

Non-stationary data stream processing with metafeature analysis

Joanna Komorniczak

Streszczenie

Rozprawa doktorska koncentruje się na zagadnieniach związanych z przetwarzaniem niestacjonarnych strumieni danych. Proponowane w rozprawie metody uczenia maszynowego pozwalają na rozwiązywanie dwóch kluczowych dla tej dziedziny zadań: *detekcji dryfów koncepcji* oraz *klasyfikacji niestacjonarnych strumieni danych* w oparciu o *metody hybrydowe*. Zaproponowane oraz zbadane w rozprawie rozwiązania opierały się na analizie *metacech* obliczanych dla rozłącznych wsadów napływających w formie strumieni danych. Celem rozprawy była weryfikacja hipotezy:

Możliwa jest propozycja metod wykorzystujących analizę metacech na potrzeby detekcji dryfów koncepcji oraz klasyfikacji niestacjonarnych strumieni danych, które wykazują jakość rozpoznawania istotnie lepszą lub statystycznie zależną od podejść znanych z literatury.

W oparciu o przeprowadzone badania hipoteza została uprawdopodobniona poprzez osiągnięcie następujących celów:

Cel 1: Propozycja niejawnego detektora dryfu koncepcji analizującego zmienność w czasie miar złożoności zadania klasyfikacji obliczonych dla rozłącznych wsadów danych. Cel został zrealizowany poprzez propozycję metody *Complexity-based Drift Detector*, która analizuje metacechy opisujące złożoność zadania klasyfikacji. Działanie mechanizmu detekcji jest niezależne od wyników osiąganych przez model klasyfikacji, co pozwala na niejawną detekcję dryfu koncepcji. Zaproponowane podejście wykorzystuje zespół klasyfikatorów jednoklasowych szkolonych za pomocą metacech w celu rozróżnienia reprezentacji wsadów pochodzących z różnych koncepcji, co pozwala na efektywne rozpoznawanie istotnych zmian zachodzących w strumieniach danych.

Cel 2: Propozycja niejawnego detektora dryfu koncepcji analizującego miary magnitudy dryfu zintegrowane zgodnie z paradygmatem uczenia zespołowego. Cel został zrealizowany poprzez propozycję metody *Statistical Drift Detection Ensemble*, wykorzystującej miary *drift magnitude* oraz *conditioned marginal covariate drift* obliczane dla rozłącznych wsadów danych. Metoda detekcji dryfu wykorzystuje podejście zespołowe do integracji składowych decyzji o rozpoznaniu zmiany koncepcji.

Metacechy wykorzystane w tej metodzie dedykowane są analizie zmian koncepcji w strumieniach danych, jawnie opisując zmiany w rozkładzie prawdopodobieństwa. Bezpośrednie wykorzystanie tych miar wykazało ich niską informatywność w przypadku analizy danych wielowymiarowych. Zaproponowane w detektorze podejście zespołowe pozwala na efektywne rozpoznawanie dryfów dzięki analizie podprzestrzeni o niskiej wymiarowości i następnie integrację składowych detekcji w obrębie podprzestrzeni wykorzystując paradygmat uczenia zespołowego.

Cel 3: Propozycja nienadzorowanego detektora dryfu koncepcji analizującego rozkład aktywacji pochodzących z ostatniej warstwy deterministycznej sieci neuronowej. Cel został zrealizowany poprzez propozycję metody *Parallel Activations Drift Detector* pozwalającej na nienadzorowaną detekcję dryfów koncepcji w oparciu o wyjścia z deterministycznej sieci neuronowej. Wykorzystane w zaproponowanym podejściu metacechy zdefiniowane są jako losowe projekcje realizowane przez sieć neuronową zainicjalizowaną losowymi wagami, niezmiennymi w trakcie przetwarzania. Metacechy określone w ten sposób niejawnie opisują położenie próbek w wielowymiarowej przestrzeni. Analiza ich zmienności za pomocą wielokrotnie powtórzonych parowych testów statystycznych pozwala na identyfikację istotnych zmian w rozkładzie danych wskazujących na dryfy koncepcji.

Ze względu na brak wiarygodnego kryterium oceny jakości zadania detekcji dryfów koncepcji w literaturze, które pozwoliłoby na zbadanie jakości rozwiązania w przypadku strumieni danych z dryfami inkrementalnymi i gradualnymi, w rozprawie doktorskiej zostały zaproponowane trzy miary oceny jakości *błędu detekcji dryfu* bazujące na odległości oraz liczności dryfów i detekcji. Dodatkowo ocena jakości wspomagana była wizualną analizą momentów sygnalizacji zmian koncepcji.

