Streszczenie

Popularność wykorzystania metod sztucznej inteligencji oraz uczenia maszynowego jest stale rosnąca i zauważalna w takich dziedzinach jak cyberbezpieczeństwo, optymalizacja, finanse, medycyna, opieka zdrowotna, prawo, media czy też edukacja.

Klasyfikacja nadzorowana jest jedną z metod uczenia maszynowego. Algorytmy klasyfikacji nadzorowanej budują pewien model, który nazywany jest też klasyfikatorem, z wykorzystaniem poetykietowanych danych tworzących zbiór uczący. Utworzony model jest następnie wykorzystywany do nadawania etykiet nowym, niepoetykietowanym wcześniej obiektom.

Ważną rolę wśród algorytmów klasyfikacji nadzorowanej odgrywają klasyfikatory łączone. Zauważono, że użycie wielu modeli, które tworzą pewien zespoł prowadzi do poprawienia jakości klasyfikacji.

Problem badawczy

Jedno z podejść do łączenia klasyfikatorów zakłada użycie ich reprezentacji geometrycznej wyrażonej jako granice decyzyjne. Wykorzystanie granic decyzyjnych, które są efektem uczenia klasyfikatora pozwala na zdefiniowanie nowych algorytmów łączenia klasyfikatorów. Algorytmy te nie wykorzytują wektora prawdopodobieństw etykiet klas czy też samych etykiet klas.

W rozprawie zaproponowano 4 autorskie algorytmy wykorzystujące granice decyzyjne wyznaczone przez drzewa decyzyjne.

Głównym celem badań było poszerzenie wiedzy na temat różnorodnych systemów wielu klasyfikatorów opartych na ich geometrycznych właściwościach, zaprojekowanie i implementacja nowych algorytmów oraz ich ewaluacja w odniesieniu do metod referencyjnych: głosowania większościowego i lasu losowego. Hipoteza badawcza przyjęta w pracy sformułowana jest w następujący sposób.

Hipoteza badawcza. *Wykorzystanie granic decyzyjnych wytrenowanych drzew decyzyjnych pozwala na zbudowanie klasyfikatora łączonego o większej wartości miary jakości klasyfikacji niż klasyfikator łączony jakim jest las losowy lub głosowanie większościowe korzystające z tego samego zbioru wytrenowanych drzew decyzyjnych.*

Opracowane algorytmy łączenia klasyfikatorów zostały porównane z innymi algorytmami dedykowanymi do zespołu klasyfikatorów: (ważonego) głosowania większościowego oraz lasu losowego. Następujące miary jakości klasyfikacji zostały wykorzystane do udowodnienia hipotezy badawczej: dokładność (ACC), współczynnik korelacji Matthews'a (MCC) oraz współczynnik F1 dla zbiorów binarnych oraz dokładność, precyzja, czułość i wskaźnik F1 (trzy ostatnie w formie mikro– i makro–uśrednionej) dla pozostałych problemów. Przeprowadzono badania eksperymentalne z wykorzystaniem testowych baz danych na licencji open–source pochodzących z platform UCI oraz KEEL. Zaproponowana hipoteza badawcza została zweryfikowna z wykorzystaniem nieparametrycznych testów statystycznych, które uwzględniały różne miary jakości klasyfikacji.

Zaimplementowano i przetestowano cztery algorytmy integracji drzew decyzyjnych wykorzystujące ich reprezentację geometryczną. Analiza statystyczna otrzymanych wyników wskazuje, że zaproponowane metody w wielu przypadkach przewyższają w działaniu referencyjne techniki integracji. W rozprawie wykazano, że zaproponowany algorytm jest uogólnieniem ważonego głosowania większościowego.

Osiągnięte wyniki

Pierwszy algorytm integracji wykorzystuje dwie odległości w przestrzeni cech rozpatrywanej jako układ współrzędnych z wyróżnionymi granicami decyzyjnymi wytrenowanych klasyfikatorów. Pierwsza odległość liczona jest od centroidu, druga – od granicy decyzyjnej. Testowany obiekt klasyfikowany jest przez wytrenowany model drzewa decyzyjnego. Następnie centroid dla wybranej klasy determinuje wartość pierwszej odległości. Obie wartości mapowane są z użyciem funkcji Gaussa. Zbadano kilka zestawów parametrów funkcji mapującej, która pełni dwa zasadnicze zadania:

- Normalizacja obu odległości do wartości z zakresu $[0, 1]$.
- Obliczona wartość odzwierciedla wpływ odległości od granicy decyzyjnej na wagę.

Ostateczną odpowiedzią klasyfikatora jest etykieta klasy, której kombinacja liniowa obu odległości jest maksymalna. Kilka możliwych parametrów rozkładu funkcji mapującej

zostało przetestowanych w celu znalezienia najwyższej wartości miary jakości klasyfikacji. Testy statystyczne wskazują na znaczący wzrost wartości miar klasyfikacji ACC oraz MCC zaprezentowanej techniki w stosunku do referencyjnych metod łączenia.

