanych z klasyfikacją danych niezbalansowanych. Dodatkowo, w trakcie pracy na rozprawą, opracowana została biblioteka *stream-learn*[1], pozwalająca na przetwarzanie niezbalansowanych strumieni danych z dryfem koncepcji. Biblioteka ta została wykorzystana do przeprowadzenia wszystkich eksperymentów związanych z danymi strumieniowymi.

Słowa kluczowe
Rozpoznawanie wzorców; uczenie indukcyjne; klasyfikacja; zespół klasyfikatorów; selekcja klasyfikatorów; przetwarzanie wstępne danych; dane trudne; dane niezbalansowane; strumienie danych; dryf koncepcji; uczenie aktywne.

[1]Ksieniewicz, P. and Zyblewski, P., 2020. stream-learn–open-source Python library for difficult data stream batch analysis. arXiv preprint arXiv:2001.11077.