

Streszczenie – Multi-objective optimization to train classifiers on feature subspaces

Joanna Klikowska

Maj 2024

Streszczenie

W pracy skupiono się na zastosowaniu optymalizacji wielokryterialnej w zadaniu selekcji cech w klasyfikacji głównie danych niebalansowanych. Selekcja cech wyznacza podprzestrzeń cech dla każdego modelu, a mechanizm ten zapewnia różnorodność modeli w zespole. Badania w tym obszarze wykazały, że optymalizacja pod kątem selekcji cech daje dobre rezultaty, a proponowane metody czasami przewyższają metody referencyjne. Dodatkową zaletą proponowanych algorytmów wykorzystujących optymalizację wielokryterialną jest możliwość wyboru najlepszego rozwiązania, czego nie oferują metody klasyczne. Przedstawione poniżej cele badawcze potwierdziły postawioną na wstępie hipotezę badawczą.

Zastosowanie optymalizacji wielokryterialnej do uczenia klasyfikatorów na podprzestrzeniach cech pozwala uzyskać modele o jakości predykcji nie gorszej niż metody referencyjne i pozwala na wybór rozwiązania dostosowanego do potrzeb użytkownika.

Aby udowodnić hipotezę, sformułowano kilka celów.

- **Opracowanie metod selekcji cech w oparciu o optymalizację wielokryterialną do konstruowania pojedynczych klasyfikatorów.**

Cel ten został osiągnięty poprzez zaproponowanie metod wykorzystujących optymalizację do przeprowadzenia selekcji cech. Metody wykorzystują algorytmy optymalizacyjne takie jak *GA* w wersji jednokryterialnej i *NSGA-II* w wersji wielokryterialnej. Zaletą stosowania optymalizacji jest uwzględnienie nie tylko jakości zbudowanego klasyfikatora, ale także kosztu funkcji, co jest szczególnie ważne w uczeniu uwzględniającym koszty. Jednoczesna optymalizacja dwóch kryteriów, maksymalizacja wydajności i minimalizacja kosztów, w przypadku optymalizacji wielokryterialnej daje zestaw rozwiązań, spośród których użytkownik może wybrać najbardziej dopasowane do swoich potrzeb. Zaproponowane metody osiągnęły jakość porównywalną z metodami klasycznymi, przy czym te ostatnie nie pozwalają na możliwość wyboru ze zbioru rozwiązań i zwracają tylko jedno rozwiązanie.

13.05.2024.

Joanna Klikowska

- **Opracowanie wielokryterialnej metody uczenia zespołów klasyfikatorów z wykorzystaniem uczenia maszyn wektorów nośnych (Support Vector Machines), które opierają klasyfikatory na podprzestrzeniach przestrzeni cech.**

Cel ten osiągnięto proponując metodę *SEMOOS*. *SEMOOS* jest zespołem składającym się z pojedynczych klasyfikatorów *SVM* wykorzystujących optymalizację wielokryterialną i algorytm *NSGA-II* do przeszukiwania przestrzeni cech i znajdowania dwóch parametrów klasyfikatorów *SVM*. *NSGA-II* zwraca zestaw takich rozwiązań, a każde z nich służy do uczenia modelu, który następnie jest dodawany do puli, tworząc klasyfikator zespołowy. Stosowanie w tej metodzie proponowanego losowania zbioru treningowego i testowego (center-based bootstrapping) oraz usuwanie modeli z finalnego zespołu klasyfikatorów (pruning) jest opcjonalne. Metoda została przetestowana na wielu niezbalansowanych zbiorach danych i uzyskała zadowalające wyniki w porównaniu z metodami referencyjnymi.

- **Opracowanie metody uczenia zespołu klasyfikatorów z wykorzystaniem klasyfikatorów bazowych drzew decyzyjnych na podprzestrzeniach przestrzeni cech i zagregowanych kryteriów.**

Cel został zrealizowany poprzez zaprezentowanie metody *DE-Forest* wykorzystującej algorytm optymalizacyjny *Differential Evolution* w celu znalezienia najlepszego wektora cech dla całego zespołu w odniesieniu do różnych zagregowanych metryk. Taki wektor jest odpowiednio przygotowany i na jego podstawie można wytrenować poszczególne modele drzew decyzyjnych, aby utworzyły zespół. Zaproponowana metoda często przewyższa metody referencyjne.

- **Opracowanie wielokryterialnej metody uczenia zespołu klasyfikatorów z wykorzystaniem klasyfikatorów bazowych drzew decyzyjnych na podprzestrzeniach przestrzeni cech w celu utworzenia lasu nielosowego.**

Cel ten osiągnięto proponując metodę *MOOforest*, czyli zespół zbudowany z poszczególnych drzew decyzyjnych przy użyciu algorytmu optymalizacji wielokryterialnej *MOEA/D*. *MOEA/D* przeszukuje przestrzeń cech dla całego zestawu w oparciu o jednocześnie dwa kryteria: *Precision* i *Recall*. Dzięki temu zwraca zbiór rozwiązań, z których przy pomocy funkcji *PROMETHEE II* wybierane jest jedno rozwiązanie. Modele są trenowane w oparciu o to rozwiązanie, które składa się na finalny zespół. Zaproponowana metoda w wielu przypadkach przewyższa metody referencyjne.

- **Opracowanie wielokryterialnej metody uczenia zespołu klasyfikatorów z wykorzystaniem klasyfikatorów bazowych na podprzestrzeniach przestrzeni cech i optymalizacji lokalnej.**

Cel został osiągnięty poprzez zaproponowanie metody *MOLO*. Jest to oryginalna metoda wykorzystująca wielokryterialną optymalizację lokalną w celu zbudowania zróżnicowanego zespołu. Każdy podstawowy model *DT*

jest trenowany na jednym wylosowanym podzbiorze. Optymalizacja przeszukuje możliwe rozwiązania na każdym etapie i dodaje jeden model do zespołu. Ograniczenia w znacznym stopniu zapobiegają rozprzestrzenianiu się poszukiwań, dzięki czemu algorytm wybiera ścieżki, które w danym momencie zapewniają najlepsze wyniki. Metodę *MOLO* przetestowano dla dwóch zestawów podwójnych kryteriów i może ona również obsłużyć przypadek trzech kryteriów. Na koniec *MOLO* zwraca kilka zestawów, spośród których użytkownik może wybrać ten, który najlepiej odpowiada jego potrzebom. Obszerne testy na danych niebalansowanych, porównujące metody proponowane i referencyjne, wykazały przewagę metody *MOLO*.

- **Eksperymentalna ocena otrzymanych metod.**

Cel ten osiągnięto poprzez porównanie wszystkich zaproponowanych metod: *SEMOOS*, *DE-Forest*, *MOOforest* i *MOLO*. Testy statystyczne i szczegółowe wyniki dla każdego zestawu danych wykazały charakterystykę każdej metody i obszary ich kompetencji.