Druga propozycja algorytmu korzysta ze statycznego podziału przestrzeni cech w procesie integracji. Przestrzeń cech podzielona jest na równe podprzestrzenie, z taką samą liczbą podziałów wzdłuż każdej z osi, których etykiety determinowane są przez klasyfikację ich punktów środkowych. Dla każdej podprzestrzeni dla danej etykiety obliczana jest waga zależna od objętości regionu klasyfikacji. Region klasyfikacji jest hiperkostką o maksymalnej objętości obejmującą punkty oznaczone tą samą etykietą przez dane drzewo decyzyjne. Dodatkowo podano teoretyczny dowód na to, że zaproponowana technika jest uogólnieniem ważonego głosowania większościowego. Implementacja algorytmu została sprawdzona z wykorzystaniem wspomnianych zbiorów danych. Zaobserwowano statystycznie znaczącą poprawę jakości klasyfikacji m. in. w takich miarach jak ACC i MCC.

W trzeciej zaproponowanej metodzie wykorzystano dynamiczny podział przestrzeni cech. Geometryczna reprezentacja wytrenowanych modeli determinuje podział przestrzeni cech. Dla każdej etykiety obliczana jest waga na podstawie klasyfikacji samej podprzestrzeni jak również jej sąsiadów. Połowę wagi stanowi wkład pochodzący z samej podprzestrzeni, wagi sąsiadujących podprzestrzeni sumują się natomiast do $\frac{1}{2}$. Wartości wag zależą od odległości między środkami sąsiadujących podprzestrzeni. Przeprowadzono analizę statystyczną i zaobserwowano poprawę w jakości klasyfikacji wyrażoną mikrouśrednioną precyzją, czułością i współczynnikiem F1 w porównaniu z referencyjnymi metodami: głosowaniem większościowym oraz lasem losowym.

W ostatniej pracy, podobnie jak w poprzedniej, wykorzystany jest dynamiczny podział przestrzeni cech, a wynik klasyfikacji w danej podprzestrzeni zależny jest od podprzestrzeni ją otaczających. Wprowadzono dodatkowe udoskonalenia:

- W definicji odległości między podprzestrzeniami wykorzystano średnią z punktów treningowych w danej podprzestrzeni zamiast jej geometrycznego środka. Odległość między średnimi jest jednocześnie odległością między podprzestrzeniami.
- W obliczaniu wagi rozpatrywane są jedynie podprzestrzenie, w których obiektów treningowych każdej klasy jest przynajmniej 5% wszystkich obiektów w danej podprzestrzeni.
- Większy zakres sąsiadujących podprzestrzeni jest analizowany w przypadku odfiltrowania całego pierścienia sąsiadów. Ta procedura powtarzana jest do momentu aż znaleziony zostanie niepusty pierścień.

Omawiany algorytm został w badaniach eksperymentalnych porównany z wykorzystaniem wspomnianych baz danych z referencyjnymi metodami (głosowaniem większościowym i lasem losowym) jak również zaprezentowanym wcześniej, pierwotnym algorytmem. Zaobserwowano statystycznie znaczącą poprawę w jakości klasyfikacji w stosunku do metod referencyjnych. W szczególności dla miar jakości klasyfikatorów: dokładności, mikro– i makrouśrednionego współczynnika F1 wykazana została statystycznie znacząca różnica, co dowodzi hipotezy badawczej.

We wszystkich badaniach zachowano spójną strukturę badań. Przeprowadzono wyczerpującą ewaluację z użyciem wielu testowych baz danych. Kilka miar jakości klasyfikacji zostało obliczonych mając na względzie różnorodną charakterystykę baz danych, na przykład współczynnik niezbalansowania: dokładność, współczynnik korelacji Matthews'a oraz współczynnik F1 dla binarnych oraz dokładność jak również mikro– i makrouśrednioną precyzję, czułość i współczynnik F1 dla pozostałych problemów. Do implementacji zaprezentowanych algorytmów wykorzystano framework Spark i język Scala. Kod źródłowy jest publicznie dostępny na portalu github, między innymi w celu weryfikacji opracowanych algorytmów przez innych badaczy zajmujących się problemami uczenia maszynowego: https://github.com/TAndronicus/dtree-merge, https://github.com/TAndronicus/dtree-merge-scoring, https://github.com/TAndronicus/dynamic-dtree, https://github.com/TAndronicus/dynamic-ring.

10.03.2021 Jędrzej Biedrzycki

Ta praca została częściowo wsparta przez polskie Narodowe Centrum Nauki, nr grantu 2017/25/B/ST6/01750.