Długotrwałe uczenie maszynowe na podstawie danych strumieniowych

Lifelong Machine Learning on Data Stream

Obecnie notuje się gwałtowny wzrost zainteresowania pracami nad algorytmami uczenia maszynowego (ML) do analizy danych strumieniowych, a modele decyzyjne muszą być dostarczane szybko, często bazując na dużych wolumenach danych. Stąd też kwestia uzyskiwania użytecznych informacji staje się dużym wyzwaniem. Z reguły dostarczane dane nie mają jednorodnej struktury, a także mogą zawierać informacje nadmiarowe, szum, obserwacje odstające, a w wielu przypadkach również niekompletne opisy obiektów. Powyższe cechy skłoniły środowisko naukowe do podjęcia wyzwania opracowania takich algorytmów uczenia maszynowego, które pozwalają na analizę dużych wolumenów danych o charakterze strumieniowym. Dotychczasowe badania koncentrowały się głównie na analizie adaptacyjnych modeli decyzyjnych, znajdujących swoje zastosowanie w zadaniach grupowania, wykrywania anomalii, uczenia częściowo nadzorowanego, czy wykrywania pojawienia się nowych klas.

Główną cechą omawianych problemów jest sekwencyjne pojawianie się dużej ilości danych, tworzących nieskończony strumień, a opracowywany model decyzyjny musi być zawsze gotowy do podjęcia decyzji, przy czym należy przy jego konstrukcji i aktualizacji uwzględniać ograniczone zasoby pamięciowe i obliczeniowe. Co więcej, możemy mieć do czynienia z tzw. niestacjonarnymi strumieniami danych, tj. przypadkiem, gdy rozkłady statystyczne danych mogą ulec zmianie, zmuszając model do uwzględnienia ich dynamiki w trakcie eksploatacji. Zjawisko to nazywa się *dryfem koncepcji* (ang. *concept drift*), a jego natura może się bardzo różnić dla poszczególnych zadań. Kolejną ważną kwestią jest dostępność etykiet klas. Wiele technik zakłada, że etykiety są zawsze dostępne, bądź że można je uzyskać z niewielkim opóźnieniem. Niestety, etykietowanie danych wiąże się ze znacznymi kosztami, zatem naiwnym jest założenie o posiadaniu pełnej wiedzy na ich temat. Stąd też, metody wykorzystujące (*częściowo nadzorowane uczenie się*) i tzw. (*uczenie aktywne*) zyskują znaczną popularność.

Systemy długotrwałego uczenia maszynowego (ang *longlife machine learning*) mogą przezwyciężać ograniczenia algorytmów uczenia statystycznego, wymagających dużej liczby przykładów uczących i są odpowiednie do uczenia się pojedynczego zadania. Jednak badania nad nimi są wciąż w początkowej fazie i niestety wiele pytań związanych z tą dziedziną pozostaje bez odpowiedzi. Kluczowe problemy – które należy rozwiązać w systemach tej klasy – dotyczą zagadnień związanych z wykorzystaniem wcześniej nabytej wiedzy, modelowania funkcji, zachowania wiedzy z poprzednich zadań, transferu wiedzy do przyszłych zadań, aktualizacji wiedzy i sposobów uwzględnienia wiedzy użytkownika w procesie uczenia. Również pojęcie *zadania*, które pojawia się w wielu formalnych definicjach modeli długotrwałego uczenia maszynowego, wydaje się trudne do zdefiniowania. Często trudno jest wyznaczyć granicę pomiędzy następującymi po sobie zadaniami. Jednym z głównych wyzwań jest tzw. dylemat *stabilności i plastyczności*, w którym systemy uczenia się muszą iść na kompromis pomiędzy opanowywaniem nowych zadań, a pamiętaniem starych rozwiązań. Jest to szczególnie widoczne w zjawisku gwałtownego zapominania (ang. *catastrophic forgeting*), które definiuje się jako całkowite zapominanie pojęć poznanych wcześniej przez sieć neuronową, gdy prezentujemy jej dane z nowego zadania. Kolejne otwarte wyzwanie dotyczy oceny modeli długotrwałego uczenia maszynowego.

Klasyfikacja strumieni danych zwykle przedstawia przypadek dużych przepływów danych, wśród których od czasu do czasu pojawiają się określone zdarzenia. Wiedza statystyczna na ich temat może pomóc w zaprojektowaniu klasyfikatorów do tego rodzaju zadań. Głównym problemem jest to, że im rzadziej występuje model docelowy, tym bardziej kosztowne jest pozyskiwanie informacji. Dlatego – w trakcie badań – planujemy uwzględnić strumienie ze zdarzeniami rzadkimi.

W trakcie projektu zamierzamy osiągnąć następujące cele:

[A] Opracowanie nowych metod radzenia sobie z dryftem koncepcji.

[B] Zaproponowanie metod analizy danych strumieniowych przy użyciu systemów rekurencyjnych.

[C] Opracowanie modeli długotrwałego uczenia maszynowego uwzględniających strumieniowy charakter danych.

W projekt zaangażowani są naukowcy z Faculty of Electrical Engineering and Computer Science (FEI), VŠB-Technical University of Ostrava (czechy) oraz Katedry Systemów i Sieci Komputerowych Politechniki Wrocławskiej.