

POPULARNONAUKOWE STRESZCZENIE PROJEKTU

Uczenie maszynowe jest jedną z najszybciej rozwijających się dziedzin sztucznej inteligencji i informatyki. Zajmuje się projektowaniem algorytmów pozwalających komputerom działać na podstawie wiedzy uzyskanej z danych, bez konieczności zaprogramowania tego działania. Wywodzi się ze statystyki i teorii prawdopodobieństwa, ale istotne są w nim również wyzwania obliczeniowe. Algorytmy uczenia maszynowego mogą być stosowane w wielu sytuacjach, zaczynając od wyszukiwarek internetowych, a kończąc na zastosowaniach w biologii.

Rzeczywiste zastosowania uczenia maszynowego charakteryzują się często bardzo dużą liczbą przykładów uczących oraz cech służących do ich opisu. Także liczba klas (etykiet), do których przypisywane są przykłady, może być bardzo duża, sięgając takich wartości jak setki tysięcy. Problemy z tak dużą liczbą etykiet nazywane są klasyfikacją ekstremalną (ang. *extreme classification*). Skalę tych problemów można zobrazować na przykładzie przypisywania kategorii do artykułów na Wikipedii. Każdy artykuł stanowi przykład, a jego cechy to słowa występujące w tekście. Etykiety to kategorie, do których należy artykuł. Zbiór danych utworzony na bazie aktualnej angielskiej Wikipedii zawierałby nie tylko miliony cech i przykładów uczących, ale także ponad milion etykiet.

Standardowe algorytmy uczenia maszynowego nie są w stanie wystarczająco efektywnie przetworzyć zbiorów danych z tak dużą liczbą etykiet. Stąd problemy klasyfikacji ekstremalnej wymagają bardziej efektywnych algorytmów, czyli algorytmów o niższej złożoności. Celem tego projektu jest redukcja złożoności algorytmów uczących, a dokładniej czasu przetwarzania i zajmowanej pamięci, bez znacznego pogorszenia jakości predykcji. Aby osiągnąć ten cel, proponujemy użycie nowego podejścia, polegającego na transformacji oryginalnego problemu klasyfikacji ekstremalnej do odpowiedniego problemu z dziedziny predykcji strukturalnej (ang. *structured prediction*). Prosty przykład takiej transformacji to zakodowanie etykiet za pomocą sekwencji bitów, które później traktowane są jako sekwencja, dla której należy przewidzieć prawidłowe etykietowanie. Jednym z zadań tego projektu jest poszukiwanie odpowiedniej struktury, za pomocą której kodowane są etykiety. Inne zadania będą dotyczyły szybkich algorytmów uczących i ich związków z istniejącymi algorytmami klasyfikacji ekstremalnej.

Transformacja problemów klasyfikacji ekstremalnej do problemów predykcji strukturalnej jest bardzo interesującym obszarem badań. Jednym z powodów, dlaczego podejmujemy ten temat, jest fakt, że badania wstępne dały obiecujące rezultaty. Podczas nich pojawiły się jednak nowe pytania i otworzył się nowy obszar badań, wymagający zrozumienia. Aby umożliwić szerokie zastosowanie opisanych metod konieczne jest lepsze ich zbadanie.

Problemy klasyfikacji ekstremalnej stają się coraz powszechniejsze. Są one komplementarne z uczeniem się reprezentacji (ang. *representation learning*), głównym celem uczenia głębokiego (ang. *deep learning*). Problem badawczy projektu jest ogólny, a jego rezultaty mogą być użyte w innych dziedzinach uczenia maszynowego.