Popularnonaukowe streszczenie projektu

Niekompletność danych można scharakteryzować jako całkowity brak wartości jednego lub więcej atrybutów (cech) opisujących dany obiekt (instancję). Niepewność danych jest pojęciem bardziej złożonym. Jest ono badane od wielu lat i wyróżnić można dwa główne podejścia do jego rozumienia – epistemiczne i ontyczne. W obu niepewny koncept opisywany jest za pomocą zbioru wszystkich możliwych jego reprezentacji. W podejściu epistemicznym bez dodatkowej wiedzy nie jest możliwe wybranie właściwej reprezentacji mimo, że taka istnieje. W rozumieniu ontycznym wszystkie reprezentacje są równie akceptowalne i dlatego nie ma konieczności ich rozróżniania. Przykładowo rozważmy pojęcie miesięcznej pensji adiunkta opisanej przez przedział [3000zł, 5000zł]. Przedział ten będzie interpretowany ontycznie jeśli reprezentuje on widełki płacowe adiunkta w pewnej instytucji. Z drugiej strony ten sam przedział może być rozumiany epistemicznie, gdy reprezentuje on nieznane wynagrodzenie konkretnego adiunkta.

W przeciągu ostatnich kilku lat nastąpił znaczny wzrost zainteresowania problemem klasyfikacji opartej na danych niekompletnych i niepewnych. Spowodowane jest to przede wszystkim coraz powszechniejszym wykorzystaniem technik eksploracji danych (ang. data mining) i uczenia maszynowego w różnych obszarach badawczych oraz życia codziennego. Powstaje zatem potrzeba przetwarzania danych, które nie były zbierane z myślą o wykorzystaniu w klasyfikacji. Dane te często są nieprecyzyjne oraz niekompletne lub ogólniej niepewne. Powszechność zjawiska niepewności danych w rzeczywistym świecie i praktycznych zastosowaniach jest niepodważalna i dlatego nie można jej ignorować.

Choć istnieje wiele metod klasyfikacji, niewiele z nich podejmuje problem klasyfikacji opartej na danych niepewnych w pełnej ogólności. Zasadniczą wadą nielicznych istniejących metod jest duży poziom skomplikowania, co uniemożliwia ich zastosowanie we wspomaganiu decyzji, gdzie konieczna jest duża przejrzystość oferowanego modelu. Proponowane badania mają na celu opracowanie łatwej w interpretacji metody klasyfikacji danych niepewnych.

Okazuje się, że w tej sytuacji wiele znanych metod klasyfikacji zawodzi, gdyż nie są w stanie operować w warunkach niekompletnej informacji. Głównym celem naukowym projektu jest opracowanie efektywnej metody klasyfikacji dla danych niepewnych. W ramach badań opracowane zostaną algorytmy przetwarzania i klasyfikacji takich danych oparte na miarach podobieństwa. Określone zostaną miary podobieństwa niepewnych zbiorów rozmytych (ang. Hesitant Fuzzy Sets, HFS), które umożliwią adekwatne i efektywne modelowanie szeroko rozumianej niekompletności i niepewności danych.

Realizacja projektu może w przyszłości przyczynić się do zmiany podejścia do problemu klasyfikacji danych niekompletnych i niepewnych. Obecnie najczęściej stosuje się metody oparte na edycji zbioru danych, które w rezultacie mogą prowadzić do zubożenia (usuwanie instancji/atrybutów) lub uproszczenia struktury danych (imputacje).

Problem danych niepewnych jest szczególnie istotny w diagnostyce medycznej. Niekompletność i niepewność danych często jest ich naturalną i nieusuwalną cechą. Stąd wszelkie próby obejścia tego problemu prowadzą do modeli, które nie odzwierciedlają rzeczywistości. Edycja zbioru danych zazwyczaj nie jest dopuszczalna w zastosowaniach medycznych. Usuwanie instancji niekompletnych z często i tak skromnego zbioru danych, znacząco zawęża możliwości badawcze. Z drugiej strony podejścia oparte na imputacji, nie mogą być stosowane do klasyfikacji konkretnych przypadków medycznych ze względu na duże ryzyko błędnej diagnozy, gdyż imputowane wartości w żaden sposób nie odzwierciedlają aktualnego stanu pacjenta.

Wyniki badań będą stanowiły podstawę do przyjęcia zupełnie innej postawy. Zamiast ignorować, czy sztucznie polepszać jakość danych wejściowych, należy bezpośrednio uwzględnić ją w modelu danych, a następnie w samym klasyfikatorze zarówno na etapie uczenia jak i testowania. Podejście takie ma wiele przewag względem klasycznej edycji danych. Klasyfikator dysponując bardziej dokładnym (choć niekompletnym) opisem rzeczywistości, będzie mógł wyuczyć się większej liczby zależności, a następnie dokonać pełniejszej klasyfikacji. Bardzo ważna jest również możliwość oszacowania na ile konkretny wynik klasyfikacji jest pewny. Oczywiście, dane bardzo niskiej jakości (niepewne) mogą prowadzić do niepewnej klasyfikacji.

Rozwój i popularyzacja tego podejścia do klasyfikacji danych niepewnych przyczyni się do zastosowania go w problemie wspomagania decyzji, szczególnie wspomaganiu diagnostyki medycznej.