Komitety maszyn wektorów podpierających z ewolucyjnie optymalizowanymi hiperparametrami modelu i zbiorami treningowymi

Politechnika Śląska
Wydział Automatyki, Elektroniki i Informatyki

mgr inż. Wojciech Dudzik

Promotor: **dr hab. inż. Michał Kawulok, prof. PŚ**

Promotor pomocniczy: **dr hab. inż. Jakub Nalepa, prof. PŚ**

Streszczenie

Maszyny wektorów podpierających (ang. support vector machine – SVM) są z powodzeniem stosowane w rozwiązywaniu różnorodnych problemów z dziedziny rozpoznawania wzorców, z wieloma udokumentowanymi przykładami w różnych dziedzinach, w tym w bioinformatyce, ekonomii i inżynierii oprogramowania. Działają poprzez znalezienie hiperpłaszczyzny separującej przykłady należące do dwóch klas, maksymalizując margines między hiperpłaszczyzną a najbliższymi przykładami (wektorami). Te najbliższe wektory danych są często wybierane jako wektory podpierające, które określają granice decyzyjne (hiperpłaszczyznę) klasyfikatora SVM i są kluczowe dla procesu klasyfikacji nowych danych. Metoda ta prowadzi do zbudowania dobrze uogólniającego klasyfikatora, który może obsługiwać dane o wysokiej wymiarowości i/lub zaszumione. Jednakże główną wadą SVM-a jest wysoki koszt obliczeniowy procesu uczenia. Proces ten ma złożoności czasowe i pamięciowe rzędu $O(t^3)$ i $O(t^2)$ odpowiednio, gdzie t to rozmiar zbioru uczącego. Ponadto hiperparametry funkcji jądrowej (ang. kernel function) mogą mieć istotny wpływ na wydajność klasyfikatora.

Mimo tych wyzwań, SVM-y nadal odgrywają ważną rolę w dziedzinie uczenia maszynowego. Wielu badaczy proponuje różnorakie metody pozwalające stosować klasyfikator SVM dla dużych danych, takie jak przyspieszenie procesu uczenia poprzez lepsze wykorzystanie akceleracji sprzętowych lub ograniczenie rozmiaru zbioru uczącego poprzez wybór najbardziej obiecujących wektorów. Jednak, jak pokazano w przeglądzie literatury, problemy optymalizacji hiperparametrów oraz doboru zbioru uczącego i cech są w większości przypadków rozpatrywane niezależnie od siebie, co może negatywnie wpłynąć na wydajność klasyfikatora SVM.

Niniejsza rozprawa doktorska prezentuje skuteczne rozwiązania optymalizacji modeli SVM w kontekście binarnej klasyfikacji. Proponowane i zweryfikowane zostały algorytmy wykorzystujące wzajemną zależność wyboru zbioru uczącego i optymalizacji hiperparametrów. Ponadto praca dyplomowa wprowadza nowe metody konstruowania komitetów modeli SVM w formie kaskad, aby poprawić jakość klasyfikacji i rozszerzyć ich możliwości, zachowując jednocześnie szybki proces uczenia. W celu osiągnięcia tych celów, proponowane

metody opierają się na algorytmach ewolucyjnych, które wielokrotnie udowodniły swoja efektywność w rozwiązywaniu problemów o wysokiej złożoności.

Przeanalizowane metody obejmują obecne w literaturze algorytmy ewolucyjne do wyboru zbioru treningowego, które mają wadę polegającą na konieczności posiadania wcześniejszej wiedzy na temat hiperparametrów SVM. Aby rozwiązać ten problem, zaproponowano metody, które rozszerzają wcześniejsze algorytmy poprzez łączenie optymalizacji hiperparametrów w schemacie naprzemiennym. Zaproponowano również metodę jednoczesnej optymalizacji hiperparametrów, zbioru treningowego i zbioru cech poprzez algorytm o nazwie SE-SVM. Dalsze prace rozwijają tą metodę o wykorzystanie nowej adaptacyjnej funkcji jądrowej RBF. Te algorytmy zostały zaproponowane w celu weryfikacji pierwszej hipotezy, że jednoczesna optymalizacja zbioru treningowego i hiperparametrów SVM zmniejsza czas treningu i klasyfikacji w porównaniu z innymi najnowocześniejszymi metodami zaproponowanymi w tym celu, nie wpływając na jakość klasyfikacji. Schemat zespołowy wykorzystuje wcześniej zaprezentowane metody do budowy kaskad SVM-ów, które powinny zapewnić najlepszą możliwą jakość klasyfikacji przy optymalnych czasach szkolenia i klasyfikacji w sensie Pareto. Ich celem jest weryfikacja drugiej hipotezy, że komitety klasyfikatorów SVM utworzone za pomocą algorytmów ewolucyjnych zapewniają poprawę jakości klasyfikacji w porównaniu z innymi uznanymi metodami, w tym istniejącymi algorytmami do budowania komitetów klasyfikatorów SVM.

Wszystkie te techniki zostały przeanalizowane eksperymentalnie i porównane z innymi popularnymi klasyfikatorami i najnowocześniejszymi technikami optymalizacji SVM-ów przy użyciu sztucznych zbiorów danych 2D i zbiorów danych benchmarkowych. Badania obejmują analizę jakościową wizualizowanych granic decyzyjnych na zbiorach danych 2D. Korzystając z tych zbiorów danych, przedstawiono nową koncepcję regionów pewnych w komitetach klasyfikatorów (w postaci kaskad). Ponadto przeprowadzona została analiza ilościowa zarówno na zbiorach danych 2D, jak i na zestawie 31 zbiorów danych referencyjnych. Wszystkie te wyniki były analizowane również za pomocą testów statystycznych.

Wyniki pokazują, że jednoczesna optymalizacja zapewnia dużą redukcję czasu treningu i klasyfikacji w porównaniu z innymi metodami bez istotnego wpływu na utratę jakości klasyfikacji w porównaniu z metodami operującymi na pełnym zbiorze uczącym. To pozytywnie potwierdza pierwszą hipotezę postawioną w pracy. Co więcej, w obu przypadkach, zarówno dla zbiorów danych 2D, jak i zbiorów danych benchmarkowych, przedstawione metody budowy komitetów klasyfikatorów (ECE-SVM) przewyższają wszystkie inne dobrze znane metody, w tym istniejące algorytmy do budowania zespołów SVM-ów. Te wyniki pozytywnie potwierdzają drugą hipotezę. Chociaż podejście ECE-SVM może wymagać większych zasobów obliczeniowych (w porównaniu z SE-SVM), potencjalne korzyści w jakości klasyfikacji sprawiają, że jest to wartościowa technika. Pokazano również, że zarówno SE-SVM, jak i ECE-SVM leżą na froncie Pareto, gdy jakość klasyfikacji jest odniesiona w stosunku do czasu klasyfikacji lub treningu.