Zastosowanie zespołów klasyfikatorów w klasyfikacji niezbalansowanych strumieni danych

Jakub Klikowski

Wrocław 01.12.2021

Streszczenie

Rozprawa koncentruje sie na zadaniach klasyfikacji binarnych niezbalansowanych strumieni danych z dryfem koncepcji. Zaproponowano wykorzystanie hybrydowych technik próbkowania z akumulacja danych i modelami klasyfikatorów jednoklasowych. Przeprowadzone badania i uzyskane wyniki jednoznacznie wskazuja, że proponowane podejścia maja znaczny potencjał, a wyniki uzyskane opracowanymi metodami czesto dorównuja lub przewyższaja znane z literatury algorytmy. Bazujac na powyższych założeniach została sformułowana nastepujaca hipoteza badawcza:

Można zaprojektować metody zespołów klasyfikatorów wykorzystujace próbkowanie danych oraz jednoklasowe klasyfikatory, które potrafia osiagnać jakość predykcji lepsza niż znane z literatury klasyfikatory niezbalansowanych strumieni danych.

Powyższa hipoteza została uprawdopodobniona poprzez realizacje nastepujacych celów badawczych:

Zaprojektowanie metody zespołu klasyfikatorów jednoklasowych w celu klasyfikacji niezbalansowanego strumienia danych.

Cel ten został osiagniety poprzez zaproponowanie algorytmu klasyfikacji niezbalansowanych strumieni danych w oparciu o zespół klasyfikatorów jednoklasowy - **O**ne **C**lass support vector machine classifier **E**nsemble for **I**mbalanced data

Stream (*OCEIS*). Metoda osiaga wyniki na zbliżonym poziomie do porównywanych metod, ale warto zauważyć, że najlepiej sprawdza sie na strumieniach rzeczywistych, co jest jej istotna zaleta. Kolejna zaleta jest brak tendencji do nadmiernej klasyfikacji obiektów z jednej z klas. Taka stabilność" znaczaco przyczynia sie do poprawy jakości klasyfikacji i uzyskania satysfakcjonujacych wyników.

Rozszerzenie metody zespołu klasyfikatorów jednoklasowych poprzez wprowadzenie ulepszonej ważonej reguły decyzyjnej dla lepszej adaptacji do niezbalansowanych strumieni danych.

Cel ten został osiagniety poprzez zaprojektowanie nowatorskiej metody klasyfikacji niezbalansowanych strumieni danych - **O**ne **C**lass support vector machine **W**eighted **E**nsemble (*OCWE*). Jest to metoda oparta na zespołach klasyfikatorów jednoklasowych, rozszerzajaca pierwotna idee z podejścia *OCEIS*. Wprowadzone modyfikacje doprowadziły do wzrostu jakości klasyfikacji niezbalansowanych strumieni danych. W porównaniu z innymi metodami referencyjnymi, algorytm *OCWE* uzyskuje podobna jakość, a w pewnych warunkach *OCWE* osiaga lepsza wydajność. Metoda ta bardzo dobrze sprawdza sie w zadaniach klasyfikacji silnie niezbalansowanych danych. Wielość parametrów zapewnia dobra zdolność adaptacji do aktualnie rozwiazywanego problemu.

Zaproponowanie klasyfikatora dla niezbalansowanych strumieni danych z hybrydowa technika próbkowania i akumulacji danych.

Cel ten został osiagniety poprzez opracowanie metody klasyfikacji strumieni niezbalansowanych. **D**eterministic **S**ampling **C**lassifier (*DSC*) wypadło korzystnie w porównaniu z innymi dedykowanymi algorytmami. Ocene zdolności predykcyjnych technik przeprowadzono na podstawie eksperymentów komputerowych. Algorytm wykorzystuje bufor pamieci w celu propagacji wybranych instancji z poprzednich fragmentów strumieni danych. Ponieważ bufor ma stały rozmiar, po jego zapełnieniu niektóre instancje musza zostać usuniete. W zaproponowanej wersji algorytmu przyjeto, że usuwane sa najstarsze przykłady.

Ulepszenie klasyfikatora z hybrydowym próbkowaniem akumulacji danych do ważonego zespołu z wykorzystaniem baggingu w celu klasyfikacji niezbalansowanego strumienia danych.

Cel ten został osiagniety poprzez zaproponowanie nowej metody klasyfikacji niezbalansowanych strumieni danych - **D**eterministic **S**ampling **C**lassifier with weighted **B**agging (*DSCB*). *DSCB* wykorzystuje koncepcje algorytmu *DSC* oraz perturbowane zbiory danych. Proponowana metoda *DSCB* to zespół klasyfikatorów, którego główna idea skupia sie na tworzeniu nowych modeli poprzez użycie ważonego *baggingu* i danych, które zostały zgromadzone z poprzednich cześci strumienia danych. Eksperymenty komputerowe potwierdziły, że zaproponowane podejście uzyskuje dobre wyniki w stosunku do jakości klasyfikacji wybranych metod referencyjnych.

Poprawa klasyfikatora niezbalansowanych strumieni danych z hybrydowym próbkowaniem akumulacji danych do metody zespołowej z detektorem dryfu koncepcji.

Cel ten osiagnieto dzieki zaproponowaniu nowatorskiej, skutecznej metody zespołu klasyfikatorów w połaczeniu ze struktura przetwarzania wstepnego dla zadania klasyfikacji dryfujacego, niezbalansowanego strumienia danych. Główna idea *DSE* opiera sie na akumulacji wybranych próbek z poprzednich fragmentów w celu późniejszego wzmocnienia i zrównoważenia danych. Pozwala to na specyficzny *oversampling*, który nie wymaga generowania sztucznych danych. Badania przeprowadzone na szerokiej gamie strumieni danych rzeczywistych i generowanych komputerowo potwierdziły skuteczność proponowanego rozwiazania. Podkreślono jego mocne strony w porównaniu z metodami referencyjnymi oraz jego wysoka odporność na szum etykiet.

Zaprojektowanie metody zespołowej do klasyfikacji niezbalansowanych strumieni danych, która łaczy hybrydowe próbkowanie akumulacji danych i klasyfikatory jednoklasowe.

Cel ten osiagnieto dzieki opracowaniu nowej metody klasyfikacji niezbalansowanych strumieni danych — *DSE-OC*, która łaczy dwa różne podejścia. Jest to połaczenie najlepszych cech algorytmów *DSE* i *OCEIS*. W jednej metodzie istnieje jednocześnie adaptacja problemów binarnych do klasyfikacji przy użyciu jednoklasowych modeli SVM z metoda, która akumuluje dane do półsyntetycznego *oversamplingu* danych niezbalansowanych. Z przeprowadzonych badań wynika, że prezentowana metoda osiaga dość dobra jakość, a przeprowadzone analiza

statystyczna wyników eksperymentów wykazała, że metoda ta jest statystycznie istotnie lepsza metody referencyjne.