Metody wyjaśniania modeli uczenia maszynowego dla statycznych i zmiennych danych

Sztuczna inteligencja i w szczególności uczenie maszynowe są tymi dziedzinami informatyki, które się rozwijają bardzo intensywnie w ostatnich kilkunastu latach. Wiąże się to zarówno z proponowaniem nowych algorytmów, jak i dostępem do danych o większym rozmiarze i złożoności, pojawieniem się specjalizowanych bibliotek programistycznych oraz łatwiejszą osiągalnością bardziej wydajnego sprzętu obliczeniowego. Doprowadziło to do sukcesu systemów inteligentnych w takich dziedzinach, jak np., rozpoznawanie obrazów, przetwarzanie języka naturalnego, rozpoznawanie mowy, robotyka, sterowanie autonomicznymi urządzeniami, diagnostyka medyczna i eksploracja różnego rodzaju danych.

W szczególności, od kilkunastu lat dane są generowane automatycznie, np. w sieciach sensorów, oraz udostępniane w postaci tzw. strumieni danych. Takie dane napływają w sposób ciągły, mają duże rozmiary, wymagają odpowiednio szybkiego przetwarzania i reakcji ze strony inteligentnego systemu. Ponadto z uwagi na niestacjonarną charakterystykę źródła danych dochodzi do zmian rozkładów prawdopodobieństwa przykładów w strumieniu jako tzw. dryfu pojęć. Ponieważ model predykcyjny był uczony na danych historycznych, traci swoje zdolności predykcyjne wraz z tempem zachodzenia dryfu. Dlatego wprowadzono wiele metod detekcji występowania dryfu oraz specjalizowanych przyrostowych algorytmów uczenia modeli adaptujących się do niego. Same modele są jednak dość złożone, jak np. zespoły klasyfikatorów i tym samym trudne do analizy i zrozumienia zwłaszcza w kontekście zmienności.

Należy zauważyć, że współczesne systemy uczące się są złożonymi modelami, które nie są przejrzyste i łatwe do analizy nawet dla specjalistów. Równocześnie w wielu zastosowaniach, np. systemach krytycznych lub przy wspomaganiu podejmowania decyzji wobec ludzi, oczekuje się wyjaśniania zasad wypracowania ich predykcji lub rekomendacji decyzji. W odpowiedzi rozwija się dziedzina wyjaśnialnej sztucznej inteligencji (z ang. XAI). W jej ramach proponuje się metody dostarczające informacje o wewnętrznym działaniu systemu albo lokalnego wyjaśnienia przyczyn predykcji dla konkretnego przypadku. Sama XAI rozwijana jest jednak dotychczas głównie/tylko dla statycznych danych.

Podsumowując, obie dziedziny, tj. wyjaśnialność modeli uczenia maszynowego oraz uczenie się modeli ze strumieni danych, rozwijają się niezależnie od siebie i brakuje badań nad adaptacją metod wyjaśnialności dla zmiennych strumieni danych. Niniejszy projekt ma na celu wypełnić tę lukę dla wybranych metod. W ramach proponowanego projektu badawczego zamierzamy realizować: (1) Badanie własności wyjaśnień kontrfaktowych; (2) Wielokryterialną analiza zbioru wielu wyjaśnień; (3) Badanie współzależności zmian w strumieniu danych, modelach oraz wyjaśnieniach; (4) Przyrostowe uogólnienie metod generowania kontrfaktowych i prototypowych wyjaśnień dla zmiennych strumieni; (5) Wykorzystanie wyjaśnień zmian modeli uczenia maszynowego do opisania dryfu pojęć; (6) Badanie sekwencji zmian wyjaśnień w strumieniach.

Oczekiwane wyniki tego projektu powinny dostarczyć nowych wskazówek metodologicznych na temat stosowania XAI do zmieniających się strumieni danych i być przydatne dla poprawy działania modeli, korekcji ewentualnych błędów oraz zwiększenia zaufania ludzi do używania takich systemów.