Porównanie wszystkich zaproponowanych metod detekcji dryfów koncepcji z innymi detektorami znanymi z literatury na szerokiej puli syntetycznych strumieni danych pozwoliło na wykazanie, w zależności od badanego kryterium, istotnie lepszej lub statystycznie zależnej jakości detekcji w porównaniu do metod znanych z literatury.

Cel 4: Propozycja metody zespołowej do klasyfikacji strumieni danych analizującej rozkład statystycznych metacech obliczonych dla kolejnych wsadów danych w celu identyfikacji nawracającego konceptu. Cel został zrealizowany poprzez propozycję metody *Metafeature Concept Selector*, która wykorzystuje zestaw statystycznych metacech do detekcji dryfów koncepcji oraz ponownej identyfikacji koncepcji występującej w przeszłości, w przypadku koncepcji nawracających. Zaproponowany algorytm wykorzystuje pulę klasyfikatorów jednoklasowych do identyfikacji koncepcji na podstawie obliczonych metacech oraz, równolegle, pulę klasyfikatorów odpowiedzialnych za efektywne rozwiązywanie realizowanego zadania rozpoznawania.

W eksperymentach badających działanie metody wykorzystano ocenę jakości identyfikacji koncepcji nawracającej za pomocą metryki *Rand* dedykowanej ocenie zadań klasteryzacji. Metryka pozwoliła na jednoznaczną ocenę realizowanego zadania niezależnie od mechanizmu klasyfikacji zintegrowanego klasyfikatora. W dalszych eksperymentach porównano działanie zaproponowanego rozwiązania z działaniem niezależnych klasyfikatorów, co wykazało statystycznie znaczącą poprawę jakości klasyfikacji przy użyciu zaproponowanego mechanizmu.

Cel 5: Propozycja metody klasyfikacji pozwalającej na kompensację stronniczości klasyfikatorów podczas przetwarzania strumieni danych z dynamicznymi zmianami stopnia niebalansowania wykorzystujących estymację prawdopodobieństwa *a priori* w oparciu o metacechy przetwarzanych porcji danych. Cel został zrealizowany poprzez propozycję *Prior Probability Assisted Classifier*, wykorzystujący estymowane prawdopodobieństwo *a priori* w celu kompensacji stronniczości klasyfikatorów w kierunku klasy większościowej. Estymacja prawdopodobieństwa *a priori* odbywa się w oparciu o metacechy obliczane dla kolejnych wsadów strumieni danych. W oparciu o estymowany stopień niebalansowania w danym wsadzie, zaproponowana metoda koryguje predykcje klasyfikatora zapewniając określoną licznosc obiektów danej klasy.

W eksperymentach porównano różne strategie estymacji prawdopodobieństwa *a priori*, z których najbardziej uniwersalną stanowiła *Dynamic Statistical Concept Analysis*, będąca autorską metodą określającą aktualne prawdopodobieństwo *a priori* problemu w oparciu o parę regresorów analizujących wartości średnie i odchylenie standardowe próbek w obrębie przetwarzanych klas. Wyniki badań

wykazały statystycznie znaczącą poprawę jakości klasyfikacji przy zastosowaniu zaproponowanego podejścia.

Cel 6: Propozycja struktury przetwarzania strumieni danych ze zmiennym w czasie stopniem trudności, pozwalającej na wybór odpowiedniej architektury sieci neuronowej na podstawie analizy pewności modelu. Cel został zrealizowany poprzez propozycję *Certainty-based Architecture Selection Framework* – struktury przetwarzania wykorzystującej stopień *pewności* modelu klasyfikacji w zadaniu klasyfikacji strumieni danych o zmiennej w czasie trudności. Wykorzystanie *pewności* klasyfikatora jako metacechy pozwoliło na redukcję narzutu obliczeniowego związanego z określeniem dodatkowych wskaźników opisujących przetwarzane dane. Na podstawie wartości *funkcji wsparcia* pochodzących z sieci neuronowej obliczane były metacechy, a następnie predykcje dla przetwarzanych obiektów.

W przeprowadzonych eksperymentach realizowane było zadanie klasyfikacji półsyntetycznych strumieni danych wizji komputerowej. Zaproponowany schemat przetwarzania pozwala na przełączanie architektur *konwolucyjnych sieci neuronowych* o różnym stopniu złożoności, wyszkolonych dla zadanego problemu. Wyniki przeprowadzonych eksperymentów pokazały, że zaproponowane rozwiązanie pozwala na znaczące zredukowanie czasu przetwarzania oraz liczby operacji przy niewielkiej redukcji jakości rozpoznawania.

Słowa kluczowe uczenie maszyn, strumienie danych, dryf koncepcji, detekcja dryfu koncepcji, klasyfikacja, zespoły klasyfikatorów, metacechy, dane niezbalansowane

17.04.2025

Joanna Komorowska